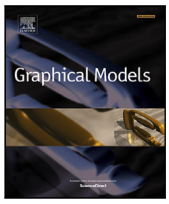
Graphical Models 138 (2025) 101255

Contents lists available at ScienceDirect

Graphical Models

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



3D data augmentation and dual-branch model for robust face forgery detection

Changshuang Zhou [a], Frederick W.B. Li [b], Chao Song [a], Dong Zheng [c], Bailin Yang [a,*]

[a] Zhejiang Gongshang University, China
[b] University of Durham, United Kingdom
[c] Zheng Dong Universal Ubiquitous Technology Co., LTD, China

ARTICLE INFO

Keywords:
Dual-branch network
3D data augmentation
Deepfake detection

ABSTRACT

We propose Dual-Branch Network (DBNet), a novel deepfake detection framework that addresses key limitations of existing works by jointly modeling 3D-temporal and fine-grained texture representations. Specifically, we aim to investigate how to (1) capture dynamic properties and spatial details in a unified model and (2) identify subtle inconsistencies beyond localized artifacts through temporally consistent modeling. To this end, DBNet extracts 3D landmarks from videos to construct temporal sequences for an RNN branch, while a Vision Transformer analyzes local patches. A Temporal Consistency-aware Loss is introduced to explicitly supervise the RNN. Additionally, a 3D generative model augments training data. Extensive experiments demonstrate our method achieves state-of-the-art performance on benchmarks, and ablation studies validate its effectiveness in generalizing to unseen data under various manipulations and compression.

1. Introduction

The emergence of generative models has enabled widespread generation of advanced deepfakes for malicious impersonation and the spread of misinformation. As deepfakes grow increasingly realistic and accessible, reliable detection is urgently needed. However, current forgery identification remains challenging due to limited labeled datasets and lack of unified representations to capture faces' rich multi-modal nature. Additionally, deepfakes exhibit subtle temporal inconsistencies that frame-level analysis fails to identify. To robustly address these pressing issues, we propose Dual-Branch Network (DBNet), a novel solution that leverages both 3D-temporal and fine-grained texture modeling in an end-to-end trainable framework to provide comprehensive spatial–temporal forgery detection against emerging deepfake threats.

Existing deepfake detection approaches can broadly be categorized into image-level methods analyzing individual frames through static analysis [1–12] and video-level techniques considering temporal features to model dynamics over sequences [13–22]. However, both categories exhibit limitations. While video-level approaches leverage sequences, they do not fully capture dynamics from intrinsic properties of 3D facial attributes. Furthermore, approaches extracting hand-crafted spatial or temporal features [23] lack flexibility, failing to generalize across manipulation techniques as attackers intentionally modify targeted cues. Specifically, prior work extracts iris color [23] but forgers can easily avoid this. There remains a need for data-driven multi-modal methods dynamically modeling appearance and dynamics through latent 3D representations to achieve robustness against a wide range of adversarial attacks.

Recent works have explored improving deepfake detection via new architectures like Vision Transformers leveraging self-supervision to capture long-range dependencies across frames [24,25], and RNN-based analysis of 2D landmark sequences extracted from videos [16,21]. While ViTs achieve high accuracy, they require large amounts of data and computations. Landmark-based methods are sensitive to variations that obscure keypoints. Inspired by the successful 3D decomposition in [19,20], it is important to encode facial structure and dynamics explicitly within a unified latent representation, rather than separately modeling appearance and motion. However, existing approaches still lack methods holistically leveraging both modalities through a specialized architecture optimized on intrinsic 3D facial attributes. This remains a problem in achieving robustness against diverse manipulations.

Meanwhile, existing methods aim to improve generalization through domain adaptation [26] and multi-task learning [8,24,27]. However,

* Corresponding author.
E-mail addresses: ybl@mail.zjgsu.edu.cn (C. Zhou), frederick.li@durham.ac.uk (F.W.B. Li), csong@zjsu.edu.cn (C. Song), zhengdong@uni-ubi.com (D. Zheng), ybl@mail.zjgsu.edu.cn (B. Yang).

https://doi.org/10.1016/j.gmod.2025.101255
Received 21 August 2024; Received in revised form 9 January 2025; Accepted 14 January 2025
Available online 4 February 2025

attributes are modeled independently without capturing intrinsic correlations between expressions, poses and identities through the underlying 3D facial structure. Liu et al. [12] focus exclusively on margins, risking errors with complex backgrounds. Data augmentation, as discussed in [6,7], typically involves static 2D manipulations, which may lack the necessary variability. Ideally, a data-driven paradigm could dynamically encode coupled attributes within a unified latent space, enabling self-supervised learning of constraints between variation factors through an explicit modeling of semantics. This may allow techniques to realistically extrapolate beyond training distributions without extensive data, improving generalization capabilities for addressing the gap in robust fake image detection.

We propose Dual-Branch Network (DBNet), a novel multi-modal framework to holistically model intrinsic 3D facial properties and address current limitations. Unlike prior work extracting landmarks and textures independently [19,21], DBNet leverages both within a unified latent 3D representation optimized through end-to-end training. Our RNN classifier encodes 3D shape dynamics rather than static landmarks to overcome limitations of face reconstruction under variations. Meanwhile, a Vision Transformer (ViT) with Patch Discriminator captures fine-grained textures at localized levels to complement global analysis. Leveraging FENeRF [28], our data-driven approach synthesizes training examples by directly editing latent codes representing attributes in a 3D face manifold. This enables infinite sample generation to model intrinsic correlations, optimized by our novel Temporal Consistency-aware Loss. We hypothesize that by synergistically combining modalities in this way, DBNet addresses gaps in robust detection. Through end-to-end training on 3D properties, we aim to validate our hypothesis and advance the field of forgery detection. Our main contributions are:

1. We leverage FENeRF to dynamically synthesize training data through disentangled editing of 3D facial identities, poses and expressions within a latent space, enhancing model robustness over static augmentation.
2. We propose Dual-Branch Network (DBNet), a novel multi-modal framework that synergistically leverages complementary 3D geometric and localized texture representations optimized through end-to-end training.
3. We introduce a specialized Temporal Consistency-aware Loss to supervise the RNN classifier, enhancing its ability to capture subtle dynamic patterns indicative of manipulations.
4. We conduct extensive ablation studies and evaluate on multiple datasets to validate the effectiveness, robustness and generalization of our data-driven spatiotemporal modeling approach against various attacks.

