Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Research Paper

Texture and artifact decomposition for improving generalization in deep-learning-based deepfake detection

Jie Gao [a,b], Marco Micheletto [b], Giulia Orrù [b,*], Sara Concas [b], Xiaoyi Feng [a], Gian Luca Marcialis [b], Fabio Roli [b,c]

[a] *Northwestern Polytechnical University, 1 Dongxiang Road, Xi'an, 710129, China*
[b] *University of Cagliari, Via Marengo 3, Cagliari, 09123, Italy*
[c] *University of Genova, via alla Opera Pia 13, Genova, 16145, Italy*

ARTICLE INFO

Keywords:
DeepFake detection
Generalization
Texture
Artifact
Ensemble learning strategy

ABSTRACT

The harmful utilization of DeepFake technology poses a significant threat to public welfare, precipitating a crisis in public opinion. Existing detection methodologies, predominantly relying on convolutional neural networks and deep learning paradigms, focus on achieving high in-domain recognition accuracy amidst many forgery techniques. However, overseeing the intricate interplay between textures and artifacts results in compromised performance across diverse forgery scenarios. This paper introduces a groundbreaking framework, denoted as Texture and Artifact Detector (TAD), to mitigate the challenge posed by the limited generalization ability stemming from the mutual neglect of textures and artifacts. Specifically, our approach delves into the similarities among disparate forged datasets, discerning synthetic content based on the consistency of textures and the presence of artifacts. Furthermore, we use a model ensemble learning strategy to judiciously aggregate texture disparities and artifact patterns inherent in various forgery types, thereby enabling the model's generalization ability. Our comprehensive experimental analysis, encompassing extensive intra-dataset and cross-dataset validations along with evaluations on both video sequences and individual frames, confirms the effectiveness of TAD. The results from four benchmark datasets highlight the significant impact of the synergistic consideration of texture and artifact information, leading to a marked improvement in detection capabilities.

1. Introduction

DeepFake technology (Tolosana et al., 2020) refers to the creation or synthesis of fake content based on deep learning methods, such as images (Carlini and Farid, 2020), audio (Conti et al., 2022; Chen et al., 2020), and video (Korshunov and Marcel, 2019; Yu et al., 2021). The generation of synthetic or manipulated content containing the faces of individuals is among the most notorious of these; the principle behind it is replacing one face in a picture or video with another, creating a convincingly realistic but fake representation. The broader public became aware of this technology in 2017 when an autoencoder–decoder was used to create pornographic content, revealing the technology's potential for misuse (Zhang, 2022). Since then, deepfakes have been used to create fraudulent identities, manipulate political statements, and undermine trust in media and information sources, leading to widespread ethical, legal, and social implications. With the continuous development of generative adversarial networks, the deepfake technology based on the idea of adversarial games (Wang et al., 2022) has also been improved, and human eyes now struggle to recognize synthesized data. Moreover, the rapid advancement and accessibility of deepfake-generating tools have democratized the ability to create convincing fake content, exacerbating the detection challenge. Traditional detection methods are becoming increasingly inadequate as they fail to keep pace with the evolving sophistication of deepfake techniques. In this context, the urgency of implementing robust countermeasures is evident. Over the years, significant research has been dedicated to developing reliable deepfake detection methods (Rana et al., 2022). Despite these efforts, traditional algorithms often struggle with generalization across the varied techniques used in digital content manipulation. For instance, FaceSwap targets the entire face for alteration (Korshunova et al., 2017), whereas NeuralTexture specifically modifies the mouth region (Thies et al., 2019), as depicted in Fig. 1. The broad spectrum of manipulation methods within DeepFake technology,

[*] Corresponding author.
E-mail addresses: jie_gao@mail.nwpu.edu.cn (J. Gao), marco.micheletto@unica.it (M. Micheletto), giulia.orru@unica.it (G. Orrù), sara.concas90C@unica.it (S. Concas), fengxiao@nwpu.edu.cn (X. Feng), marcialis@unica.it (G.L. Marcialis), fabio.roli@unige.it (F. Roli).

https://doi.org/10.1016/j.engappai.2024.108450
Received 10 November 2023; Received in revised form 12 February 2024; Accepted 12 April 2024
Available online 19 April 2024

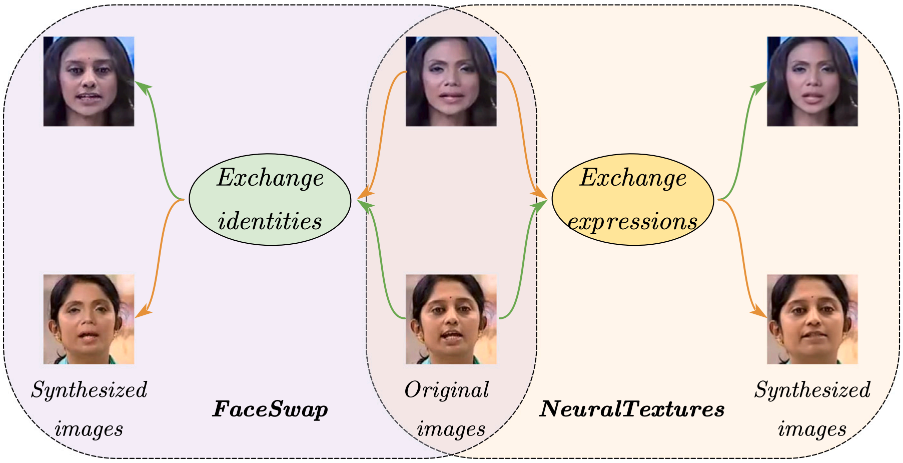


Fig. 1. Fake examples generated by different deepfake methods (FaceSwap and NeuralTextures, respectively) on the same source and target images. Regardless of the fake generation technology, face swap for the FaceSwap deepfake (left) and reenactment for the NeuralTextures deepfake (right), unnatural texture inconsistencies or artifacts are present.

from whole-face changes to localized adjustments, underscores the critical need for detection systems adept at discerning each approach's subtle differences. Furthermore, the challenge is compounded when excessive extraction of facial data overshadows the more subtle indicators of forgery, complicating the task of differentiating between authentic and manipulated content.

In our previous work (Gao et al., 2023), we noticed this problem and adopted the inconsistency of internal and external faces to explore the generalization of deepfake. We used the masks obtained by the face detector to directly separate the internal and external faces. However, this kind of mask does not provide variety and flexibility, ignoring the interactive nature of two key elements: artifacts and textures. Artifacts in this context refer to unintended distortions or anomalies introduced during the deepfake creation process, such as unusual pixel patterns or edge artifacts. Conversely, textures pertain to the natural patterns and details found in genuine images, such as skin complexion and hair. Hence, in this paper, we propose a novel Texture and Artifact Detector (**TAD**) to distinguish real and fake images. The proposal of TAD is based on three observations: (1) synthetic manipulations usually destroy the texture consistency of original images; (2) synthetic manipulation adds artifacts; (3) texture inconsistencies and artifacts often interact with each other. Therefore, our TAD aims to separate and fully exploit texture inconsistencies and artifact information. Specifically, the TAD framework consists of two parts: a texture inconsistencies detection network and an artifact detection network. Given that different manipulation types alter texture uniquely and produce varied artifacts, we propose an ensemble learning strategy for TAD. By integrating multiple learning models, the ensemble approach aims to boost the network's robustness, allowing it to handle a wider range of manipulations effectively. Moreover, this approach ensures that TAD is not only precise and versatile but also scalable, with the ability to adapt and expand with new manipulations.

To sum up, our contributions can be summarized as follows:

- We propose a novel Texture and Artifact Detector (TAD) for deepfake detection, which aims to separate mutually exclusive texture inconsistencies and artifact information, thereby weakening their mutual influence and improving the model's generalization ability.
- On the one hand, we capture texture differences with the aid of a self-supervised learning strategy. Specifically, by utilizing multiple deformable convolution stacks, we design trainable soft masks to separate source and target images in the texture encoder and further use the texture decoder to complete image restoration as a constraint to enforce the extraction of texture feature vectors. On the other hand, we utilize a shallow network to capture artifact information and design an artifact detector.
- An ensemble strategy is adopted to integrate texture discrepancies and artifact patterns of different forgery types, aiming to learn a broader range of forgery traces.
- We provide a comprehensive experimental validation of our model, including intra-dataset and cross-dataset tests and analyses on video sequences and individual frames. This extensive evaluation ensures a comprehensive assessment of the model's performance across varied testing conditions, highlighting its suitability for practical deployment.

The rest of this article is organized as follows: Section 2 introduces some related work on deepfake generation and deepfake detection, with a focus on works based on artifacts and texture inconsistencies detection. Section 3 presents texture and artifact models and their fusion via ensemble learning. Section 4 introduces the dataset and the SOTA methods used in this paper. Section 5 conducts large-scale comparison, validation, and analysis experiments. Finally, in Section 6, we obtain the corresponding conclusions and provide an outlook on future work.

2. Related work

The fundamentals of deepfake generation techniques are critical to understand the generalization issues and our proposed solution. In this section, we delve into the methodologies surrounding deepfake creation and its detection, focusing on techniques based on texture and artifact identification.

2.1. Deepfake generation

Two models predominantly used for deepfake synthesis are the Variational Auto-Encoders (VAEs) and the Generative Adversarial Networks (GANs). As depicted in Fig. 2, the traditional deepfake generation using VAE involves two encoder–decoder pairs. The encoders map both the source and target images into a feature space, while the decoders reconstruct these features back into images. By such training, the decoder associated with the source image becomes adept at retaining unique individual details and can manipulate other image features to make them closely resemble the original source image. This process allows for a high degree of control over the generated output, making VAEs particularly useful for tasks that require maintaining specific characteristics of the input images. Moreover, it is well-known that VAEs are able to perform very well on small-scale datasets (Bond-Taylor et al., 2022), but when it comes to bigger datasets, samples could become blurry and unrealistic. This is due to the fact that the obtained

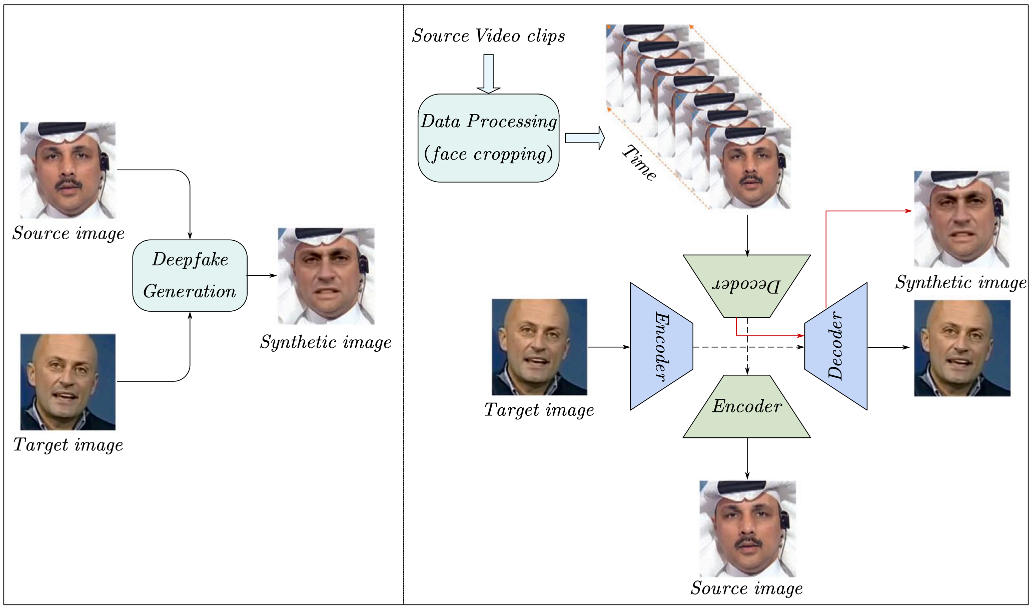


Fig. 2. The classic pipeline of deepfake generation. The left side of the figure shows the source image and target image through the deepfake generation process to synthesize fake data. The right side of the figure illustrates the specific process of deepfake generation, which includes two encoder–decoders.

model could be too simple and map different data points to the same encodings.

On the other side, Generative Adversarial Networks (Goodfellow et al., 2014) have been proposed in 2014. They model an adversarial game composed of generative and discriminative modules. The generative model can synthesize images with random noise, and the discriminative model is used to determine whether the synthetic image can be distinguished from the natural image. The idea of competition makes GANs able to generate very realistic images. The competitive nature of GANs drives them to produce highly detailed and convincing images, making them superior for applications that demand high realism. However, GANs can be challenging to train due to their adversarial nature. For example, if one of the models gets stuck in one local minimum, just a limited portion of the distribution is correctly learnt.

So far, most deepfake synthesis techniques, both face swap and reenactment, have been developed based on VAEs and GANs. The development of GAN technology has made manipulated images more and more realistic, almost indistinguishable from authentic images, in human eyesight. Over recent years, numerous methods for deepfake generation have surfaced and been incorporated into a range of applications. These include FaceShifter (Li et al., 2019), FaceSwap,[1] DeepFaceLab (Perov et al., 2020), DeepNude,[2] and ZOO,[3] among others. The widespread use of these applications allows anyone without professional skills to quickly generate highly realistic fake content with the help of a computer or mobile phone for malicious manipulation.