2. Related work

We provide context on deepfakes and existing detection strategies, originally stemming from generative facial synthesis. Deepfakes have advanced using increasingly sophisticated techniques. Current approaches include image-based and video-based methods, each with limitations in comprehensive spatiotemporal analysis. Neural radiance fields like FENeRF [28] enable disentangled editing of attributes within a latent space capturing underlying 3D facial semantics. This provides background on syntheses, generative modeling and prior work motivating our dual-branch DBNet framework for robust fake detection through holistic representation of intrinsic properties.

2.1. Deepfake generation

Early methods posed constraints like fixed pose [29], migrating faces across similar poses only. Three-dimensional representations addressed this by enabling reenactment across arbitrary poses [30–32], though realism suffered without corresponding regions [33]. Deepfakes beginning with FaceShifter [34] and GAN-based techniques [35,36] improved realism through self-supervised learning of intrinsic face properties and scene context. However, they remain constrained by limited, homogeneous training data. Uncommon expressions or poses from unseen identities challenge generalizability. For detection, generated faces may not be reliably distinguished based on appearance alone, as techniques capture photorealism across seen data. Detectors must therefore leverage higher-order inconsistencies like temporal dynamics poorly represented in restricted datasets. This implies challenges for existing approaches reliant on posed attributes.

2.2. Deepfake detection

Image-Level Deepfake Detection Early image-level approaches exploited spatial artifacts but suffered from overfitting and neglecting temporal cues [1,23,34,37,38]. More recent works aimed to learn generalized features across representations rather than focusing on specific manipulations [2,3]. While others emphasized modeling global relations in spectral domains to capture nonlocal inconsistencies [4,5].

Some methods localized manipulated regions through attention maps [39,40] or identity features [41] to inform decisions. More recently, techniques leveraged local perturbations [10] and cross-modal learning [11] to improve robustness against various fake qualities. Liu et al. [12] explored distributed systems to address dataset biases. However, image-level analysis remains limited due to neglecting important temporal dynamics between frames indicative of manipulations. Addressing this gap through modeling intrinsic spatiotemporal properties is crucial for robust deepfake detection.

Video-Level Deepfake Detection Early works started analyzing temporal cues in isolation [13–16] like mouth movements [22] or eyeblinks [42,43], lacking a holistic approach. Recent methods jointly modeled spatial–temporal relationships via inconsistency learning [44], 3D/Transformers encoding dependencies [45,46]. Some introduced attention/relational learning with dynamic masking [17] or dual-branch fusion [18]. Additionally, a few leveraged rich 3D structures through component decomposition/selection [19,20,47] or geometric feature enhancement [21].

However, analyzing modalities in isolation or focusing on localized artifacts limits generalization against complex manipulations. Most prior arts also constrain detection to specific manipulation types. The key gap remains in methodologies offering generalized detection through joint spatiotemporal modeling of intrinsic relationships without such constraints. Our work aims to address these challenges through a holistic learning framework.

Deepfake Detection with Vision Transformer ViTs demonstrated success across vision tasks [48–50]. Many early works applied ViTs to deepfake detection [45,51], achieving competitive performance but sacrificing efficiency. More recent methods leveraged ViTs' long-range modeling through two-branch architectures [5,52] or auxiliary tasks [24] to capture multi-domain inconsistencies.

However, each approach has limitations, such as only targeting identity [41] or disrupting frame continuity [53]. Bai et al. [24] required muscle annotations while Khormali et al. [54] ignored efficiency. UIA-ViT [55] and FA-ViT [25] also demand large datasets. Resizing frames in TALL-Swin [56] risks losing spatial details. This highlights the need for ViT-based methods balancing performance, generalizability and efficiency without constraints or additional data requirements. While achieving progress, gaps remain in holistically addressing aforementioned challenges through joint spatiotemporal modeling.

2.3. Data augmentation

Overfitting remains a challenge, where dropout helps but data augmentation plays a crucial role in enriching distributions. Early works applied basic 2D transformations [57,58] with limitations in generating diverse varieties beyond original datasets. Some explored temporal manipulations like dropout, repetition and blending [59] or random

erasing of frames/regions [58,60,61] for video representations. However, temporal modifications do not introduce new poses or identities. Generative models were also used, such as Transformers conditioned on noise [62] or reference images [27] to synthesize data. However, variability remained limited. 3D decomposition implicitly augmented data [20] but relied on reconstruction accuracy.

A key gap is the inability to dynamically synthesize infinite varieties of poses and identities beyond original samples. 3DMM-based reconstruction [63] generates rotated faces but depends on precise landmark detection. This highlights the need for a data-driven 3D generator suited for limited monocular face datasets. Inspired by advances in 3D face analysis [20], we aim to synthesize samples through 3D reconstruction. We leverage FENeRF [28], an implicit neural representation trained on images and maps. By editing attributes in its disentangled latent space, FENeRF supports unlimited generation without multi-view constraints, informing our augmentation approach.

Specifically, FENeRF [28] builds upon NeRF [64], which achieves high fidelity 3D reconstruction through attribute editing but relies on multi-view inputs requiring large datasets. FENeRF addresses this limitation with an implicit neural representation trained on paired images and maps. By directly modeling the 3D generator rather than fitting 3DMMs, FENeRF supports dynamic editing of attributes in latent space, enabling unlimited augmentation tailored for face datasets that primarily contain monocular views.

3. Method

We propose Dual-Branch Network (DBNet) for face forgery detection, as shown in Fig. 1. DBNet leverages a 3D generator pretrained with FENeRF to encode identity, pose, and expression in shape and texture latent codes z_s and z_t, respectively. These codes are utilized to reconstruct facial geometry and address overfitting through data augmentation. DBNet extracts landmark coordinate sequences A and difference sequences B over time from a suspicious video V. An RNN model $R(\cdot)$ analyzes the temporal dynamics by taking A and B as input to produce a probability $g = R(A, B)$ of suspiciousness. Concurrently, a Vision Transformer $VPD(\cdot)$ independently analyzes the spatial features from the entire video V to output another probability $v = VPD(V)$. The dedicated branch for extracting 3D facial landmarks, along with their spatial coordinates and temporal differences, is integral to DBNet's design, enhancing robustness. This architecture enables effective 3D data augmentation, providing additional insights. By leveraging comprehensive 3D representations through FENeRF, DBNet's dual-branch structure integrates 3D information with traditional cues, significantly improving forgery detection performance. This approach surpasses conventional methods that rely on basic data augmentations like Cutout or Erase, ensuring that the 3D data informs both the RNN and Vision Transformer components. Ultimately, the integration of the probabilities g and v through weighted fusion, characterized by $P = \theta \cdot g + (1 - \theta) \cdot v$, allows DBNet to leverage both temporal consistency and spatial understanding. This comprehensive assessment of spatial–temporal cues leads to robust detection through DBNet.