2.2. Deepfake detection

Deepfake technology represents an ongoing “arms race”. The emergence of hyper-realistic fake content presents significant risks to nations and societies, prompting researchers to devise counteracting detection strategies. Convolutional Neural Networks (CNNs) have emerged as a promising tool in this context due to their powerful image analysis capabilities. For instance, Rossler et al. (2019) proposed to transfer a well-known State of the Art (SOTA) CNN named XceptionNet to deepfake detection by replacing the final fully connected layers with two outputs. Similarly, Bonettini et al. (2021) employed an ensemble of trained CNN models to enhance detection accuracy. Li et al. (2020a) focused on examining the texture variations present at blending boundaries, which represent the areas in an image where two different visual components, such as a person’s face and a reference image, merge together; a common feature of deepfake images. Notably, their approach is promising since it allows using only real images to train a CNN. Moreover, Zhao et al. (2021) fully explored and designed a textural feature enhancement block and multi-attention maps to prompt the network to focus on multiple spatial regions. These CNN-based approaches excel in identifying deepfakes by analyzing textural patterns indicative of manipulated content, especially in texture-rich areas such as skin and hair (Liu et al., 2020). However, they frequently face challenges in generalizing to new manipulations due to their tendency to overfit to textures encountered during training.

One of the most effective approaches to address this problem is training models with synthetic data, as demonstrated by Shiohara and Yamasaki (2022), who designed a self-mixing image generation method. Chen et al. (2022) expanded the forged dataset (Faceforencis++ (Rossler et al., 2019)) using an adversarial training strategy and obtained a more generalized model. Sun et al. (2022) designed several data generation strategies for constructing positive and negative paired samples and made full use of inter-instance and intra-instance contrastive learning to make positive samples closer and negative samples further away. By enriching the diversity of training data, these methods potentially increase the generalizability of detection models. Nonetheless, the reliance on synthetic data may introduce biases that could affect real-world applicability.

Other works enhance generalization in deepfake detection by analyzing various manipulation artifacts beyond edge distortions and lighting mismatches, such as compression artifacts, unnatural facial expressions, and inconsistent blink patterns (Matern et al., 2019; Concas et al., 2022b; Liy and InIctuOculi, 2018; Sun et al., 2021). The strength of these methods lies in their ability to exploit specific deepfake generation weaknesses. However, they may struggle with more sophisticated deepfakes that have reduced artifacts due to advanced synthesis techniques.

According to the “arms-race” definition, deepfake detection technology chases the deepfake synthesis technology. While current detection

[1] https://github.com/MarekKowalski/FaceSwap/, (Last accessed: 06.11.2023).

[2] https://github.com/yuanxiaosc/DeepNude-an-Image-to-Image-technology, (Last accessed: 06.11.2023).

[3] https://www.zaoapp.net/, (Last accessed: 06.11.2023).

methods have shown promise, they exclusively target specific aspects of deepfakes, either artifacts or texture inconsistencies. This specialization can be a double-edged sword, providing high accuracy in specific contexts but potentially lacking in adaptability to new and evolving deepfake techniques. Within this context, our research introduces a novel approach, aiming to amalgamate the strengths of texture analysis and artifact detection to construct a more resilient and versatile deepfake detection framework capable of confronting the challenges posed by the next generation of synthetic media.

2.3. Textures and artifacts in deepfake detection

It is well known that the deepfake generation process can lead to the destruction of the original texture and the introduction of tampering artifacts. These defects are often used in the basic theory of deepfake detection. In the following parts, we analyze their advantages and disadvantages and draw comprehensive conclusions.

2.3.1. Texture inconsistencies

Generally speaking, the texture features of the face are highly discriminative, which is why biometric recognition technology for identity authentication based on facial features has good classification and recognition performance (Ahonen et al., 2006; Benavides et al., 2016). But what exactly is the facial texture? Although biometric systems based on facial texture have been widely used in various fields, such as security, education, finance, and transportation, facial texture cannot be formally described.

From the perspective of human vision, the main structures that can be observed on the surface of the human face are the texture wrinkles and faintly visible hairs of the skin. Skin textures are the tiny, polygonal hills and ravines on the surface of human skin.

Different from image features such as grayscale and color, the texture is represented by the grayscale distribution of pixels and their surrounding spatial neighborhoods, that is local texture information. In addition, the repeatability of local texture information to different degrees represents the global texture information.

There are three reasons why facial texture is considered important in deepfake detection tasks.

1. The texture of real and fake faces is different. Sun et al. (2020) pointed out that the facial texture of a natural face is unique, and its distribution is discriminative. Although different local facial regions may expose different textures, they have overall regularity, such as geometric coherence or high-order smoothness of spatial variation. This property has been used many times for face recognition (Ahonen et al., 2006). However, during the deepfake generation, the individual perpetrating the manipulation substitutes the source face or a portion of it with that of the intended target. This action disrupts the textural integrity of the original character's face, particularly at the face edges of for face swap manipulations and at the eyes and mouth for reenactment manipulations (Fig. 1). Based on our observations as well as previous studies, we can conclude that the texture consistency and continuity properties of synthetic fake data are destroyed. In addition, Liu et al. (2020) found the inconsistency of real and fake face textures through observation, and they believed that global texture statistics were more robust to image editing.
2. Texture features are crucial for face recognition. Numerous discoveries (Geirhos et al., 2018) have pointed out the important role of object texture in object recognition by CNN. For example, Gatys et al. (2017) believe that even if the global shape structure is completely destroyed, CNN can still classify texture images. Even in the field of image restoration and 3D reconstruction, the importance of texture is unquestionable. For instance, in the field of 3D face reconstruction (Gecer et al., 2021), the key problem is how to reconstruct more realistic facial texture details, especially high-frequency details in texture and, subsequently, identity characteristics. In some deepfake technologies dedicated to face-swapping operations, the inner and outer faces have different identities, which is another reason texture can be used as a powerful basis for discrimination between real and fake (La Cava et al., 2023).
3. Post-processing methods deal with local details with difficulty. Existing post-processing methods make synthetic data difficult for the naked eye to distinguish real from fake. However, commonly used post-processing techniques such as Poisson fusion and color constraints focus on global facial information rather than local details, which makes the whole image seem realistic but neglects the local consistency. Accordingly, Zhao et al. (2021) can model the deepfake detection task as a fine-grained classification problem.

2.3.2. Artifacts

Although texture consistencies are crucial for neural network-based deepfake detection, artifacts due to post-processing operations can affect the extraction of texture features. The blurring effect caused by artifacts can make texture information weak and difficult to detect.

The term "artifacts" refers to all types of image disturbances and various other non-random disturbances that appear on the image during reconstruction (Liang et al., 2022). For instance, when crafting deepfake materials, several distinct types of artifacts often arise. These can include inconsistencies stemming from the fact that the target and source faces may have been captured with different acquisition devices, each leaving its unique type of noise or "device fingerprint" on the images (Wang et al., 2020). Mistakes during the face-merging process can also cause slight pixel misalignments, resulting in visible flaws. Additionally, when generative adversarial networks (GANs) are used to create forgeries, the upsampling step in the process tends to introduce specific types of artifacts (Zhang et al., 2019). Furthermore, synthetic videos that are manipulated on a frame-by-frame basis tend to accumulate inconsistencies over time, leading to temporal artifacts that can be tracked through the video sequence (Nguyen et al., 2021). These unavoidable signatures of tampering serve as a critical foundation for identifying and confirming the presence of deepfake content.

In this regard, many studies have shown that artifact-based detection methods constitute a significant branch of current frame-level deepfake detection methods (Li and Lyu, 2019a). For example, Zhou et al. (2017) proposed a two-stream network for face tampering detection to detect tampering artifacts and capture local noise residuals. Based on the classical frequency domain analysis, Durall et al. (2019) converted the unseen artifacts in the air domain to the frequency domain for detection, and achieved good results. Focusing on JPEG compression artifacts left behind during image acquisition and editing, Kwon et al. (2022) utilize discrete cosine transform (DCT) coefficients to locate image manipulations. This technique was refined by Perelli et al. in a tensor-based approach with excellent preliminary results (Concas et al., 2022b). Dong et al. (2023) claim that the lack of generalization capabilities is caused by the fact that they unintentionally learn information about face identity. They name this phenomenon "Implicit Identity Leakage" and try to overcome it by relying on local artifacts areas in the images in order to give less importance to the global identity.

Other works rely on temporal artifacts like the inconsistencies between subsequent frames or lack of consistency between audio and video: Knafo and Fried (2022), for example, apply a pre-trained backbone on a large different video dataset and adapt it to the deepfake detection task. Similarly, Haliassos et al. (2022) the authors design a two-stage approach named RealForensics where they first use self-supervision to learn the correspondence between audio and video modalities in natural videos, then they use the learned representations as targets to be predicted along with the classical binary classification task. The same authors (Haliassos et al., 2021) exploit the semantic

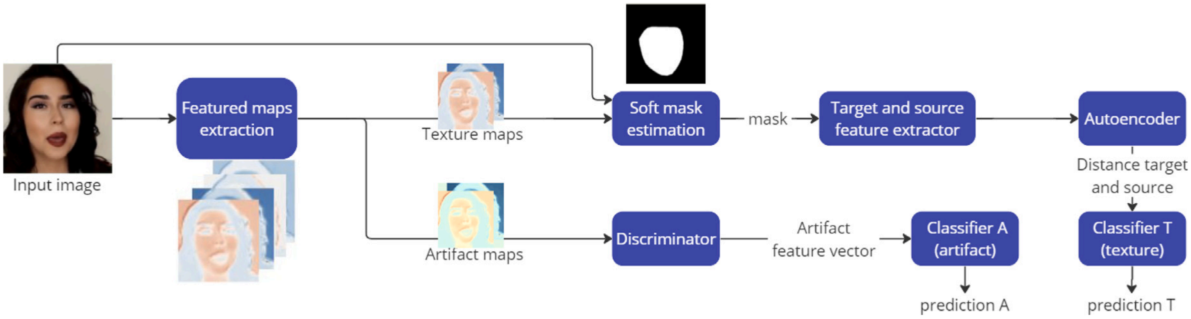


Fig. 3. Running example of the proposed method. The sets of feature maps extracted from the input image I are divided and used to calculate the texture difference between source (background) and target (foreground) and to analyze the presence of artefacts using a discriminative approach.

irregularities of mouth movements present in most fake videos: they pre-train a network for the lipreading task thus learning internal representations related to natural mouth motion. Then, a temporal network is used with fixed mouth embeddings of real and fake samples in order to discriminate between them. Pretty different is the approach by Cai et al. (2023). It is not strictly related to the deepfake detection problem, and provides a general method to learn universal facial representations that can be used in different fields like Facial Attribute Recognition, Facial Expression Recognition and Lip Synchronization. Cai et al.'s work includes a facial video masked autoencoder to learn generic facial embeddings and reconstructs spatio-temporal details from the densely masked facial regions. The goal is to extract local and global details to encode transferable features.

Although these methods perform well to some extent on synthetic data with obvious non-smooth boundaries, their performance drastically declines when the data synthesis is of high quality. This is the natural consequence of the basic assumption made by Li and Lyu (2019a), that cannot always (or anymore) be true according to the "arms-race" nature of the deepfake detection problem. Thus, relying solely on artifacts is insufficient to cope with the increasingly realistic synthetic data in the future.

2.4. Summary

The facial texture has a certain periodicity and continuity for a natural and "authentic" person. Conversely, the texture consistency of the fake images is destroyed due to the combination of facial regions from different people. At the same time, another proven fact is that there will be artifacts in synthetic images due to manipulation traces, and this observation is often used as the basis for deepfake detection. However, existing deepfake detection methods always consider only one at a time. Consequently, they ignore the other or directly utilize data-driven methods for binary classification, without considering the distinction between artifacts and textures. These methods cannot be generalized to various forgeries as demonstrated by the insufficient cross-domain performance (Li and Lyu, 2019a; Yu et al., 2021; Zhang, 2022).

Accordingly, this paper proposes a novel approach for deepfake detection based on textures and artifacts. We delineate a model that distinctly segregates textural from artifact information. By leveraging their complementary characteristics, we anticipate enhancing the model's generalization ability across various scenarios.

3. The proposed approach

3.1. Overview

Based on the observations in Section 2.3, it becomes apparent that both textures and artifacts play significant roles in detecting manipulations, each offering unique insights. Nevertheless, without a clear demarcation between the two, challenges arise. Specifically, emphasizing texture inconsistencies can be misleading when artifacts are present since they potentially undermine the texture's authentic representation. Conversely, when the focus is solely on artifacts for classification, it might be better to ignore texture information entirely (Sun et al., 2020).

We define I as an RGB image containing a face. The state of nature for I is partitioned into the set of authentic and manipulated images, denoted as real and fake, respectively. Thus, $I \in real, fake$. A boolean label y is associated with I to indicate its authenticity:

$$\begin{cases} y = 0, I \in real \\ y = 1, I \in fake \end{cases} \tag{1}$$

Wanting to determine if this face has been manipulated, we divide I into two parts to describe it more concretely and intuitively: the texture vector T and the artifact vector A, respectively. The interaction between these vectors within I is not explicitly known; therefore, we introduce a generic function $G(\cdot)$ to denote their combination:

$$I = G(T, A). \tag{2}$$

This approach allows us to model I without asserting a specific form of interplay between T and A. A running example of our proposed solution is shown in Fig. 3.

First, a set of feature maps is extracted from the input image I through a simple sequence of convolutional layers, which extracts features, normalizes them to stabilize the learning process, and introduces non-linearities. This process is applied for capturing complex patterns within the data and results in feature maps with 64 channels. The idea that different channels obtained from applying a convolutional block capture different types of features is a well-established concept in the deep learning literature (Zeiler and Fergus, 2014).

For this reason, the proposed approach divides the feature maps into two separate sets and processes them independently to extract complementary information about texture inconsistencies and artifacts. The first 32 channels and the last 32 channels of the 64-channel output from the convolutional block are used to describe textures and artifacts separately. Inspired by the broader understanding of hierarchical feature processing in CNNs, this exploratory design choice is underpinned by the hypothesis that, within the diverse set of features captured by the 64 channels, specific groupings might naturally align with textural or artifact characteristics due to the varying nature of the filters.