3.1. 3D editable face for data augmentation

We leverage the FENeRF technique [28] to generate editable 3D faces for data augmentation. FENeRF represents faces in a latent space that disentangles identity, pose, and expression attributes through a novel 3D-aware generator. Using two decoupled latent codes, it produces view-consistent and locally-editable portrait images with spatial-aligned 3D volumes and shared geometry. This enables programmatic editing of 3D facial geometry in the latent domain to synthesize additional training samples with diverse pose and identity configurations beyond the original dataset. To ensure temporal consistency in generated videos, we apply the same modifications uniformly across all frames of a video, maintaining the shape and identity to reflect the same person throughout. Different augmentation strategies, such as altering only expression or both shape and expression, are applied across videos to ensure dataset diversity. Unlike conventional 2D image manipulations, FENeRF's approach through intrinsic 3D structural modeling is crucial. Dynamically augmenting the data by editing 3D attributes through FENeRF facilitates improving the model's generalization against overfitting. This addresses the critical challenge of robustly detecting faces across various poses, lighting conditions, and identities not present in the original training distribution for forgery detection. Note that FENeRF-generated samples differ from deepfakes as they are designed for controlled augmentation of training data, improving detection capabilities by varying facial attributes. While these synthetic samples are technically 'fake', their purpose is to enhance model robustness, not to deceive. In contrast, deepfakes use advanced techniques to create misleading content for deceptive purposes.

FENeRF trains a generator producing 3D face models. It takes two disentangled latent codes - a texture code z_t for appearance and a shape code z_s for geometry. These disentangled codes are crucial for robustness, allowing changes to looks without impacting structure. Separating z_t and z_s enables FENeRF to edit traits independently through its disentangled editing. This facilitates generating extensive training data by varying the disentangled codes separately, improving the model's ability to detect manipulated faces across different poses and identities. The generator is formulated as:

$$G : (\hat{X}, d, z_s, z_t, e_{coord}) \rightarrow (\sigma, c, s). \tag{1}$$

where the input landmarks X are calibrated to $\hat{X}$ using Kalman filtering [65], while d and e_{coord} represent viewing direction and the learned positional feature. The outputs σ, c, s represent the density, color, and semantic fields capturing 3D structure. Specifically, σ is the viewpoint-independent density field, c is the color field defined by z_t, and s is the semantic field defined by z_s, enabling precise and independent manipulation of facial attributes such as identity and expression.

Once trained, the FENeRF generator can synthesize additional samples by editing the disentangled codes z_s and z_t to induce variations in attributes. This mitigates overfitting without disrupting 3D face geometry. Internally, it represents faces as density σ, color c, and semantic s fields. Crucially, z_s and z_t can be edited independently: modifying z_s alters shape while keeping texture via z_t constant, and vice versa. Stereo rendering reconstructs images and semantics by accumulating color $C(r)$, defined as the line integral of c weighted by σ and transmittance T in Eq. (2), and semantics $S(r)$ in Eq. (3) along camera rays. Using separate equations ensures changes to one attribute, like texture with z_t, do not inadvertently impact shape defined by z_s, preserving photo-realism.

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), d)dt \tag{2}$$

Similarly, equation (3) defines $S(r)$ as a line integral of semantic s weighted by σ and T.

$$S(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))s(r(t), d)dt \tag{3}$$

Transmittance T(t) captures density integration along the ray. In practice, we approximate Eqs. (2) and (3) in a discretized form following NeRF [64]. The generator trained on CelebAMask-HQ and FFHQ datasets is used to reconstruct FaceForensics++ (FF++) faces, generating additional data through disentangled shape/texture code interpolation. This data augmentation aids our method in generalizing to detect various manipulated images by expanding the training distribution while respecting geometric constraints. Fig. 3 visualizes example generated faces.