As a second contribution, we implement a specifically designed, learnable soft mask module to sharpen the differentiation within the texture vector T, demarcating the foreground (the modified facial or target region) from the background (the source image context). Differentiating these fundamental components becomes significant in light of the marked inconsistencies seen at the fusion boundaries when two distinct images are merged, a characteristic widely observed in several face manipulation methods (Li et al., 2020a). In this context, the integration of adaptive learning facilitates meticulous differentiation. Furthermore, our method effectively addresses a broad problem: the possible lack of predefined masks in facial forgery datasets.

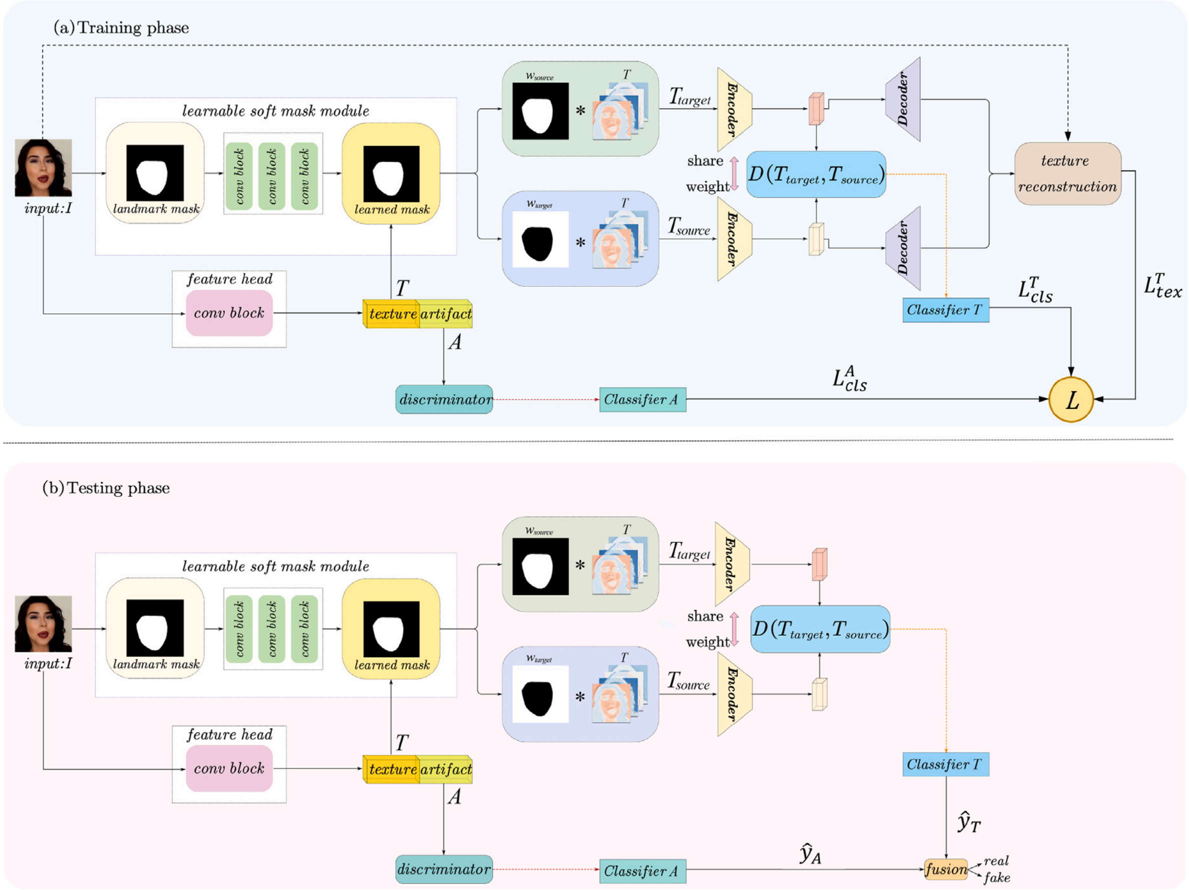


Fig. 4. Visualization of our proposed architecture. Figure (a) shows the training phase, and Figure (b) indicates the test phase pipeline. The input image is divided into two parts: texture and artifact; among them, the texture is further divided into two sub-parts: the foreground part, which comes from the target image, and the background part, which reserves the source image. Finally, the texture and artifact classification results are converted into probability values and fused to obtain the final inference result.

For instance, there are no given masks in the CelebDF dataset (Li et al., 2020b), which limits the application of supervised training. Some works have noticed this problem (Li et al., 2020a; Shiohara and Yamasaki, 2022); they used self-designed data sets with masks for training and achieved good results. However, redesigning and producing datasets with masks is inefficient and uneconomical for the general public. To this end, by leveraging a self-supervised strategy, our methodology reduces the reliance on masks within manipulated datasets, optimizing the data processing phase. The detailed exposition of the texture and artifact models, detailed in Fig. 4, is presented in the following sections.

3.2. Texture model construction and mask estimation

Starting from the findings of Li et al. (2020a), we suppose that the texture vector T of a given face image can be characterized as a linear combination of two primary components:

- T_{target}, the texture vector of the foreground face with the intended facial attributes (the “target” for manipulation);
- T_{source} the texture vector corresponding to the background or context of the image, serving as the source of those attributes.

This decomposition can be mathematically described as:

$$T = T_{target} + T_{source} = w_{target} \odot T + w_{source} \odot T \tag{3}$$

where the operation $\odot$ denotes an element-wise multiplication, while w_{target} and w_{source} are masks used to segment and emphasize their respective regions within the image. Intuitively, w_{target} preserves the face area intended for manipulation (face swap area), while w_{source} emphasizes the remaining area, retaining the image’s original or unaltered context.

They satisfy the following relationship:

$$w_{target} = 1 - w_{source} \tag{4}$$

For the construction of texture models, our goal is to connect the relationship between the T_{source} and T_{target} in Eq. (3) with real and fake data using unique constraints, which can effectively separate the natural image from the synthetic image. Therefore, we aim to separate the foreground from the background effectively.

As shown in Fig. 4, we generate an initial mask based on facial landmark detection points, perform adaptive learning on this basis to obtain a possible manipulated mask, and use it to divide the foreground (target image) and background (source image) with its complement. Subsequently, we employ two texture encoders to analyze these segmented regions. A parameter-sharing approach is adopted for computational efficiency, ensuring that the two encoders operate using the same set of parameters. Then, we compute the cosine similarity distance between the foreground and background textures in the feature space. This measure assesses the similarity or divergence between the two regions. In particular, the source and target images align closely for genuine input data, making T_{target} and T_{source} proximate in the feature space. In contrast, synthetic data yield a larger distance.

Based on the model above, we propose the corresponding novel consistency loss function, which aims to increase the inter-class distance between real and fake data:

$$D\left(T_{target}, T_{source}\right) = 0.5 * (1 - cos(T_{target}, T_{source})) \tag{5}$$

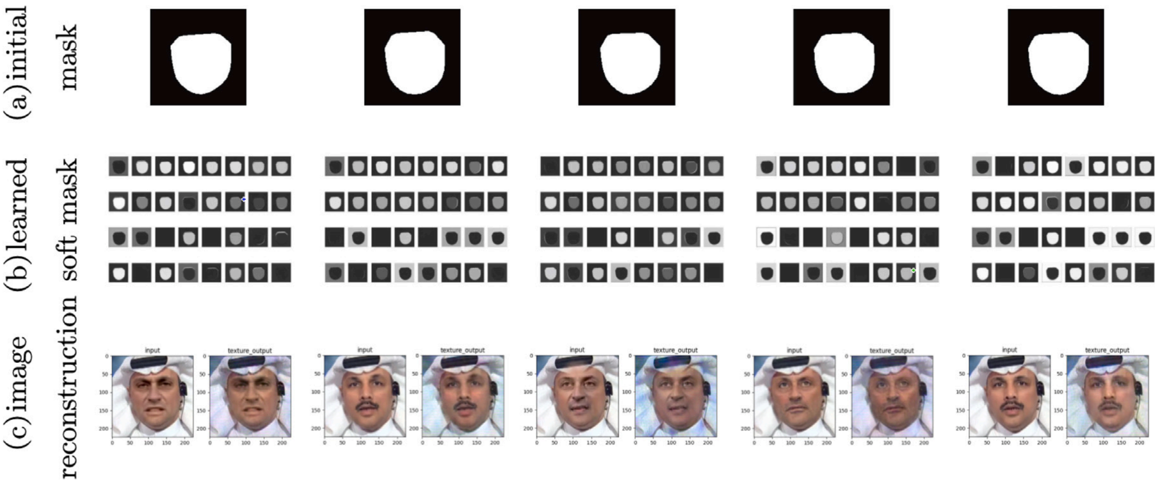


Fig. 5. Proposed soft mask learning process for image reconstruction and texture representation.

Given the operational range $[0,1]$ of our defined distance function D, we introduce a threshold parameter τ to define the classification criterion clearly. The predicted label $\hat{y}_T$ for generic input image I is defined by the following:

$$\hat{y}_T = D\left(T_{target}, T_{source}\right) < \tau \tag{6}$$

Therefore, according to Eq. (6), images with a texture distance falling below τ are categorized as real, while those exceeding τ are identified as manipulated. Considering that the value scope of τ is [0,1], we set $\tau = 0.5$ as a reasonable trade-off in our experiments. For final authenticity discrimination, we build a texture model classifier CLS_T as shown in Eq. (7).

$$L^{T}_{cls}(y, \hat{y}_T) = -y \log \hat{y}_T + (1-y)\log(1-\hat{y}_T) \tag{7}$$

This equation represents the binary cross-entropy (BCE) loss, a commonly used loss function for binary classification tasks. In this context, $\hat{y}_T$ denotes the classifier's prediction, while y indicates the actual authenticity label (ground truth).

The final step of our texture branch consists of two texture decoders reconstructing the input image's texture, operating with shared parameters. During this phase, the model precisely adjusts the learnable weights w_{target} and w_{source} to enhance the precision of the mask, ensuring a robust identification of potential texture inconsistencies. Additionally, this approach eliminates the need for ground truth masks in fake data, addressing a notable constraint in many datasets. Therefore, to reconstruct the texture information of the image, we adopt the style loss and perception loss; we also use pixel loss as a regularization term to speed up the convergence of the model, as shown in Eq. (8).

$$L_{tex}(I, \tilde{T}) = \sum_{l} \gamma_l \{ \| G_r(\phi_l(\tilde{T})) - G_r(\phi_l(I)) \| \} + \| \phi_l(\tilde{T}) - \phi_l(I) \| + 0.1 * MSE\left(I, \tilde{T}\right) \tag{8}$$

where $\tilde{T}$ refers to the reconstructed image texture and I refers to the input image. The term l represents the number of convolutional layers of the neural network, and G_r represents the Gram matrix, which is often used to measure the style loss function. The features extracted from the l_{th} convolutional layer are represented by ϕ_l, with γ_l being its respective weighting coefficient. Lastly, MSE denotes the Mean Squared Error loss.

In Fig. 5, we provide five examples of soft mask learning and image reconstruction on different types of manipulations. In particular, (a) shows the initial mask obtained by the face detector, (b) shows the soft masks obtained after the learnable soft mask module and (c) shows the fake input image and the corresponding reconstructed image. It is important to note that for each of the 32 representations obtained during the feature extraction phase, a different soft mask is obtained, which allows the texture to be extracted more precisely.

Finally, the total loss of the texture model is given by Eq. (9).

$$L_T = L^{T}_{cls}(y, \hat{y}_T) + L_{tex}(I, \tilde{T}) \tag{9}$$

3.3. Artifact model construction

In our framework, we hypothesize that the given input image might have undergone artifact modifications. These artifacts could arise from deliberate tampering or unintended alterations such as data compression. Given these artifacts' intricate nature and multifaceted components, an exhaustive decomposition becomes impractical. Instead, we opted for a streamlined approach by employing a straightforward discriminator to discern genuine images from tampered ones. Mirroring our approach for texture modeling, we design an artifact-specific classifier (Classifier "A" in Fig. 4). Classifier "A" is tailored to assess the presence of manipulation artifacts within an image. We denote the predicted probability that an image I is authentic by $\hat{y}_A$, as determined by Classifier "A". A score closer to zero indicates a higher probability that the image is authentic, whereas a score closer to one is indicative of manipulation. The established threshold for classification is set at 0.5. The training of Classifier "A" is still conducted by leveraging the BCE loss:

$$L_A = L^{A}_{cls}(y, \hat{y}_A) = -y \log \hat{y}_A + (1-y)\log(1-\hat{y}_A) \tag{10}$$

3.4. Texture and artifact detector

This paper aims to combine the strengths of texture and artifact paradigms into a composite framework.

The initial phase of this integration involves their respective loss functions to ensure that the system is fine-tuned to both textural variations and artifact-induced inconsistencies. Consequently, the aggregate loss is:

$$L = L_T + L_A = L^{T}_{cls}(y, \hat{y}_T) + L_{tex}(I, \tilde{T}) + L^{A}_{cls}(y, \hat{y}_A) \tag{11}$$

This holistic training approach ensures that no single component of the model operates in isolation. Instead, they collaboratively evolve to deliver an optimized detection capability.

The second phase involves the integration of the output of these models. Fusion methodologies have been instrumental in advancing pattern recognition, enhancing generalization capabilities, and addressing the challenges of intra-class variations and inter-class similarities. These fusion strategies can be incorporated at diverse stages of a

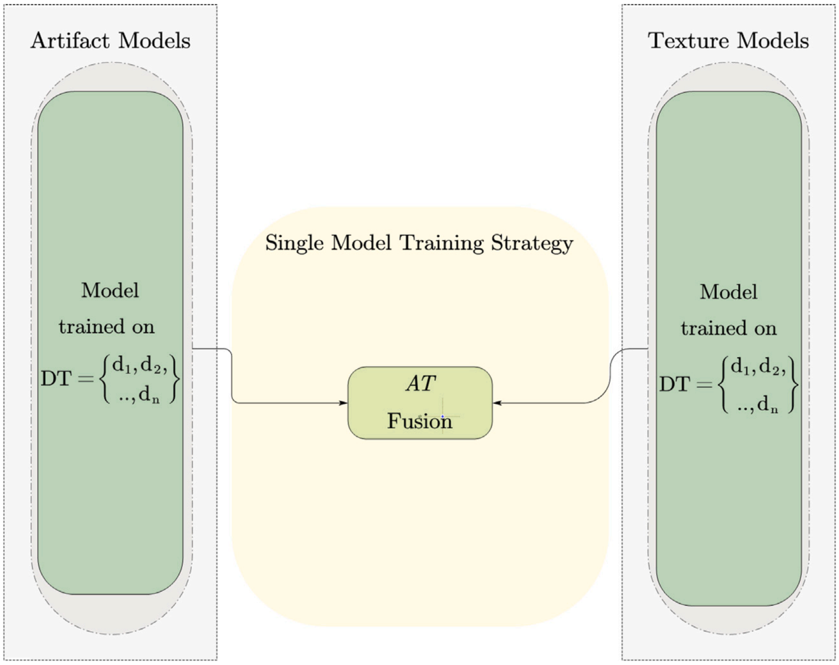

Artifact Models
Model
trained on
$DT = \{d_1, d_2, .., d_n\}$
Single Model Training Strategy
AT
Fusion
Texture Models
Model
trained on
$DT = \{d_1, d_2, .., d_n\}$

Fig. 6. Single Model Training Strategy: both the 'Artifact Model' and the 'Texture Model' are trained on a collective set of deepfake manipulations, denoted as $(D = \{d_1, d_2, \dots, d_n\})$.

classification system, from the sensor to the decision level. In the domain of deepfake detection, score-level fusion emerges as a promising technique (Tolosana et al., 2022; Bonettini et al., 2021; Concas et al., 2022a). It generates a unified, robust metric by combining the numerical outputs or "scores" from different models. Guided by this insight, our approach employs score-level fusion to combine the outputs from texture and artifact models.

We investigate fusion at the score level, as this allows exploiting the complementarity without increasing the system's complexity. Moreover, it offers multiple advantages. Among others, it permits exhaustive data utilization from models while retaining their inherent distinctiveness. To take full advantage of the complementarity between textures and artifacts, we propose two architectures: one lighter and simpler aimed at applications with limited computational resources and one more complex aimed at applications with a high need for generalization, such as the identification of never-before-seen manipulations.

3.4.1. Single-model training strategy

The first proposed architecture, shown in Fig. 6, is based on obtaining a single model for the texture part and one for the artifact part. This architecture allows the two pieces of information to be fused with low computational complexity and low memory storage, since, whatever the number of manipulations involved in the training phase, only two models are generated and used for inference. Starting from a DT training dataset composed of different manipulations $d_1, d_2, .., d_n$, two models are trained, one for the detection of texture inconsistencies, as explained in Section 3.2, and one for the detection of artifacts, as explained in Section 3.3. The outputs of these models are merged in Artifact-Texture (AT) fusion.

3.4.2. Ensemble training strategy

In the pursuit of improving the efficacy of deepfake detection, we identified potential limitations inherent to the single-model training approach. While it provided valuable insights, its isolated nature posed challenges regarding versatility and a comprehensive understanding of the various manipulations. To address these concerns, we present the ensemble-model training strategy (Fig. 7). Since different texture inconsistencies and different artifacts characterize each manipulation, this strategy aims to obtain specialized models on each known type of deepfake. For instance, if N manipulations are considered, N texture models and N artifact models are generated. We then combine all the texture models with a score-level fusion, called Texture fusion, and all the artifact models with the Artifact fusion. Subsequently, we combine the aggregate information of the textures and the aggregate information of the artifacts, performing a final AT fusion.

This strategy offers potential benefits in terms of versatility and generalization. The rationale behind this layered fusion approach is multifaceted. As highlighted in Fig. 7, it allows for incremental learning from new forgery types. This means that the fusion mechanism can accommodate more fake patterns beyond the ones considered in the initial training. Furthermore, by integrating outputs from diverse models, the ensemble method addresses the challenges posed by artifact and texture differences that single models might fail to capture comprehensively.

4. Experimental set-up

4.1. Training and testing protocol

4.1.1. Single-model vs ensemble training strategy

The experimental analysis of this work aims to evaluate the pros and cons of the TAD method with the two different training strategies, single-model and ensemble, in comparison with the state of the art. Although seemingly straightforward, the single-model strategy unveils crucial insights about how individual models, trained on specific types of manipulations, perform across various other manipulations. Furthermore, it elucidates aspects of the generalizability and adaptability of these fusion strategies on different manipulated content.

In the experiments on the single-model strategy, we evaluated both training on single manipulations and multiple manipulations. In the ensemble strategy, we trained on multiple manipulations to exploit the information of texture inconsistencies and the presence of artifacts to improve generalization capabilities. For both training strategies, we tested both in intra-dataset, i.e. using the same dataset but different sets for training and testing, and in cross-dataset, i.e. testing on a different dataset than the training set. This last test is the worst case in deepfake detection but is the most realistic, as the test videos have

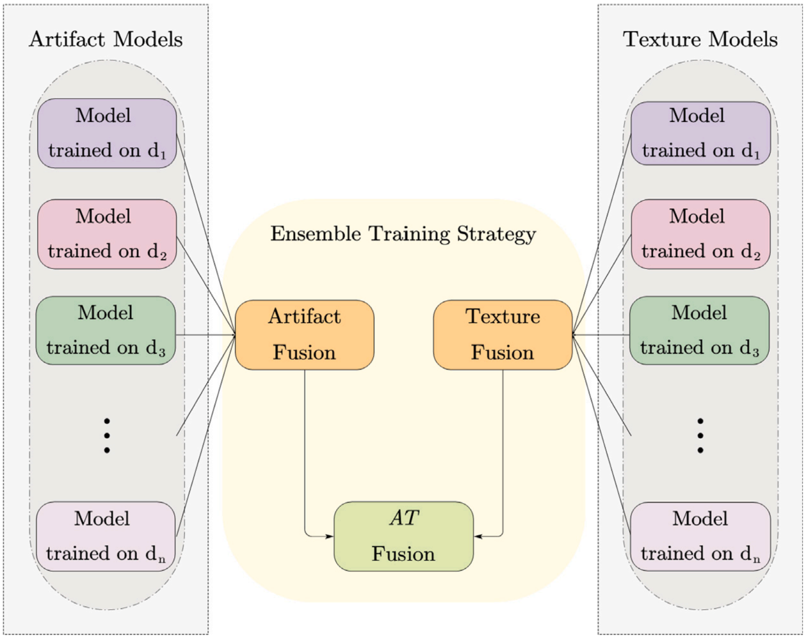


Fig. 7. Ensemble Training Strategy: Individual models within 'Artifact Models' and 'Texture Models' are trained individually on specific deepfake manipulations, denoted as $(d_1, d_2, \ldots, d_n)$. During inference, the outputs from each model are initially combined using 'Artifact Fusion' and 'Texture Fusion.' Subsequently, a final 'AT Fusion' is performed on the combined results.

unknown characteristics such as resolution, compression and types of manipulation.

The intention behind these experiments is twofold: first, to evaluate the adaptability of a model trained on one specific manipulation type when exposed to others; second, to identify and understand the potential vulnerabilities and strengths, paving the way for optimizing future fusion methodologies and offering a more comprehensive perspective.

Moreover, our evaluation strategy encompasses two primary modes: frame-level and video-level. In the frame-level evaluation, we assess each frame extracted from a video as an independent sample. In the video-level evaluation, we derive a combined score for the entire video by computing the average score across the considered frames for the video-level evaluation.

4.1.2. Ablation study for fusion rules

To evaluate which fusion rules are most suitable for the Artifact fusion and the Texture fusion in the ensemble-model strategy and for the AT fusion in both strategies, we analyzed five fusion rules (Concas et al., 2022a). In particular, we evaluated three non-parametric fusion rules, that is average, maximum, minimum, and two-parametric fusion rules by Multi-Layer Perceptron (MLP) and accuracy-based weighted average. Using the notation used in Concas et al. (2022a), where $P_i(\text{deepfake}|I)$ represents the score of the ith model on the same frame I we can compute the non-parametric fusions of N models with the following formulas:

$$P_{avg\_fusion}(\text{deepfake}|I) = \frac{1}{N}\sum_{i=1}^{N} P_i(\text{deepfake}|I) \tag{12}$$

$$P_{max\_fusion}(\text{deepfake}|I) = max_i(P_i(\text{deepfake}|I)) \tag{13}$$

$$P_{min\_fusion}(\text{deepfake}|I) = min_i(P_i(\text{deepfake}|I)) \tag{14}$$

The Multi-Layer Perceptron fusion takes as input the match scores generated by the multiple models to obtain more complex fusion rules according to the hidden and output layers (Concas et al., 2022a):

$$P_{MLP\_fusion}(\text{deepfake}|I) = sigmoid(\sum_{i=1}^{N} P_i(\text{deepfake}|I) \cdot w_i) \tag{15}$$

Where w_i are the weights estimated during the MLP training phase and $sigmoid(x) = 1/(1+e^{-x})$.

The accuracy-based fusion uses the accuracy obtained by each model on the training set as weight for linearly combining the scores assigned to each sample:

$$P_{acc\_weight\_fusion}(\text{deepfake}|I) = \frac{\sum_{i=1}^{N} P_i(\text{deepfake}|I) \cdot w_i}{\sum_{i=1}^{N} w_i} \tag{16}$$

4.2. Data sets

The following benchmark data sets are adopted:

- The **FaceForensics++** (Rossler et al., 2019) (FF++) dataset stands as a benchmark for facial forgery evaluation, consisting of 1000 authentic videos and 5000 forged videos produced using five advanced face manipulation techniques: DeepFakes[4] (DF), Face2Face (Thies et al., 2016) (F2F), FaceSwap (Korshunova et al., 2017) (FS), FaceShifter (Li et al., 2019) (FSh), and NeuralTexture (Thies et al., 2019) (NT). The FF++ dataset is categorized based on compression levels: the original (raw), lightly compressed (c23), and heavily compressed (c40) versions. In our pursuit of cross-domain generalization, the majority of our experiments were conducted using the c23 version to ensure wide applicability. We subsequently extended our testing to the more challenging c40 version to probe the limits of our network against severe compression artifacts typical of real-world video data.
For each forgery method within the FF++ dataset, we designated 750 videos for training, while the remaining 250 videos were reserved for testing. The data preparation involved utilizing the Multi-task Cascade Convolutional Neural Networks (MTCNN) for face detection and extraction, from which we derived 32 frames per video. This resulted in a training set encompassing 24,000 authentic images and 120,000 forged ones, distributed as 24,000 images for each forgery technique. The test set comprised 8000 genuine images alongside 40,000 counterfeit ones, with each manipulation method contributing 8000 images.

[4] https://github.com/deepfakes/faceswap (Last accessed: 06.11.2023).

In our study, we have intentionally chosen the FF++ dataset as the sole training set for all our deepfake detection models. This decision ensures that our experiments are both cross-dataset and cross-manipulation. All other datasets utilized in this research are exclusively for testing purposes. Moreover, beyond the training of our core models, it is important to highlight that, once trained, the scores derived from the models are utilized to set the Multi-Layer Perceptron (MLP) and accuracy-based fusion methods.

- **WildDF** (Zi et al., 2020) dataset consists of 7314 face sequences tailored to advance the creation of efficient real-world deepfake detectors. Videos in this dataset are uniquely challenging due to their manipulations using undisclosed techniques and their presentation against varied backgrounds. For our study, we exclusively extracted 300 authentic and 300 fake videos for testing. As with our other datasets, data processing involved using the MTCNN face detector to crop faces to extract 32 consecutive face frames from each video.
- **DFDC** (Dolhansky et al., 2019) dataset was introduced as part of the DeepFake Detection Challenge organized by Facebook (now known as META). The dataset is one of the largest publicly available collections for face-swapping videos, featuring over 100,000 clips from 3426 actors. The videos in this dataset were generated using a spectrum of techniques, from deepfakes to GANs and non-learned methods. Given the imbalance in the real versus fake video count, we selected an equal number from both categories to ensure sample balance. In alignment with the setup detailed in Otto (2020), we designated the last 15 compressed files out of 50 for testing. Due to computational considerations, additional data processing steps were undertaken. Utilizing the MTCNN face detector, we processed the data for cropping and refining. The final test set comprised approximately 2000 genuine and 2000 manipulated video clips, each contributing 32 facial frames.
- **CelebDF** (Li et al., 2020b) (Version 2) is an advanced dataset that encompasses 590 genuine videos and 5639 deepfake counterparts. The genuine videos are derived from public YouTube content featuring 59 celebrities from varied age groups, genders, and ethnic backgrounds. We have chosen the more comprehensive second version for our experimental assessments, with two iterations of this dataset available. We exclusively employ 190 real videos and an equivalent number of deepfake videos from this dataset for testing purposes. Consistent with the preceding datasets, we use MTCNN to extract 32 facial frames from each video, resulting in 12,160 testing images.

4.3. Baseline methods

In our experimental setup, we have chosen a set of diverse baseline methods for comparison with the proposed method. These methods were selected based on their prominence in the literature and performance in similar domains. We aim to provide a comprehensive benchmark against our proposed method by evaluating these architectures in our experiments. Below, we outline the key characteristics of each method and the rationale for its inclusion in our study.