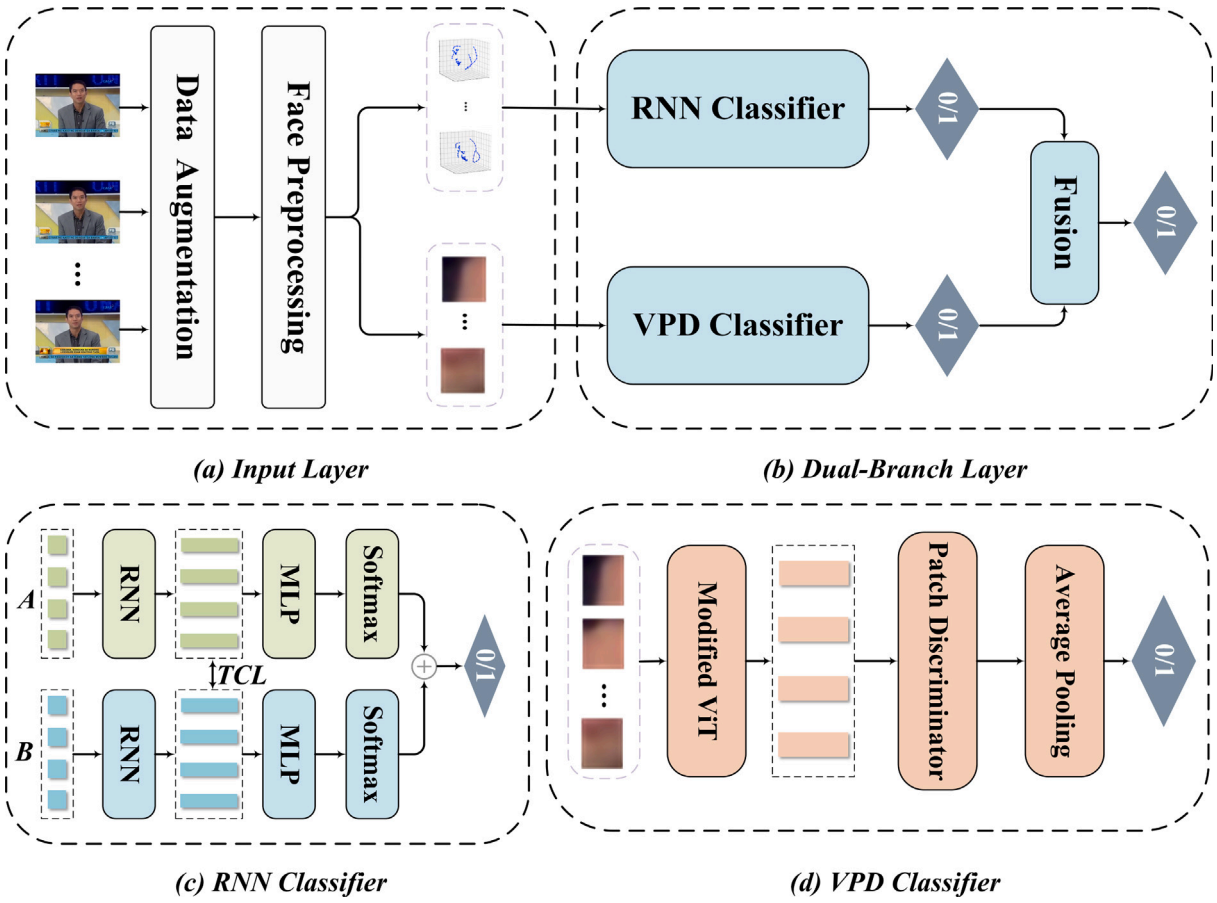


Fig. 1. Overview of the DBNet architecture. (a) The input video is processed to extract landmark coordinate sequences A and difference sequences B over time. (b) The Dual-Branch layer contains two distinct classifiers: an RNN classifier (c) analyzes the temporal dynamics by taking A and B as input, while a Vision Transformer classifier (d) analyzes the spatial features by taking image patches as input. Both classifiers output probabilities of suspiciousness, which are fused to classify the input video.

3.2. Detection

Our detection model leverages both temporal dynamics and spatial texture cues for comprehensive face forgery analysis. Alongside the RNN classifier that encodes temporal geometric sequences from 3D facial landmarks, a Vision Transformer with Patch Discriminator (VPD), i.e. ViT+PD, independently examines localized textures at a fine-grained level. The VPD operates by dividing the input facial images into small patches and processing them through a transformer architecture to capture subtle, spatially-distributed texture inconsistencies that may indicate manipulation. This fine-grained spatial analysis complements the RNN's temporal modeling, as texture anomalies can provide crucial evidence of face forgeries that may not be apparent from solely analyzing dynamic facial movements. The outputs from the RNN and VPD classifiers are then fused, enabling a holistic understanding that integrates both temporal and spatial perspectives for robust face forgery detection. This end-to-end trainable framework that jointly leverages 3D facial dynamics and localized texture representations aims to achieve comprehensive detection capabilities against diverse forgery attacks.

3.2.1. Face preprocessing

The preprocessing stage begins by segmenting the input video into D clips of length L frames. It then uses the DLIB face detector [66] to precisely extract the facial regions-of-interest (ROIs) from each frame by locating faces. The extracted ROIs are cropped and resized to serve as input to both the VPD classifier and the 3D landmark generation process. Obtaining accurate 3D facial landmarks is crucial for analyzing face geometric properties. While methods like 3DDFA [67] can directly estimate 3D landmarks from 2D images by leveraging 3D Morphable Models (3DMM) [63] and Principal Component Analysis (PCA), we opt for the FENeRF approach [28]. FENeRF utilizes a 3D spatial alignment to correlate the landmarks with corresponding facial semantics, preserving the high-frequency geometric details essential for comprehending the structural integrity of the face. Although this prioritizes accuracy over computational speed, the enhanced 3D representation provided by FENeRF is crucial for the subsequent spatio-temporal analysis in our face forgery detection framework.

The coordinates of the α-th landmark in the ith frame are represented as a 3D vector $p_i^\alpha = [x_i^\alpha, y_i^\alpha, z_i^\alpha]^T$, where $x_i^\alpha, y_i^\alpha, z_i^\alpha$ denote the Cartesian coordinates. This preprocessing extracts facial regions and identifies 3D landmarks to facilitate our detection of spatial and temporal inconsistencies through downstream RNN and transformer classifiers.

3.2.2. RNN classifier

To effectively capture the subtle temporal artifacts indicative of face manipulations, such as unnatural expressions and movements, we leverage 3D facial landmarks as a key input. The 3D landmarks provide a detailed geometric representation of the face, including shape, position, and depth information, offering a more comprehensive description of the facial dynamics compared to 2D landmarks. Recognizing the strength of Recurrent Neural Networks (RNNs) in modeling temporal sequences and dependencies in video data, we employ an RNN classifier that takes the 3D facial landmark sequences as input. By analyzing the temporal patterns and inconsistencies exhibited in these 3D geometric features, the RNN classifier is well-suited to identify the telltale signs of face manipulation, which often manifest as transient

and time-dependent artifacts introduced during the forgery process. This RNN-based approach allows our model to effectively detect these subtle temporal anomalies that frame-level analysis may miss, making it a crucial component in our holistic face forgery detection framework.

As in Fig. 1, the RNN classifier models and analyzes temporal sequences of 3D facial landmark coordinates detected from each frame. It extracts two input sequences — the Coordinate Sequence A representing facial movement patterns, and the Difference Sequence B representing speed patterns. These are fed into a combination of RNN modules (denoted as R_c and R_d) and MLP to estimate the likelihood of manipulation by detecting any abnormal movements or temporal discontinuities in landmark trajectories over the video. By capturing dynamics from 3D geometry over time, this component analyzes subtle inconsistencies to aid our detection of face forgeries from temporal cues in videos.

We utilize the 3D facial landmarks $p_i^\alpha = [x_i^\alpha, y_i^\alpha, z_i^\alpha]^T$ detected from each frame. To construct inputs, the landmarks sequence $L_i = [p_i^1, p_i^2, \dots, p_i^{n_p}]^T$, where n_p represents the number of landmarks, is used to generate the feature vector $\alpha_i \in \mathbb{R}^{3\times n_p}$ for each frame i through:

$$\alpha_i = [x_i^1, y_i^1, z_i^1, x_i^2, y_i^2, z_i^2, \dots, x_i^{n_p}, y_i^{n_p}, z_i^{n_p}]^T \tag{4}$$

The Coordinate Sequence $A \in \mathbb{R}^{n_l\times n_{sa}}$ is constructed by stacking α_i over all frames as:

$$A = [\alpha_1, \alpha_2, \dots, \alpha_{n_{sa}}]^T \tag{5}$$

where n_l denotes the number of coordinate values per frame and n_{sa} is the total number of frames. This sequence A is input to the RNN module R_c, producing the feature $F_A = R_c(A)$ representing facial dynamics. F_A is then fed to an MLP and softmax to compute the probability P_A of being fake through:

$$P_A = \text{Softmax}(MLP(F_A)) \tag{6}$$

The second feature vector β_i represents speed of facial movements between adjacent frames. Specifically, as in the following equation, β_i is obtained by computing coordinate differences between facial landmarks in frames i and $i+1$.

$$\beta_i = \alpha_{i+1} - \alpha_i \tag{7}$$

The Difference Sequence $B \in \mathbb{R}^{n_l\times n_{sb}}$ contains β_i over all frame pairs using Eq. (8), where n_{sb} is the total number of pairs.

$$B = [\beta_1, \beta_2, \dots, \beta_{n_{sb}}]^T \tag{8}$$

B is input to RNN module R_d, producing feature $F_B = R_d(B)$ modeling speed patterns to capture temporal discontinuities. F_B is fed to an MLP and softmax to compute probabilities P_B using:

$$P_B = \text{Softmax}(MLP(F_B)) \tag{9}$$

The final output g of our RNN classifier averages probabilities P_A and P_B, i.e., $g = (P_A+P_B)/2$, to jointly analyze coordinate and speed trajectories for robust forgery detection, integrating spatial dynamics from P_A, which reflects facial movement patterns, with temporal changes from P_B, capturing speed inconsistencies. This combined approach enhances the model's sensitivity to subtle manipulations that might be overlooked when examining each trajectory separately.

3.2.3. Temporal consistency-aware loss

Existing works typically use the binary cross-entropy loss for training, which disregards important temporal cues in video sequences by treating each frame independently. This limits generalization to new forgery types exhibiting subtle landmark trajectory changes over time. We propose a Temporal Consistency-aware Loss (TCL) to address this. TCL explicitly captures the dynamics of facial landmark sequences, a key indicator of forgeries, thereby improving the RNN classifier's performance. Unlike cross-entropy, it does not solely optimize for binary classification. TCL also accounts for intra-class variations to enhance the model's ability to detect unseen manipulation techniques by better adhering to the inherent properties of temporal face data. Through direct modeling of landmark motion patterns, our loss function enhances training to leverage temporal consistency as an important forgery cue.

Our proposed loss contains two main components: Coordinate Sequence Loss (CSL) and Difference Sequence Loss (DSL). The CSL, denoted as L_C, aims to capture subtle differences in facial coordinate patterns between genuine and fake sequences. It measures the MAE distance between the F_A features of real and fake videos.

The MAE is sensitive to even minor deviations in the coordinates, which is crucial for detecting the spatial changes that can occur in manipulated videos.

$$L_C = \text{MAE}(F_{A-\text{real}}, F_{A-\text{fake}}) \tag{10}$$

where $F_{A-\text{real}}$ and $F_{A-\text{fake}}$ refer to the F_A features of genuine and forged videos, respectively. CSL prompts the model to discern coordinate alterations induced by manipulation. Such spatial cues are important indicators of forgery artifacts. By focusing on the spatial arrangement of facial features, CSL helps the RNN to identify subtle spatial discrepancies that are indicative of manipulations.

Similarly, DSL, denoted as L_D, leverages MAE to quantify differences in F_B features representing landmark speed patterns.

$$L_D = \text{MAE}(F_{B-\text{real}}, F_{B-\text{fake}}) \tag{11}$$

Since facial movements are smooth and follow certain physiological constraints, any unnatural changes in speed or acceleration can be indicative of forgery. DSL complements CSL by additionally capturing temporal inconsistencies in speed dynamics, aiding accurate forgery detection. Together, these loss terms directly model key spatial and temporal characteristics to effectively train our model.

In addition to CSL and DSL, we introduce a Periodic Consistency Loss (PCL) denoted as $L_{periodic}$. This measures the consistency between landmark speed patterns of the current frame and neighboring frames using MAE.

$$L_{periodic} = \text{MAE}(F_{B-\text{cur}}, F_{B-\text{next}}) \tag{12}$$

where $F_{B-\text{cur}}$ and $F_{B-\text{next}}$ refer to the F_B features of the current and subsequent frames, respectively. PCL enhances robustness to temporal inconsistencies induced by manipulations by enforcing periodic consistency in landmark movements. Facial expressions and movements exhibit inherent rhythms and patterns like blinking that are too regular, too frequent, or even absent, which are not typically replicated in manipulated videos.

The overall Temporal Consistency-aware Loss L_{TC} combines CSL, DSL and PCL with weighting λ_1, λ_2 and λ_3:

$$L_{TC} = \lambda_1 L_C + \lambda_2 L_D + \lambda_3 L_{\text{periodic}} \tag{13}$$

The hyperparameters $\lambda_1 = 0.25$, $\lambda_2 = 0.25$, and $\lambda_3 = 0.5$ in the Temporal Consistency-aware Loss (TCL) were determined through empirical tuning to balance spatial (L_C), temporal (L_D), and periodic ($L_{periodic}$) features during training. The equal weighting of λ_1 and λ_2 emphasizes learning from spatial and temporal inconsistencies, enhancing model robustness, while the higher λ_3 prioritizes periodic consistency to detect subtle temporal anomalies indicative of manipulations. This strategic weighting ensures a comprehensive understanding of the dynamics in effective face forgery detection.

3.2.4. VPD classifier

While 3D landmarks effectively model facial dynamics, they alone are insufficient to discern textures indicative of manipulations. Traditional CNNs operating on full images also incur high computational costs. We propose utilizing a Vision Transformer with Patch Discriminator (VPD) classifier to complement our approach. The VPD brings a unique capability to analyze spatial hierarchies within image patches

as input, capturing both global and local features that are essential for understanding the context and details of facial expressions and textures.

By independently analyzing localized textures, the VPD extracts complementary static spatial representations to our RNN's modeling of dynamics. This dual approach ensures that both the temporal sequence of facial movements and the spatial details within specific moments are thoroughly examined, providing a more holistic understanding of potential manipulations. It aids detection by identifying subtle inconsistencies not discernible from landmarks or entire images alone. The joint model leverages both global motion patterns and fine-grained local cues for comprehensive and efficient forgery analysis.