- **XceptionNet**: A traditional CNN known for its depthwise separable convolutions with residual connections, trained initially on ImageNet (Chollet, 2017). Recognized for its robust performance, we trained it from scratch in our study.
- **EfficientNetB4** (Tan and Le, 2019): Chosen for its optimal balance of parameters, runtime, and classification performance.
- **EfficientNetB4Att** (Bonettini et al., 2021): This is a variant of the EfficientNetB4, inspired by attention mechanisms seen in natural language processing and computer vision. Through the attention mechanism, the network learns which part of its input is most relevant for its task, enhancing its ability to pinpoint the most informative portions of the input.
- **DSP-FWA**: An enhanced version of Li and Lyu (2019b). Utilizing a dual spatial pyramid strategy at both the image and feature levels, it capitalizes on the distinctive artifacts left by DeepFake algorithms due to resolution mismatches and transformations in the source video.
- **MCX-API** (Xu et al., 2023): A method that harnesses the strength of pairwise learning and diverse color space representations to ensure detection robustness across varied DeepFake generation techniques aimed at wide-scale employment.

4.4. Parameter settings

For a balanced comparison, we uniformly apply the same parameter settings across all methods in this study. We employ the stochastic gradient descent as optimization technique with a learning rate of 0.001 and momentum of 0.9. Each training batch is configured with a batch size of 4, and we train each model for 10 epochs. Our experimental setup utilizes the PyTorch framework in a Python environment, with experiments conducted on an NVIDIA GeForce RTX 2080 operating under the Windows system.

5. Experimental results

To demonstrate the effectiveness of the proposed method, we conduct comparative experiments to verify the classification performance in the case of intra-dataset and cross-dataset scenarios both for the single-model (S-TAD) and the ensemble (E-TAD) training strategy of the TAD system. The next sections report the results of each of these strategies and their comparison. Additionally, ablation studies are reported.

5.1. Single-model fusion experiments

This section presents the S-TAD frame-level experimental results using the FF++ dataset. This approach, comprehensively explained in Section 3.4.1, produces two dedicated Artifact and Texture models, each fine-tuned via end-to-end training. Subsequently, the results from the Artifact and Texture models are fused to provide a comprehensive assessment. For the sake of clarity, we focus on the MLP fusion strategy for presenting the results in this section. Preliminary tests have indicated that the MLP fusion outperforms other fusion strategies in the AT fusion of the S-TAD. For completeness and a holistic understanding, the results derived from other fusion strategies are detailed in the ablation study (Section 5.4).

Table 1 presents the accuracies across intra-manipulation, cross-manipulation, and cross-dataset scenarios for various models. Accuracy is defined as the percentage ratio between the number of correctly classified frames and the number of classified frames (decision threshold $\tau = 0.5$). While the EfficientNet family (B4 and B4ATT) of methods demonstrates optimal results in the intra-manipulation scenario, our S-TAD method's performance is still comparable with the other SOTA methods. Taking DF as an example, the best result is the **B4** model with an accuracy of 99.06%; our model's accuracy is 97.04%. However, S-TAD outperforms nearly all compared models on cross-manipulation tasks.

Regarding the response to varied data sources, the S-TAD method exhibits enhanced cross-dataset generalization in several instances compared to the other techniques. This suggests that the fusion approach employed by S-TAD is notably effective in better generalization across diverse datasets. Furthermore, a clear correlation between the choice of training data and the detector's performance can be observed. For instance, models trained on the NT dataset often achieve superior results. This underlines the critical role of strategically selecting training data, accentuating its impact on the detector's adaptability and accuracy across multiple datasets. However, upon a comprehensive examination of the results (Cross Avg column in Table 1), our method consistently stands out, demonstrating a marked efficacy with an incremental advantage ranging from 4.76% to 8.60%.

Table 1
Generalizability of the proposed single-model TAD (S-TAD) compared with SOTA methods in terms of frame-level accuracy (%) on different manipulation techniques. Gray background indicates intra-manipulation results, and the others indicate cross-manipulation results (Deepfake (DF), Face2Face (F2F), FaceSwap (FS), FaceShifter (FSh) and NeuralTextures (NT)) and cross-dataset results (WildDF, CelebDF and DFDC). The best results are in bold. The last column indicates the average of the cross-manipulation and cross-dataset accuracy.

Training data	Method	Testing set									Cross Avg.
		FF++						Cross-dataset			
		DF	F2F	FS	FSh	NT	All	WildDF	CelebDF	DFDC	
DF	Xception	98.38	51.25	49.87	52.03	52.92	61.33	52.55	55.55	56.89	53.01
	DSP-FWA	89.79	51.73	52.03	54.94	53.17	60.28	55.19	56.76	53.75	53.94
	B4	**99.06**	52.83	49.77	54.53	55.69	62.89	58.14	62.80	**57.84**	55.94
	B4ATT	98.98	52.13	49.66	52.82	54.24	61.93	58.79	**63.31**	57.06	55.43
	MCX-API	95.37	53.34	48.82	53.61	54.56	62.11	55.28	57.96	53.69	53.89
	Our(S-TAD)	97.04	**61.13**	**63.58**	**65.06**	**70.88**	**64.91**	**61.47**	56.17	52.95	**61.61**
F2F	Xception	56.28	98.17	50.73	49.74	51.43	61.87	51.78	50.84	51.59	51.77
	DSP-FWA	55.70	82.98	51.63	50.76	51.80	58.59	52.42	55.25	50.43	52.57
	B4	58.98	**98.91**	51.28	50.02	51.75	62.83	50.32	51.91	50.67	52.13
	B4ATT	59.71	98.65	51.38	49.74	51.74	63.19	50.21	51.08	50.94	52.11
	MCX-API	57.09	90.08	50.76	49.89	50.76	60.64	52.35	51.17	51.40	51.92
	Our	**69.60**	96.04	**57.24**	**53.42**	**61.41**	**64.98**	**56.22**	**60.38**	**51.82**	**58.58**
FS	Xception	51.22	52.67	97.72	50.14	49.42	60.56	42.07	50.03	**54.53**	50.01
	DSP-FWA	**61.13**	51.52	70.56	**52.94**	50.59	58.13	55.98	54.34	50.95	53.92
	B4	50.65	53.16	98.03	50.27	50.09	60.89	48.78	50.05	53.21	50.89
	B4ATT	51.11	**54.53**	**98.33**	50.18	50.20	**61.32**	47.64	50.13	53.30	51.01
	MCX-API	50.44	50.33	94.35	49.58	48.11	58.88	47.79	49.84	52.68	49.82
	Our	51.61	52.71	93.34	47.90	**59.61**	55.49	**57.32**	**60.11**	52.80	**54.58**
FSh	Xception	51.51	48.73	49.16	96.06	51.18	58.36	43.15	53.03	50.65	49.63
	DSP-FWA	51.54	49.88	49.69	98.06	50.34	59.62	50.23	51.85	51.32	50.69
	B4	50.52	49.75	49.94	98.23	50.09	59.46	49.34	54.44	50.57	50.66
	B4ATT	51.23	49.71	**50.17**	**98.44**	50.31	59.81	50.60	55.72	50.46	51.17
	MCX-API	56.72	51.20	48.15	93.90	55.01	61.33	48.65	54.52	53.04	52.47
	Our	**67.99**	**52.68**	49.43	95.79	**58.43**	**61.75**	**61.15**	**57.66**	**53.93**	**57.32**
NT	Xception	67.06	61.21	48.46	54.81	89.50	64.24	59.54	62.37	53.25	58.10
	DSP-FWA	62.36	59.18	50.18	55.79	87.61	64.49	62.59	59.62	51.13	57.26
	B4	72.10	59.49	48.61	58.99	**93.80**	66.43	59.80	**64.02**	54.74	59.68
	B4ATT	67.99	60.09	48.36	57.55	93.61	64.49	58.26	62.78	54.23	58.47
	MCX-API	70.88	58.98	47.41	56.64	82.62	64.78	58.38	60.85	52.71	57.98
	Our	**82.63**	**63.39**	**62.49**	**65.81**	86.06	**67.20**	**67.09**	58.54	**54.83**	**64.97**

Table 2
Generalizability of the proposed ensemble TAD (E-TAD) compared with SOTA methods in terms of frame-level accuracy (%) on different manipulation techniques. Gray background indicates intra-dataset results Deepfake (DF), Face2Face (F2F), FaceSwap (FS), FaceShifter(FSh) and NeuralTextures (NT)), and the others indicate cross-dataset results (WildDF, CelebDF and DFDC). The best results are in bold. The last column indicates the average of the intra-dataset and cross-dataset accuracy. The SOTA models are trained on the entire FF++ dataset.

frame-level results	Test set										Avg.
Method	Intra-dataset (FF++)						Cross-dataset				
	DF	F2F	FS	FSh	NT	Intra Avg.	WildDF	CelebDF	DFDC	Cross Avg.	
Xception	83.73	81.87	79.91	77.99	79.08	80.52	58.03	66.35	57.23	60.54	73.02
DSP-FWA	70.48	66.86	60.21	69.14	60.15	65.37	53.39	57.11	55.02	55.17	61.54
B4	91.76	89.62	88.35	88.09	**87.14**	88.99	63.47	**73.75**	**59.30**	**65.51**	80.18
B4ATT	92.63	91.00	87.78	89.36	86.81	89.52	62.65	69.29	58.28	63.41	79.72
MCX-API	71.44	67.46	64.95	68.66	63.43	67.19	50.78	56.43	54.94	54.05	62.26
Our (E-TAD)	**97.11**	**96.04**	**91.58**	**95.24**	86.61	**93.32**	**64.12**	64.42	56.40	61.65	**81.44**

5.2. Ensemble fusion experiments

In light of advancing cross-domain generalization capabilities, encompassing cross-manipulation and cross-dataset domains, the exploration was extended beyond the single-mode fusion strategy. The underlying hypothesis investigated was the potential complementary nature of texture and artifact information across distinct fake modes. To address this, we designed an ensemble fusion strategy based on the five synthesis methods from the FF++ dataset, as described in Section 3.4.2.

Similar to the S-TAD analysis, we avoid reporting all the results on the different fusion approaches, which are detailed in Section 5.4.

This protocol trained all SOTA methods on the entire FF++ dataset to provide a comprehensive comparison, ensuring the same exposure to our model's forgery cues. This evaluation, using frame-level accuracy and AUC as evaluation metrics, is depicted in Table 2 and graphically represented in Figs. 8 to 13.

Within the intra-dataset context, experimental results indicate that our method now surpasses most of the benchmarked state-of-the-art techniques. The current findings are particularly significant in light of our earlier assessments, where our method did not consistently rank foremost in this scenario. Furthermore, our model maintains consistent average accuracy compared to the single-model approach, whereas other models display a more pronounced reduction in performance. This difference in outcomes can be traced to the distinct training strategies. While these models previously exhibited robustness when trained on specific forgery methods, a significant accuracy degradation occurs when trained comprehensively across all methods. The E-TAD model was, therefore, successful in learning the differences between manipulations by exploiting the information on texture inconsistencies and artifacts. This is evident from the analysis of the score distributions (Fig. 10): the E-TAD method allows obtaining fake-real distributions that are very distant from each other and easily separable.

This behavior is also evident in the cross-dataset results (Table 2 and Figs. 11 and 12). The E-TAD remains the top performer on the WildDF dataset, acknowledged as one of the most challenging due to its real-world, uncontrolled conditions. However, the E-TAD remains suboptimal in terms of performance on CelebDF and DFDC datasets. This is confirmed by the further analysis reported in Table 3 which, although with slightly different experimental protocols, reports a comparison of the AUC on CelebDF on a large number of SOTA methods

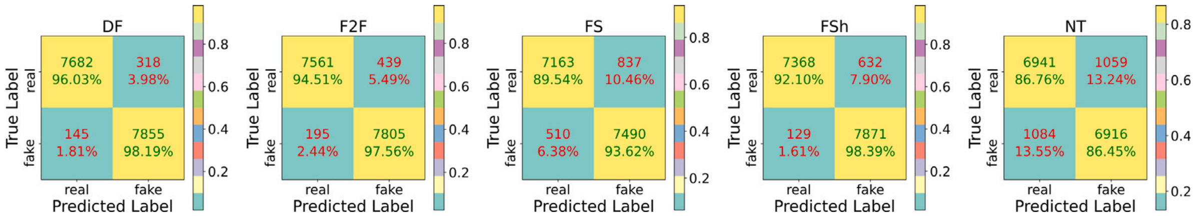


Fig. 8. Frame-level confusion matrix of our method (E-TAD) on the FF++ dataset (intra-dataset results). The decision threshold is set to 0.5.

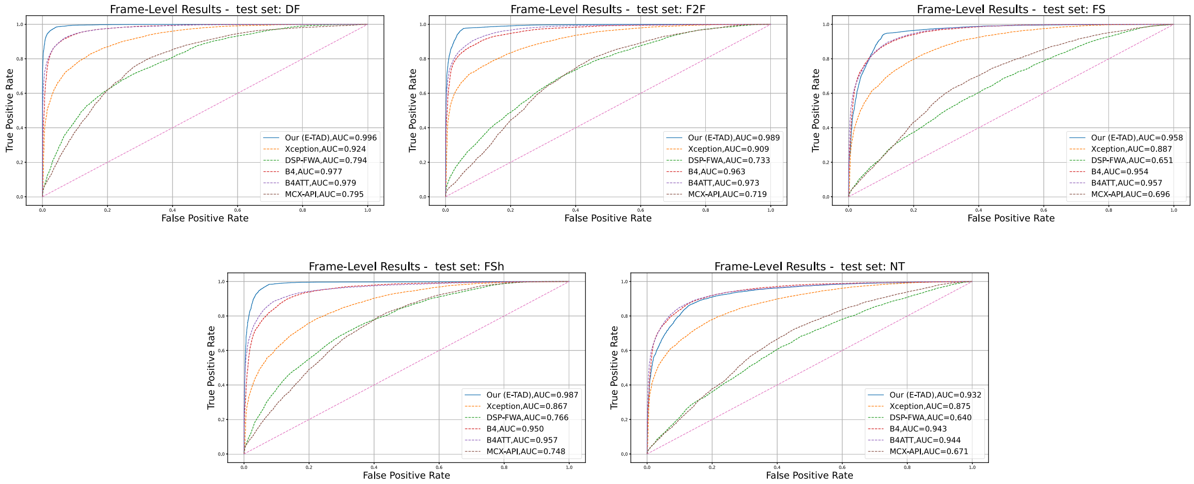


Fig. 9. Comparison of frame-level ROC curves of E-TAD and SOTA methods trained on the FF++ training set and tested on different manipulations of the FF++ test set (intra-dataset results).

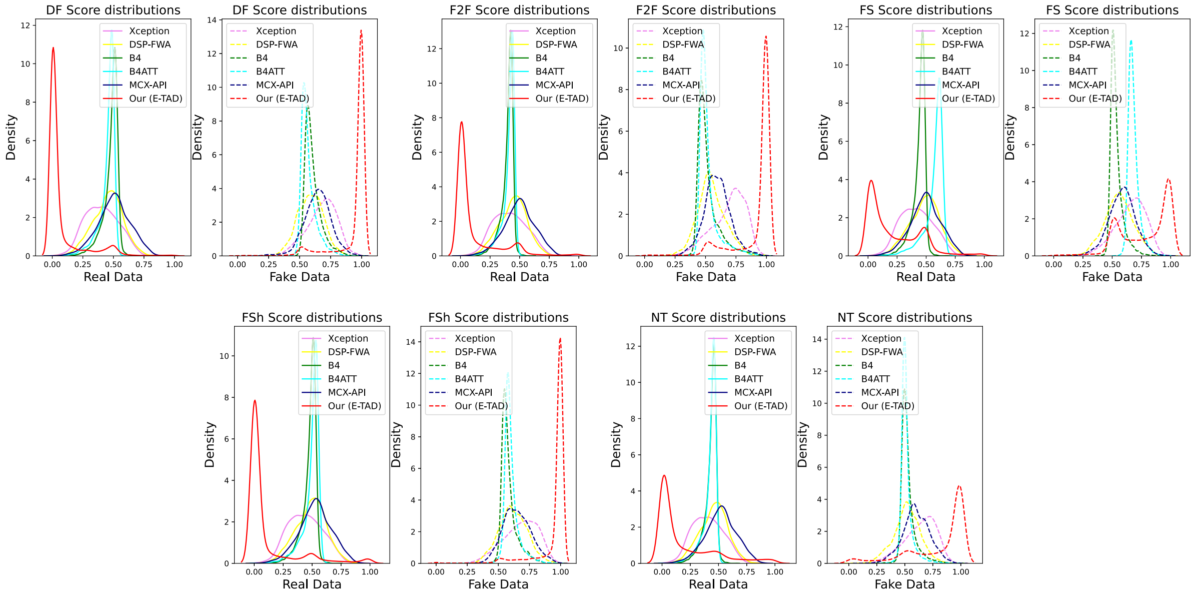


Fig. 10. Frame-level score distribution of our method (E-TAD) and SOTA methods on the FF++ dataset (intra-dataset results).

Table 3
AUC (%) frame-level cross-dataset evaluation on CelebDF of the proposed E-TAD method compared with other SOTA methods trained on the entire FF++ dataset. SOTA results are cited directly from (Zhao et al., 2021). These AUCs could be derived from different experimental protocols and should therefore be read as an indicative value.

Method	Two-stream (Zhou et al., 2017)	Meso4 (Afchar et al., 2018)	HeadPose (Yang et al., 2019)	FWA (Li and Lyu, 2019b)	VA-MLP (Matern et al., 2019)	VA-LogReg	Multi-task (Nguyen et al., 2019a)	Capsule (Nguyen et al., 2019b)	DSP-FWA (Li and Lyu, 2019b)	MAT (Zhao et al., 2021)	DCL (Sun et al., 2022)	E-TAD (proposed)
AUC on CelebDF	53.8	54.8	54.6	56.9	55.0	55.1	54.3	57.5	64.6	67.4	**82.3**	70.0

Table 4
Generalizability of the proposed ensemble TAD (E-TAD) compared with SOTA methods in terms of video-level accuracy (%) on different manipulation techniques. Gray background indicates intra-dataset results (Deepfake (DF), Face2Face (F2F), FaceSwap (FS), FaceShifter (FSh) and NeuralTextures (NT)), and the others indicate cross-dataset results (WildDF, CelebDF and DFDC). The best results are in bold. The last column indicates the average of the intra-dataset and cross-dataset accuracy. The SOTA models are trained on the entire FF++ dataset.

video-level results	Test set										
Method	Intra-dataset (FF++)						Cross-dataset				Avg.
	DF	F2F	FS	FSh	NT	Intra Avg.	WildDF	CelebDF	DFDC	Cross Avg.	
Xception	86.60	84.60	83.20	80.60	81.00	83.20	61.67	70.53	57.70	63.30	75.74
DSP-FWA	71.40	68.20	61.00	70.40	60.80	66.40	51.83	57.11	55.61	54.85	62.04
B4	94.20	92.20	91.60	91.60	90.20	92.00	65.17	**78.68**	**59.76**	**67.87**	82.93
B4ATT	95.20	92.80	90.20	92.40	90.00	92.10	64.00	72.63	58.56	65.06	81.97
MCX-API	73.20	70.40	67.00	70.40	65.20	69.20	51.17	55.53	55.31	54.00	63.53
Our (E-TAD)	**98.80**	**97.40**	**95.20**	**96.40**	**91.20**	**95.40**	**66.50**	68.95	57.10	64.18	**83.94**

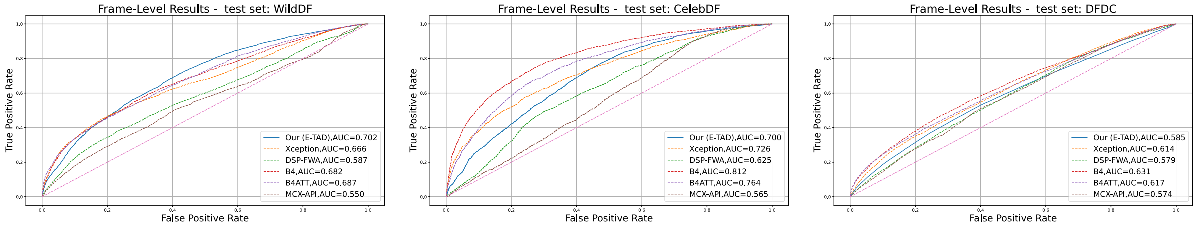


Fig. 11. Comparison of frame-level ROC curves of E-TAD and SOTA methods trained on the FF++ training set and tested on WildDF, CelebDF and DFDC datasets (cross-dataset results).

trained on the entire FF+ dataset. Upon detailed analysis, we discerned that the WildDF dataset, characterized by pronounced blur and intricate overlap of textures and artifacts, is particularly well-suited for our model. In contrast, the CelebDF dataset contains meticulously created, nearly imperceptible artifacts, while the DFDC dataset showcases more artifacts than texture variations. These distinct characteristics of each dataset might influence our model's performance, despite the enhancements achieved through the ensemble fusion strategy. Moreover, we primarily focus on the inconsistency inside and outside the texture and the presence or absence of artifacts without venturing into high-level semantic details. In contrast, methods like EfficientNet delve deep into advanced semantic information through block stacking and skip connections. This difference in approach might explain why our fusion performance does not match the detection prowess of a single model approach, especially on datasets like CelebDF and DFDC. Moreover, the distributions' analysis shown in Fig. 13 shows that for all the methods analyzed, there is an almost total overlap of the two fake and real distributions, demonstrating that the cross-dataset classification problem cannot be traced back only to the presence of different manipulations. In fact, if the problem was the classification of never-seen-before manipulations, the live accuracy would be similar to the intra-dataset protocol. The drop in accuracy between intra-dataset and cross-dataset results (Figs. 8 and 12) demonstrates that the limit of the proposed and other SOTA methods analyzed is related to not being able to generalize across different video formats. This limit will be analyzed in future work.

The video-level analysis provides further evidence of our model's robustness by expanding upon the frame-level results. While individual frames may present inherent variability in detection, the aggregation of performance over a video consistently demonstrates enhanced robustness and accuracy. As shown in Table 4, our E-TAD method confirms its competitive performance, achieving an enhanced average accuracy of 83.94%.

The comparison between the two training strategies of the proposed method, shown in Table 5, highlights the superiority of the ensemble strategy, which manages to exploit the difference in manipulations in texture inconsistencies and artifacts and obtain a deepfake detector capable of generalizing. This E-TAD framework's improved accuracy comes with a proportionate increase in inference time compared to S-TAD. However, We consider the reported increase reasonable in contexts where accurately identifying never-seen-before manipulations is crucial.

5.3. *Limitations of the proposed method*

The results presented in Table 6 provide a comprehensive view of the E-TAD network's capabilities in detecting deepfake content across various image conditions, highlighting its performance through true positive (TP) and true negative (TN) rates across different FF++ deepfake generation techniques (DF, F2F, FS, FSh, NT). The selections are tailored to assess the E-TAD network under real-world conditions. Training image quality (c23) establishes a baseline, while brightness adjustments (−50%, +50%) and the JPEG compression parameter "s" with settings at 80 for medium and 50 for low quality, simulate common scenarios such as varying lighting and digital compression effects. Moreover, including the heavily compressed c40 variant from the FF++ dataset provides a rigorous test of the network against typical compression artifacts encountered in video deepfakes. Examples of these manipulations are illustrated in Fig. 14.

When examining the effects of brightness alterations, the E-TAD network maintains generally high true positive rates, indicating robustness in identifying real images. However, under conditions of reduced brightness (−50%), the network's effectiveness diminishes notably, with the performance on F2F-generated deepfakes experiencing the most substantial decline. Similarly, the TN rates are affected by changes in brightness. The most pronounced decrease is observed under dim lighting conditions, with the F2F method being the most impacted again. This indicates that the network's capability to detect manipulated content is compromised in poorly lit environments. Conversely, the influence of increased brightness (+50%) is comparatively minor, suggesting that the key visual indicators essential for identifying deepfake content remain largely unaffected in conditions of enhanced illumination. The E-TAD demonstrates varying degrees of resilience when evaluating the impact of quality degradation, as evidenced by medium ($s = 80$) and low quality ($s = 50$) settings. The TP rates remain relatively stable across various quality levels, showcasing the network's consistent ability to recognize authentic images. However, a discernible decline in recognizing fake samples is observed as the image quality deteriorates, particularly at the lowest quality setting ($s = 50$). This trend is most pronounced for deepfakes generated by the NT method, where the network's effectiveness is notably compromised. While the TP rates are relatively unaffected by the change in quality for most generation methods, we observe a notable decrease in TN

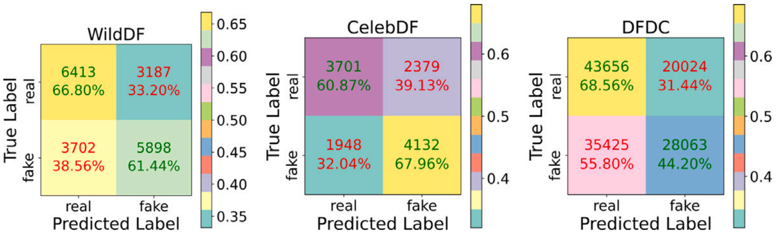


Fig. 12. Frame-level confusion matrix of our method (E-TAD) on the WildDF, CelebDF, and DFDC datasets (cross-dataset results). The decision threshold is set to 0.5.

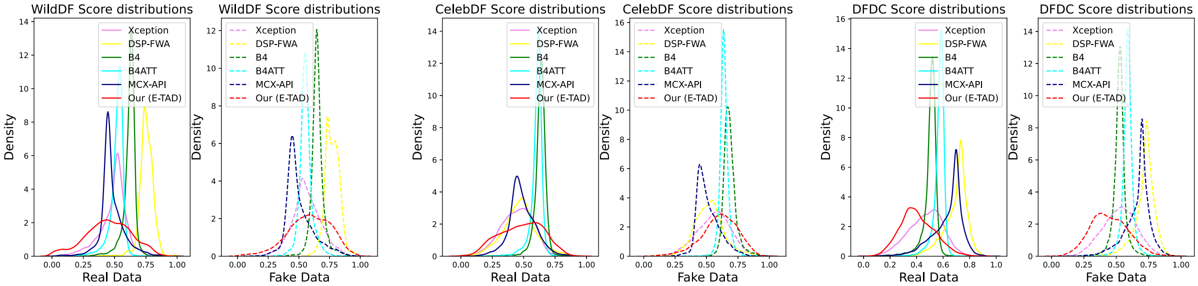


Fig. 13. Frame-level score distribution of our method (E-TAD) and SOTA methods on the WildDF, CelebDF, and DFDC datasets (cross-dataset results).

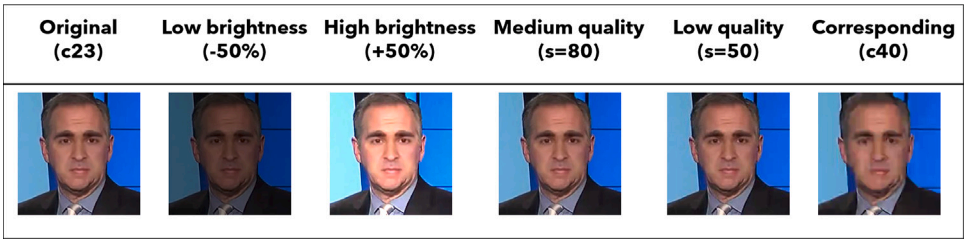


Fig. 14. Visual comparison of an individual's image processed under various conditions using the FF++ dataset. From left to right, the images represent the original quality (c23), the effects of a 50% reduction in brightness, a 50% increase in brightness, medium JPEG compression (s = 80), low JPEG compression (s = 50), and the heavily compressed version (c40) from the FF++ dataset.