The VPD classifier employs a Vision Transformer (ViT) for feature extraction from input image patches. Unlike traditional usage, our ViT omits the final MLP and softmax layers. In addition, we introduce a sparse input strategy where only two frames are sampled from each 2-second video clip. This increases efficiency while maintaining accuracy. Specifically, we choose the first frame of each second and divide them into non-overlapping patches $T \in \mathbb{R}^{n_p \times n_p}$. T is input to the ViT to obtain initial features $T' \in \mathbb{R}^{n_p \times n_p \times d}$, where d is the dimension of each patch feature. We downsample T' to $T' \in \mathbb{R}^{p \times d}$. The Patch Discriminator $PD(\cdot)$ is then applied to assign a manipulation probability $\hat{v}_i$ to each patch to capture the local inconsistencies that may indicate forgery.

$$\hat{v} = PD(T') = [\hat{v}_1, \dots \hat{v}_p] \tag{14}$$

Finally, average pooling integrates the patch predictions to generate the overall classification result v as follows. This integration of local patch analyses contributes to the overall detection by providing a detailed view of the spatial distribution of potential manipulations. This fine-grained analysis aids detection by modeling subtle local inconsistencies.

$$v = \frac{1}{p} \sum_{i=1}^{p} \hat{v}_i \tag{15}$$

To improve robustness and enable fine-grained detection, we implement a patch-level data augmentation technique denoted as $PL(\cdot)$. Given real/fake clips $D_{real/fake}$, two frames $I_i \in D$ are randomly selected and their KNN matches I_i' found from the corresponding datasets. The frames $\{I_1, I_2, I_1', I_2'\}$ are divided into patches T. Background patches are removed and proportions shuffled at 40% realism, yielding the blended patch set $M = PL(T) \in \mathbb{R}^{n_p \times n_p}$. M is input to our VPD classifier denoted as $VPD(\cdot)$. For the ith patch in the jth frame, its prediction is given by:

$$\hat{v}_{ij} = VPD(M) \tag{16}$$

We employ a binary cross-entropy loss L_b at the patch level for training.

$$L_b = -\frac{1}{N}\frac{1}{m}\sum_{i=1}^{N}\sum_{j=1}^{m}[v_{ij}\log(\hat{v}_{ij}) + (1 - v_{ij})\log(1 - \hat{v}_{ij})] \tag{17}$$

where v_{ij} is the ground truth label. This formulation enhances robustness to local artifacts through fine-grained modeling and balanced data augmentation.

3.2.5. Final prediction

Our detection model achieves the final prediction P through a weighted fusion of the individual outputs g and v from the RNN classifier and VPD classifier. This is formulated as:

$$P = \theta \cdot g + (1 - \theta) \cdot v \tag{18}$$

where g and v encode the temporal geometric cues and localized textures respectively. The learnable parameter θ determines the contribution of each classifier to P. This simple yet effective mechanism aggregates their complementary strengths. The RNN classifier captures dynamic patterns, while the VPD analyzes finer inconsistencies. Directly fusing features from these classifiers with a fully connected layer, as initially tested, did not yield satisfactory results due to the mismatch in feature dimensions and content. By optimizing θ, our model combines g and v to improve classification, facilitating robust detection by jointly leveraging 3D-temporal and texture modeling.

This integration generates a synergistic enhancement in forgery detection capabilities. The VPD supplies a distinct set of spatial features that augment the RNN's analytical process. For example, upon identifying a localized textural anomaly, the VPD can alert the RNN to focus its analysis on the dynamics within that area, searching for related movement irregularities. This collaborative mechanism also addresses VPD's challenge in detecting subtle dynamic changes over time, which are effectively identified by the RNN's temporal analysis.

4. Experiments

4.1. Datasets

To rigorously evaluate our method, we leverage four prominent public facial datasets. Comprehensive benchmarking against datasets simulating diverse attacks aids in designing detection techniques robust to emerging threats.

The Deepfake Detection Challenge dataset (DFDC) [68] contains 1133 authentic videos and 4080 synthetic videos generated using DeepFake, GAN-based and traditional techniques. Its scale and breadth of manipulations develop robust models. Celeb-DeepFake version 1 (CDF1) [69] comprises 408 genuine videos and 795 Deepfake videos. As an early dedicated benchmark, it established baseline performance measurements for the field. An extension, Celeb-DF version 2 (CDF2) [69] advances the challenge with 590 real and 5639 manipulated videos. Comparison to CDF1 gauges generalization against higher-quality Deepfakes. FaceForensic++ (FF++) [38] incorporates 1000 original videos and alterations using Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT) at three resolutions, namely original (raw), HD (c23), and LD(c40). It assesses manipulations and compression artifacts. For training, we utilize FF++'s diverse manipulations and degradations. CDF1 and CDF2 serve as hold-out tests to track progress versus advancing Deepfake technologies over time. Precise dataset definitions strengthen experimental reliability and real-world translatability.

4.2. Implementation details

Precisely documenting model configurations allows reproducibility and fair analysis. In pre-processing, L is set to 60 and video clips are cropped to 224×224 after resizing to standardize input size.

The RNN classifier adopts a stacked GRU architecture. Between the input and GRU, the dropout rate dr_1 is set to 0.25 based on prior works [16]. Additional dropout layers utilize dr_2=0.5 for regularization. The Adam optimizer uses a learning rate of 0.001 and batch size of 1024. Training occurs for 500 epochs to optimize sequence modeling.

In the VPD classifier, the Vision Transformer (ViT) divides inputs into 14×14 patches before projecting each to a 192-dimensional embedding space. The Patch Discriminator's Adam parameters mirror the RNN at 0.001 learning rate and 1×10^{-6} weight decay. Its batch normalization and LeakyReLU activation facilitate texture learning. Only 40 epochs suffice due to the shallow discriminator design.

4.3. Evaluation

4.3.1. Comparison with state-of-the-art

We compare our proposed method DBNet with other state-of-the-art approaches on multiple publicly available datasets for facial forgery detection. Specifically, we consider the FaceForensics++ (FF++) [38] dataset for in-distribution evaluation where models are trained and tested on the same dataset distribution. To test the generalization capability, we evaluate on other out-of-distribution datasets including

Table 1
Performance on In-Datasets and Cross-Datasets. Our models are trained on FaceForensics++ and tested on various datasets. Our model performs best on DFDC, CDF1, and CDF2.

Method	Year	in-set	cross-set		
		FF++	DFDC	CDF1	CDF2
Xception [38]	2019	94.86	69.70	62.30	65.50
F3-Net [2]	2020	98.10	67.45	63.57	68.69
Face X-ray [6]	2020	98.52	70.04	72.98	74.22
FTCN [45]	2021	99.73	74.01	–	75.58
LRNet [21]	2021	99.67	74.82	70.77	71.49
FD2Net [20]	2021	98.76	67.70	–	70.10
M2TR [5]	2022	97.84	69.94	68.57	69.94
E2E Learning [71]	2022	99.34	75.99	–	74.62
UIA-ViT [55]	2022	99.33	69.28	74.33	69.41
3DFS [47]	2022	97.19	70.57	–	–
FM-Net [19]	2023	98.70	72.35	–	72.04
F2Trans [4]	2023	99.74	70.39	76.29	77.61
CADDM [72]	2023	99.50	–	80.27	76.75
TALL-Swin [56]	2023	**99.87**	76.78	79.39	–
TI2Net [73]	2023	99.75	72.03	66.64	68.22
DBNet	Ours	99.28	**77.01**	**80.50**	**78.45**

Deepfake Detection Challenge (DFDC) [68], Celeb-DF v1 (CDF1) and Celeb-DF v2 (CDF2) [69], presenting a variety of challenges due to their unique data distributions and potential variations in forgery techniques. FF++ is one of the largest and most influential datasets containing over 1000 real and manipulated videos (spliced into frames). DFDC, CDF1 and CDF2 datasets contain manipulated videos generated by different techniques, providing more challenging cross-dataset evaluations. We utilize the AUC score to evaluate the predictive performance as it accounts for true positive and false positive rates in a balanced manner.

As in Table 1, our DBNet achieves competitive AUC performance compared to state-of-the-art methods on the FF++ in-distribution test set, demonstrating the effectiveness of jointly modeling dynamics and textures. Our AUC score on the FF++ dataset, while not the highest, reflects the influence of facial landmark detection accuracy on our model, as more precise landmark detection enhances feature extraction and classification quality. We currently utilize DLIB [70] for its efficient and accurate landmark detection capabilities, which also allow for easy implementation. As in Table 2, the performance of various models indicates a direct correlation between landmark accuracy and AUC scores, further confirming its critical role in our implementation. Improving this will be a focus in future work.

Our method's robustness is demonstrated by its leading performance in cross-dataset detection, effectively handling sophisticated and diverse forgeries. When testing on out-of-distribution datasets, DBNet consistently outperforms other approaches, obtaining AUC scores of 77.01% on DFDC, 80.50% on CDF1, and 78.45% on CDF2. This validates that leveraging 3D coordinates and localized textures enables our model to better extract discriminative features transferable to different data distributions. Its architecture leverages 3D coordinates and localized textures, enabling it to capture key features of facial forgeries that are consistent across different datasets. The model's exposure to a diverse range of examples during training, possibly through data augmentation and regularization techniques, could have contributed to its ability to generalize effectively. In summary, the superior cross-dataset generalization ability of DBNet underscores the robustness and efficacy of our approach for facial forgery detection. The dual-branch network design and specialized losses comprehensively exploit intrinsic properties of facial attributes for this challenging task.

4.3.2. Analysis of data augmentation using FENeRF

This section analyzes and compares the effectiveness of our proposed FENeRF data augmentation against Random-Erase and Face-Cutout. Random-Erase [60] repaints groups of pixels of different shapes on an image using face landmark information while Face-Cutout [58] further refines this approach by randomly selecting face landmarks and uses convex hull to obtain polygons, strategically choosing the largest mask based on pixel differences (Fig. 2). While they modify shape by masking landmarks, FENeRF reconstructs 3D faces conditioned on landmarks X, view d, and latent codes z_s, z_t representing shape and texture respectively as defined in Eq. (1). This enables editing of not just shape (via z_s) but also expression (via z_s) through the 3D representation. Additionally, poses can be modified by rotating 3D faces, infeasible with Face-Cutout (Fig. 3). Incorporating Random-Erase and Face-Cutout after generating FENeRF samples could potentially enhance model performance, and we plan to explore this combination in future work.

We assess facial manipulation detection accuracy using Area Under the ROC Curve (AUC) scores on the DFDC, CDF1, and CDF2 datasets, as reported in Table 3 for models trained on augmented Facial Forensics++ (FF++) images. While all tested augmentation techniques, including the 2D transformations from the Albumentations library that included techniques like RandomCutOut, RandomHorizontalFlip, and Rotate, demonstrated moderate improvements, our proposed Facial Expression Natural Augmentation with Representation Flow (FENeRF) achieved consistent and significant gains across all tested datasets and architectures. Notably, with 20% additional augmented samples, FENeRF provided AUC improvements of up to 5.5% over baseline methods like Xception, F3-NET, and FM-NET, highlighting its ability to generalize across different model architectures. The improvements

Table 2
AUC Scores across Facial Landmark Detection Models.

Method	AUC Score
OpenFace [74]	99.05
DLIB [70]	**99.28**
MGCNet [75]	99.34
3DDFA V2[67]	99.59
3DDFA V3[76]	99.72

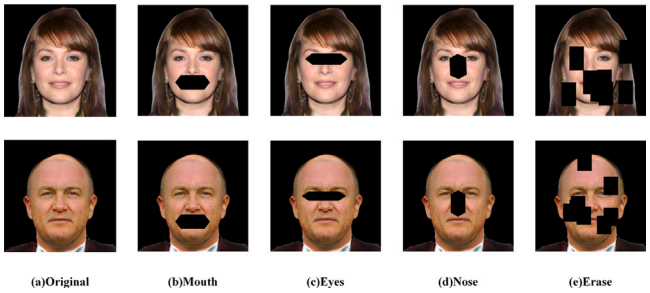


Fig. 2. Example of the Face-Cutout and Random-Erase method. (a) represents the original fake face, while (b), (c), and (d) represent the cutouts of the mouth, eyes, and nose respectively. (e) represents the Random-Erase method.

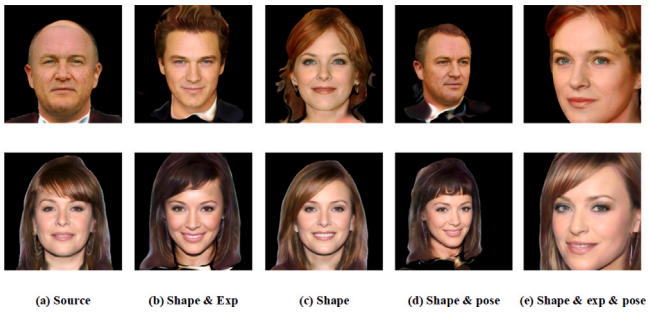


Fig. 3. FENeRF's work on the FF++ dataset. (a) represents the original unaugmented fake face with no data enhancement, (b) modified shape and expression information, (c) modified shape information, (d) modified shape and pose information, and (e) simultaneously modified shape, expression, and pose information.

Table 3
AUC scores (%) of different models under various data augmentation techniques tested on DFDC, CDF1, and CDF2 datasets.