Table 5
Comparison between performance and inference times of the two training strategies, single model (S-TAD) and ensemble (E-TAD), of the proposed system.

Method	Test set (Intra dataset accuracy)						Time (100 frames)
	DF	F2F	FS	FSh	NT	All FF++	
S-TAD	70.28	61.88	56.30	68.18	58.60	63.00	**25 s**
E-TAD	**97.11**	**96.04**	**91.58**	**95.24**	**86.61**	**86.14**	40 s

rates at the lowest quality setting, with NT showing a significant drop. In conclusion, the E-TAD network's efficacy in the c40 compression scenario within the FF++ dataset demonstrates a noticeable decline, particularly in its true negative (TN) rates. These rates experience substantial reductions, approximately between 17% and 47%, across different deepfake generation methods, highlighting the network's challenges in maintaining detection accuracy under heavily compressed conditions.

Future work should aim to enhance the network's resilience, possibly through advanced data augmentation techniques that introduce a wider array of lighting and compression scenarios during training. Additionally, exploring adaptive neural network architectures that dynamically adjust to varying image qualities may offer improved defense against sophisticated deepfake methods.

5.4. Ablation experiments

In evaluating the S-TAD and E-TAD frameworks, we systematically analyzed five fusion methodologies to identify the optimal integration approach for the outputs from both the texture and artifact models.

Table 7 presents the quantitative results for the single-model training strategy. Individual Artifact (A) and Texture (T) model performances vary across training and test sets. However, their combined output consistently outperforms individual results. This reinforces our premise about the inherent complementarity between texture and artifact models, enhancing detection rates and providing robustness against emerging forgery techniques. The MLP method demonstrates superior performance of the fusion techniques assessed, especially in intra-dataset evaluations. This can be attributed to MLP's parameter-based adaptive strategy, efficiently merging the outputs of the Texture and Artifact models to maximize detection accuracy. In particular, this method employs a two-layer architecture for both A and T fusions, where the first layer expands the inputs from two to 100 neurons, and the second layer compresses them into a singular output neuron, optimizing decision accuracy.

For the E-TAD framework, we separately evaluated the Texture fusion (Table 8), Artifact fusion (Table 9), and their combined or Artifact-Texture fusion (Table 10). In Tables 8 and 9, the efficacy of MLP as a fusion technique becomes evident, consistently outperforming all other methods across intra and cross-dataset evaluations. This trend

Table 6
Performance metrics of the E-TAD network evaluated on the FaceForensics++ dataset under varying conditions. The table reports true positive (TP) and true negative (TN) rates. Metrics are provided for baseline video quality (c23) and low quality (c40), decreased (−50%) and increased (+50%) brightness levels, medium (s = 80) and low (s = 50) image qualities.

Method	Original (c23)		Low brightness (−50%)		High brightness (+50%)		Medium quality (s = 80)		Low quality (s = 50)		Corresponding (c40)	
	TP	TN	TP	TN	TP	TN	TP	TN	TP	TN	TP	TN
DF	96.03	98.19	95.65	63.36	89.50	93.50	95.39	96.94	94.95	90.95	92.24	75.33
F2F	94.51	97.56	29.49	99.76	97.10	73.64	95.15	94.27	96.06	83.62	87.75	50.18
FS	89.54	93.62	78.35	77.00	89.50	84.16	91.42	85.20	92.20	65.81	59.82	62.00
FSh	92.10	98.39	73.86	94.90	92.25	88.01	91.22	98.15	87.05	95.38	86.58	81.44
NT	86.76	86.45	69.05	75.91	92.81	71.67	93.16	63.99	97.90	21.14	83.25	45.86

Table 7
Ablation study to evaluate the best Artifact-Texture (AT) fusion rule for the S-TAD method. Gray background indicates intra-manipulation results, and the others indicate cross-manipulation results (Deepfake (DF), Face2Face (F2F), FaceSwap (FS), FaceShifter (FSh) and NeuralTextures (NT)) and cross-dataset results (WildDF, CelebDF and DFDC). The best results are in bold.

Training data	**Fusion rule**	**Test set**								
		FF++						**Cross-dataset**		
		DF	**F2F**	**FS**	**FSh**	**NT**	**All**	**WildDF**	**CelebDF**	**DFDC**
DF	**A**	94.96	56.23	48.93	53.92	60.29	63.66	56.30	55.17	52.97
	T	93.80	52.06	48.77	57.18	61.29	62.42	60.53	51.27	51.44
	Avg	**97.06**	54.61	49.01	54.30	60.41	63.55	59.85	55.30	52.71
	Max	94.03	56.38	47.97	58.69	65.73	64.81	59.69	52.88	53.14
	Min	94.73	51.91	49.73	52.40	55.85	61.28	57.14	53.56	51.27
	MLP	97.04	**61.13**	**63.58**	**65.06**	**70.88**	**64.91**	**61.47**	**56.17**	**52.95**
	Acc-based	97.05	54.70	49.01	54.35	60.38	63.54	59.92	55.31	52.76
F2F	**A**	59.26	90.84	55.36	47.73	54.13	61.15	49.62	53.92	50.17
	T	59.41	92.34	51.13	52.02	55.08	61.94	**56.71**	**61.41**	**51.86**
	Avg	59.14	94.76	53.70	48.66	53.83	61.62	52.85	56.95	49.86
	Max	65.08	89.89	55.63	50.30	57.33	64.09	53.53	60.69	51.61
	Min	53.60	93.29	50.86	49.45	51.87	59.01	52.80	54.65	50.42
	MLP	**69.60**	**96.04**	**57.24**	**53.42**	**61.41**	**64.98**	56.22	60.38	51.82
	Acc-based	59.09	94.87	53.84	48.84	53.80	61.61	54.53	57.15	49.91
FS	**A**	47.09	52.69	92.88	47.24	46.16	57.25	45.46	53.27	49.27
	T	51.35	48.52	63.51	47.57	49.43	51.88	50.06	47.85	49.52
	Avg	47.94	50.78	93.34	47.54	47.44	**57.29**	52.00	51.01	49.96
	Max	49.15	50.56	83.93	45.76	46.46	55.36	46.49	50.62	48.69
	Min	49.29	52.67	72.46	**49.04**	49.12	53.77	49.04	50.50	50.11
	MLP	**51.61**	**52.71**	93.34	47.90	**59.61**	55.49	**57.32**	**60.11**	**52.80**
	Acc-based	48.41	50.86	**93.59**	47.58	47.48	57.26	51.88	50.81	49.89
FSh	**A**	53.31	49.74	48.78	94.02	51.49	58.79	49.02	58.13	52.40
	T	55.14	49.78	48.86	88.72	50.99	58.36	51.94	53.78	50.85
	Avg	52.83	49.79	48.98	**95.83**	50.84	58.73	53.81	57.06	51.17
	Max	57.02	49.92	48.26	95.44	52.32	60.10	50.01	**58.22**	52.70
	Min	51.43	49.61	49.38	87.29	50.16	57.04	50.94	53.69	50.55
	MLP	**67.99**	**52.68**	**49.43**	95.79	**58.43**	**61.75**	**61.15**	57.66	**53.93**
	Acc-based	52.71	49.79	48.99	95.78	50.84	58.73	53.80	57.36	51.34
NT	**A**	72.95	61.73	47.96	57.18	82.18	65.44	57.17	57.15	50.75
	T	80.90	54.29	44.06	62.36	82.38	64.75	65.88	56.74	51.17
	Avg	80.30	59.42	45.11	60.19	86.22	67.01	60.72	58.36	50.97
	Max	79.47	61.88	45.19	62.64	82.38	67.11	56.44	55.49	51.46
	Min	74.39	54.15	46.83	56.91	82.18	63.09	66.60	58.40	50.47
	MLP	**82.63**	**63.39**	**62.49**	**65.81**	86.06	**67.20**	67.09	**58.54**	**54.83**
	Acc-based	80.66	59.90	45.18	60.33	**86.24**	67.00	61.21	58.35	50.97

corroborates the earlier observation in the S-TAD framework and the strength of MLP in integrating diverse insights, whether they derive from texture or artifact clues. In this case, the first MLP layer scales the inputs from the five models to 100 neurons, and then condenses them into a single neuron for final analysis, mirroring the fusion approach employed for individual A and T models. Based on these assumptions, we investigated whether these MPL-fused scores can be combined further to improve the generalization ability of the whole model. As illustrated in Table 10, the AT fusion notably augments the model's adaptability. This is particularly pronounced for the "FS" manipulation within the intra-dataset evaluation. Despite the suboptimal performance of Texture Fusion for "FS", its combination with Artifact Fusion results in a marked improvement, further proving the complementary nature of these two attributes. Overall, the average-based fusion method now slightly outperforms the MLP approach.

The efficacy of parametric approaches is consistent with the suggested fusion strategy's goal: regardless of the training/validation data employed, the system aims to maximize generalizability and flexibility across different application situations. It is also important to point out that it is not feasible to generically identify what the fusion weights are or whether any models dominate the system's final decision.

Table 8
Ablation study to evaluate the best Texture fusion rule for the E-TAD method. Gray background indicates intra-dataset results, and the others indicate cross-dataset results. The best results are in bold.

Texture Fusion Rule	Test set							
	Intra-dataset (FF++)					Cross-dataset		
	DF	F2F	FS	FSh	NT	WildDF	CelebDF	DFDC
Avg	69.28	53.38	49.76	59.09	54.76	58.41	56.99	50.19
Max	77.43	76.02	57.47	73.69	72.01	61.10	54.40	52.85
Min	50.25	50.07	49.93	49.98	50.13	50.23	50.16	50.08
MLP	**93.52**	**92.34**	**66.85**	**91.80**	**82.51**	**65.21**	**59.36**	**55.37**
Acc-based	79.66	60.71	50.08	65.06	61.46	60.36	57.65	50.20

Table 9
Ablation study to evaluate the best Artifact fusion rule for the E-TAD method. Gray background indicates intra-dataset results, and the others indicate cross-dataset results. The best results are in bold.

Artifact Fusion rule	Test set							
	Intra-dataset (FF++)					Cross-dataset		
	DF	F2F	FS	FSh	NT	WildDF	CelebDF	DFDC
Avg	62.18	56.35	51.59	53.92	53.56	54.42	58.36	51.32
Max	78.71	77.94	78.71	77.70	75.10	51.15	55.43	51.76
Min	50.24	50.08	50.00	50.00	50.06	50.51	50.04	50.31
MLP	**95.18**	**90.48**	**93.24**	**94.57**	**83.04**	**59.49**	**64.11**	**54.50**
Acc-based	75.42	64.54	56.68	60.23	56.70	54.67	58.63	51.38

Table 10
Ablation study to evaluate the best Artifact-Texture fusion rule for the E-TAD method. Gray background indicates intra-dataset results, and the others indicate cross-dataset results. The best results are in bold.

AT Fusion Rule	Test set								Avg.
	Intra-dataset (FF++)					Cross-dataset			
	DF	F2F	FS	FSh	NT	WildDF	CelebDF	DFDC	
Avg	**97.11**	**96.04**	91.58	**95.24**	**86.61**	64.12	**64.42**	56.40	**81.44**
Max	94.04	90.72	81.33	92.14	83.16	62.98	62.20	55.16	77.72
Min	94.66	92.10	78.77	94.23	82.38	61.72	61.27	54.71	77.48
MLP	96.36	94.46	**92.47**	95.11	86.29	64.07	64.11	**56.47**	81.17
Acc-based	94.80	93.44	73.06	93.06	83.96	**65.89**	61.95	56.28	77.81

6. Conclusions

In this paper, we mainly focus on the problem of insufficient generalization of DeepFake detection. We provide a dual-focused lens through which the authenticity of digital content can be evaluated by integrating artifact analysis and texture inspection. Specifically, we construct multiple loss functions to separate textures and artifacts effectively and finally use their probabilistic fusion to distinguish real from fake. In particular, through a self-supervised learning strategy, we obtained a mask estimation method to separate the background and foreground of each image and analyze texture inconsistencies in detail. This information is then merged with artifact detection to locate manipulations in the face. This helps to improve the model's performance across domains, which is a significant step towards robust cross-domain detection.

Furthermore, we formulate an incremental ensemble learning strategy, augmenting our model's inherent flexibility and scalability. This strategic augmentation enhances the model's prompt detection capabilities, concurrently ensuring sustained adaptability—an imperative quality within the dynamic and evolving terrain of DeepFake technology. The specialized ensemble learning training affords our model the capacity to weigh insights gleaned from distinct detectors judiciously, each honed through training on diverse manipulations.

Our experimental results affirm the effectiveness of our texture-artifact separation approach in DeepFake detection, underscoring its potential in this field. Despite the promising results, we acknowledge a persistent challenge: a noticeable reduction in accuracy when moving from intra-dataset to cross-dataset scenarios, highlighting a current limitation in generalizing, probably, across heterogeneous video formats. The E-TAD network also shows constraints in low-light conditions and with heavily compressed videos. Future work will aim to mitigate these issues by integrating data augmentation techniques that simulate a broader range of lighting and compression effects and by enriching the training dataset with more varied artifacts. This endeavor will enhance the network's generalization capabilities across different video formats, addressing the differences inhibiting cross-format generalization.

CRediT authorship contribution statement

Jie Gao: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. **Marco Micheletto:** Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing. **Giulia Orrù:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Formal analysis. **Sara Concas:** Investigation, Methodology, Writing – original draft. **Xiaoyi Feng:** Conceptualization, Supervision, Writing – review & editing. **Gian Luca Marcialis:** Conceptualization, Supervision, Writing – review & editing. **Fabio Roli:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work is partially supported by the China Scholarship Council (No. 202206290093), by SERICS (PE00000014) under the Italian Ministry of University and Research (MUR) National Recovery and Resilience Plan funded by the European Union - NextGenerationEU and by the European Union – Next Generation EU through the Italian Ministry of University and Research (MUR) within the National Recovery and Resilience Plan (PNRR) - Mission 4 "Education and Research" - C2/1.1 - Fund for Relevant Projects of National Interests (PRIN) - "BullyBuster 2 - the ongoing fight against bullying and cyberbullying with the help of artificial intelligence for the human wellbeing" (Proj. Code: P2022K39K8 - CUP: F53D23009240001).

References

Afchar, D., Nozick, V., Yamagishi, J., Echizen, I., 2018. Mesonet: a compact facial video forgery detection network. In: 2018 IEEE International Workshop on Information Forensics and Security. WIFS, IEEE, pp. 1–7.

Ahonen, T., Hadid, A., Pietikainen, M., 2006. Face description with local binary patterns: Application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 28 (12), 2037–2041.

Benavides, C., Villegas, J., Roman, G., Aviles, C., 2016. Face classification by local texture analisys through CBIR and SURF points. IEEE Lat. Am. Trans. 14 (5), 2418–2424. http://dx.doi.org/10.1109/TLA.2016.7530440.

Bond-Taylor, S., Leach, A., Long, Y., Willcocks, C.G., 2022. Deep generative modelling: A comparative review of VAEs, GANs, normalizing flows, energy-based and autoregressive models. IEEE Trans. Pattern Anal. Mach. Intell. 44 (11), 7327–7347. http://dx.doi.org/10.1109/TPAMI.2021.3116668.

Bonettini, N., Cannas, E.D., Mandelli, S., Bondi, L., Bestagini, P., Tubaro, S., 2021. Video face manipulation detection through ensemble of cnns. In: 2020 25th International Conference on Pattern Recognition. ICPR, IEEE, pp. 5012–5019.

Cai, Z., Ghosh, S., Stefanov, K., Dhall, A., Cai, J., Rezatofighi, H., Haffari, R., Hayat, M., 2023. MARLIN: Masked autoencoder for facial video representation LearnINg. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1493–1504.

Carlini, N., Farid, H., 2020. Evading deepfake-image detectors with white-and black-box attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 658–659.

Chen, T., Kumar, A., Nagarsheth, P., Sivaraman, G., Khoury, E., 2020. Generalization of audio deepfake detection. In: Odyssey. pp. 132–137.

Chen, L., Zhang, Y., Song, Y., Liu, L., Wang, J., 2022. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18710–18719.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1251–1258.

Concas, S., La Cava, S.M., Orrù, G., Cuccu, C., Gao, J., Feng, X., Marcialis, G.L., Roli, F., 2022a. Analysis of score-level fusion rules for deepfake detection. Appl. Sci. 12 (15), 7365.

Concas, S., Perelli, G., Marcialis, G.L., Puglisi, G., 2022b. Tensor-based deepfake detection in scaled and compressed images. In: 2022 IEEE International Conference on Image Processing. ICIP, IEEE, pp. 3121–3125.

Conti, E., Salvi, D., Borrelli, C., Hosler, B., Bestagini, P., Antonacci, F., Sarti, A., Stamm, M.C., Tubaro, S., 2022. Deepfake speech detection through emotion recognition: A semantic approach. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 8962–8966.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Ferrer, C.C., 2019. The deepfake detection challenge (dfdc) preview dataset. arXiv preprint arXiv:1910.08854.

Dong, S., Wang, J., Ji, R., Liang, J., Fan, H., Ge, Z., 2023. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3994–4004.

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

Gao, J., Concas, S., Orrù, G., Feng, X., Marcialis, G.L., Roli, F., 2023. Generalized deepfake detection algorithm based on inconsistency between inner and outer faces. In: International Conference on Image Analysis and Processing. Springer, pp. 343–355.

Gatys, L.A., Ecker, A.S., Bethge, M., 2017. Texture and art with deep neural networks. Curr. Opin. Neurobiol. 46, 178–186.

Gecer, B., Ploumpis, S., Kotsia, I., Zafeiriou, S., 2021. Fast-GANFIT: Generative adversarial network for high fidelity 3D face reconstruction.. IEEE Trans. Pattern Anal. Mach. Intell..

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W., 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y., 2014. Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014. December 8-13 2014, Montreal, Quebec, Canada, pp. 2672–2680, URL https://proceedings.neurips.cc/paper/2014/hash/5ca3e9b122f61f8f06494c97b1afccf3-Abstract.html.

Haliassos, A., Mira, R., Petridis, S., Pantic, M., 2022. Leveraging real talking faces via self-supervision for robust forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14950–14962.

Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M., 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5039–5049.

Knafo, G., Fried, O., 2022. FakeOut: Leveraging out-of-domain self-supervision for multi-modal video deepfake detection. arXiv preprint arXiv:2212.00773.

Korshunov, P., Marcel, S., 2019. Vulnerability assessment and detection of deepfake videos. In: 2019 International Conference on Biometrics. ICB, IEEE, pp. 1–6.

Korshunova, I., Shi, W., Dambre, J., Theis, L., 2017. Fast face-swap using convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3677–3685.

Kwon, M.J., Nam, S.H., Yu, I.J., Lee, H.K., Kim, C., 2022. Learning JPEG compression artifacts for image manipulation detection and localization. Int. J. Comput. Vis. 1–21.

La Cava, S.M., Orrù, G., Drahansky, M., Marcialis, G.L., Roli, F., 2023. 3D face reconstruction: The road to forensics. ACM Comput. Surv. 56 (3), http://dx.doi.org/10.1145/3625288.

Li, L., Bao, J., Yang, H., Chen, D., Wen, F., 2019. Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457.

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

Li, Y., Lyu, S., 2019a. Exposing deepfake videos by detecting face warping artifacts. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, IEEE.

Li, Y., Lyu, S., 2019b. Exposing DeepFake videos by detecting face warping artifacts. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. CVPRW.

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 2020b. Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3207–3216.

Liang, J., Zeng, H., Zhang, L., 2022. Details or artifacts: A locally discriminative learning approach to realistic image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5657–5666.

Liu, Z., Qi, X., Torr, P.H., 2020. Global texture enhancement for fake face detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8060–8069.

Liy, C.M., InIctuOculi, L., 2018. Exposingaicreated fakevideosbydetectingeyeblinking. In: Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security. WIFS, Hong Kong, China, pp. 11–13.

Matern, F., Riess, C., Stamminger, M., 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops. WACVW, IEEE, pp. 83–92.

Nguyen, H.H., Fang, F., Yamagishi, J., Echizen, I., 2019a. Multi-task learning for detecting and segmenting manipulated facial images and videos. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems. BTAS, IEEE, pp. 1–8.

Nguyen, X.H., Tran, T.S., Nguyen, K.D., Truong, D.T., et al., 2021. Learning spatio-temporal features to detect manipulated facial videos created by the deepfake techniques. Forensic Sci. Int.: Digit. Invest. 36, 301108.

Nguyen, H.H., Yamagishi, J., Echizen, I., 2019b. Capsule-forensics: Using capsule networks to detect forged images and videos. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 2307–2311.

Otto, C., 2020. Comparing the performance of deepfake detection methods on benchmark datasets.

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C.S., RP, L., Jiang, J., et al., 2020. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. arXiv preprint arXiv:2005.05535.

Rana, M.S., Nobi, M.N., Murali, B., Sung, A.H., 2022. Deepfake detection: A systematic literature review. IEEE Access 10, 25494–25513.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M., 2019. Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11.

Shiohara, K., Yamasaki, T., 2022. Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18720–18729.

Sun, Z., Han, Y., Hua, Z., Ruan, N., Jia, W., 2021. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3609–3618.

Sun, X., Wu, B., Chen, W., 2020. Identifying invariant texture violation for robust deepfake detection, CoRR abs/2012.10580. arXiv:2012.10580.

Sun, K., Yao, T., Chen, S., Ding, S., Li, J., Ji, R., 2022. Dual contrastive learning for general face forgery detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 2. pp. 2316–2324.

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. PMLR, pp. 6105–6114.

Thies, J., Zollhöfer, M., Nießner, M., 2019. Deferred neural rendering: Image synthesis using neural textures. ACM Trans. Graph. 38 (4), 1–12.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., 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. pp. 2387–2395.

Tolosana, R., Romero-Tapiador, S., Vera-Rodriguez, R., Gonzalez-Sosa, E., Fierrez, J., 2022. DeepFakes detection across generations: Analysis of facial regions, fusion, and performance evaluation. Eng. Appl. Artif. Intell. 110, 104673.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., Ortega-Garcia, J., 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. Inf. Fusion 64, 131–148.

Wang, Z., Guo, Y., Zuo, W., 2022. Deepfake forensics via an adversarial game. IEEE Trans. Image Process. 3541–3552.

Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A., 2020. CNN-generated images are surprisingly easy to spot... for now. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8695–8704.

Xu, Y., Raja, K., Verdoliva, L., Pedersen, M., 2023. Learning pairwise interaction for generalizable DeepFake detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 672–682.

Yang, X., Li, Y., Lyu, S., 2019. Exposing deep fakes using inconsistent head poses. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP, IEEE, pp. 8261–8265.

Yu, P., Xia, Z., Fei, J., Lu, Y., 2021. A survey on deepfake video detection. Iet Biom. 10 (6), 607–624.

Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (Eds.), Computer Vision – ECCV 2014. Springer International Publishing, Cham, pp. 818–833.

Zhang, T., 2022. Deepfake generation and detection, a survey. Multimedia Tools Appl. 81 (5), 6259–6276.

Zhang, X., Karaman, S., Chang, S.F., 2019. Detecting and simulating artifacts in gan fake images. In: 2019 IEEE International Workshop on Information Forensics and Security. WIFS, IEEE, pp. 1–6.

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

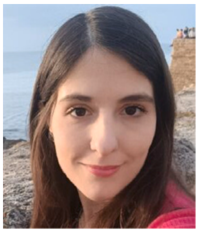
Zhou, P., Han, X., Morariu, V.I., Davis, L.S., 2017. Two-stream neural networks for tampered face detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, IEEE, pp. 1831–1839.

Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G., 2020. Wilddeepfake: A challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2382–2390.

Jie Gao received her M.S. degree from the School of Electronics and Information, Northwestern Polytechnical University, China, in 2020. After that, she continued pursuing her Ph.D. in the School of Electronics and Information at Northwestern Polytechnical University. She is currently conducting a doctoral joint training program at the Biometric Unit of the Pattern Recognition and Applications Laboratory (PRA Lab), University of Cagliari, Italy. Her current research interests include bioinformatics security and adversarial learning.

Marco Micheletto is currently a assistant professor with the Pattern Recognition and Applications Laboratory (PRA Lab), University of Cagliari. His research interests include integration of fingerprint comparison systems with presentation attack detector, fingerprint liveness detection, electroencephalography signal processing for biometric purposes and deepfake detection.

Giulia Orrù is currently assistant professor at the University of Cagliari (Italy), Pattern Recognition and Applications Laboratory (PRA Lab). Her research focuses on face recognition, fingerprint presentation attack detection, crowd analysis, adaptive biometric systems and deepfake detection. She acts as a referee for international journals and conferences on pattern recognition, biometrics and image processing.



Sara Concas received a B.Sc. Degree in Electrical and Electronic Engineering from University of Cagliari (Italy) in November 2018. Then she received her M.S. degree in Computer Engineering, Cybersecurity and Artificial Intelligence from the University of Cagliari on February 2021, with the thesis “Deepfake detection using quality measures”. Since 2020 she collaborates with the Pattern Recognition and Applications Laboratory (PRA Lab) and since October 2021 she is a PhD student.

Xiaoyi Feng received the B.S. and M.S. degrees from Northwest University, Xi’an, China, in 1991 and 1994, respectively, and the Ph.D. degree from Northwestern Polytechnical University, Xi’an, China, in 2001. From 2007 to 2008, she was a visiting scholar at the University of Oulu, Finland. She has been a Professor at the School of Electronics and Information, Northwestern Polytechnical University since 2008. She has authored or co-authored more than 100 papers in journals and conferences. Her current research interests include face analysis, affective computing, signal processing, adversarial learning, and computer vision.

Gian Luca Marcialis is currently an Associate Professor of computer engineering at the University of Cagliari (Italy) with national habilitation to full professorship. He is the Head of the Biometric Unit with the Pattern Recognition and Applications Laboratory leaded by Prof. Fabio Roli. His research interests include biometric-based personal recognition; in particular, fingerprint and face recognition, multi-modal fusion, self update algorithms, EEG-based features, and behavioral detection in crowds. He is an IAPR member. He acts as a referee for the main international journals and conferences on pattern recognition, biometrics, and image processing. He also acts as an external project referee for public and private institutions. He is the Chair of the International Fingerprint Liveness Detection Competition, aimed at assessing biennially the state of the art on fingerprint presentation attack detection (https://sites.unica.it/livdet/2023/).

Fabio Roli is currently a Full Professor of computer engineering with the University of Genova, Italy, and the Founding Director of the Pattern Recognition and Applications Laboratory with the University of Cagliari (https://pralab.diee.unica.it/). He is also a partner of the company Pluribus One that he co-founded (https://www.pluribus-one.it). He has been doing research on the design of pattern recognition and machine learning systems for 30 years. He has provided seminal contributions to the fields of multiple classifier systems and adversarial machine learning, and he has played a leading role in the establishment and advancement of these research themes. He is also a fellow of the International Association for Pattern Recognition. He was a recipient of the Pierre Devijver Award for his contributions to statistical pattern recognition.