Methods	Testing datasets		
	DFDC	CDF1	CDF2
Xception [38]	69.40	62.30	65.59
Xception+Erase [60]	71.93	65.08	68.21
Xception+Cutout [58]	72.11	66.24	68.87
Xception+FENeRF (10%)	**73.62**	**68.01**	**70.18**
Xception+FENeRF (20%)	**74.25**	**68.88**	**70.84**
F3-NET [2]	67.45	63.57	68.69
F3-NET+Erase [60]	71.08	65.99	71.78
F3-NET+Cutout [58]	70.92	66.48	72.12
F3-NET+FENeRF (10%)	**71.65**	**67.62**	**73.10**
F3-NET+FENeRF (20%)	**72.19**	**68.51**	**74.03**
FM-NET [19]	72.35	70.99	72.04
FM-NET+Erase [60]	73.96	72.15	74.01
FM-NET+Cutout [58]	73.44	73.00	74.22
FM-NET+FENeRF (10%)	**74.59**	**74.87**	**76.18**
FM-NET+FENeRF (20%)	**75.03**	**75.59**	**76.87**
Ours	72.55	75.21	78.45
Ours+Erase [60]	75.02	76.14	80.41
Ours+Cutout [58]	75.64	77.37	80.29
Ours+FENeRF (10%)	**76.97**	**80.17**	**81.15**
Ours+FENeRF (20%)	**77.01**	**80.50**	**81.57**

remain prominent in the more challenging CDF2 dataset, showcasing the effectiveness of FENeRF in handling diverse manipulation artifacts.

To systematically analyze the influence of augmented data quantity, we experimented with different proportions of FENeRF-augmented samples. A 10% augmentation ratio already yielded substantial gains, improving AUC scores by up to 4.8%, while a 20% ratio further strengthened detection capabilities across all models, maintaining the positive trend. This consistent improvement across DFDC, CDF1, and CDF2 highlights FENeRF's ability to generate training samples that better capture 3D facial subtleties, such as expression and pose variations, which are critical for enhancing manipulation detection. For example, FM-NET, which already achieved strong baseline results, exhibited marked improvements when augmented with FENeRF, with AUC scores increasing by up to 4.8% on the CDF2 dataset. These results indicate that FENeRF is not only effective but also scalable across different model architectures and datasets.

FENeRF's ability to disentangle and edit multiple facial attributes such as identity, expression, and pose in the latent space allows it to generate challenging yet realistic augmented samples, setting it apart from simpler techniques like Random-Erase and Cutout, which primarily modify shapes by masking landmarks. This capability leads to a more diverse and representative training set, significantly enhancing models' classification and generalization abilities. The strong performance of models like FM-NET and "Ours" across all datasets, including the challenging CDF2, further validates the effectiveness of FENeRF in improving robustness against diverse manipulation techniques. In conclusion, the results demonstrate that FENeRF does not only bridge the gap between 2D augmentation and realistic 3D transformations but also establishes a benchmark for improving manipulation detection through innovative data augmentation strategies.

4.3.3. Effect of 3D coordinates

We investigate the impact of utilizing 3D facial landmark representations in our approach. To objectively assess this, we remove the VPD classifier and directly compare performance using only 2D vs 3D coordinates as input to the RNN branch. Table 4 reports the AUC scores achieved on DFDC, CDF1, and CDF2 datasets. We observe an improvement of 1.15% on DFDC, 3.18% on CDF1, and 3.72% on CDF2 when 3D coordinates are employed instead of the conventional 2D coordinates. This validates the effectiveness of explicitly modeling the intrinsic 3D geometry captured by our landmark representation.

Table 4
AUC scores (%) of LRNet(2D) and our method (w/o VPD) tested on DFDC, CDF1, and CDF2 datasets. The best results are boldfaced.

Methods	Testing datasets		
	DFDC	CDF1	CDF2
LRNet(2D) [21]	74.82	75.21	71.49
Ours (w/o VPD)	**75.97**	**78.39**	**75.21**

Prior works such as LRNet [21] exclusively rely on 2D landmarks cropped from images, which lack crucial depth information retained in our 3D coordinates. By encoding variation across the third dimension, our approach provides richer cues benefiting the dynamic modeling task. The strong AUC scores observed in CDF1 and CDF2 further emphasize the robustness of our 3D representation in diverse and challenging scenarios.

The results highlight 3D coordinates as a favorable input for the RNN, empowering robust temporal sequence learning. In particular, the consistent performance across all datasets, including the challenging CDF2 dataset, demonstrate the adaptability and generalizability of our approach. In summary, the results conclusively demonstrate the value of leveraging 3D structural information over simple 2D landmarks. This enhances our framework's capability for comprehensive spatiotemporal analysis and forgery detection across multiple datasets.

4.3.4. Robustness to video compression

Given the prevalence of compression artifacts in real-world videos, we evaluate our approach's resilience against this source of variability. Fig. 4 reports AUC scores on the FF++ dataset compressed at levels c23 and c40, achieved by our model versus state-of-the-arts M2TR, TALL-Swin, F3Net, and FM-Net.

At the lightly compressed c23 level, TALL-Swin exhibits the best accuracy of 99.87%. However, as seen from the steeper performance drop at c40, it is significantly impacted by compression. In contrast, our method demonstrates the highest c40 AUC of 94.53%, dropping only 4.75% from c23. This validates our greater robustness to compression artifacts introduced by the encoding process. Our localized patch-level training helps filter out redundant high-frequency components while focusing on semantic-level manipulations. Specifically, the degradation value of our model is reduced from 5.36% to 4.75%. This imbues the ability to distinguish intrinsic variabilities from external distortions, enabling stable operation even under heavy compression. Other works relying on global features or isolated patterns remain comparatively susceptible. In summary, through quantitative comparisons, we verify our approach withstands quality degradation better than peers. This resilience stems from effective modeling at localized granularities, conferring an imperative trait for real-world applicability challenged by diverse levels of compression in video collections.

4.3.5. Robustness to different manipulation types

We evaluate our model's robustness across diverse manipulation techniques. Table 5 reports accuracy on each FaceForensics++ method when trained only on the remaining three. Our model achieves the highest average accuracy of 89.31%, outperforming Xception [38], F3-Net [2], MASDT [77], and FD2Net [20]. Notably, while FM-Net [19] attains the best performance on the FS dataset (83.67%), our method surpasses it on two out of the four manipulation methods, including the challenging NT dataset where we achieve 80.31%, marking a significant improvement over all competitors. This demonstrates the robustness imparted by our multi-modal representation.

Through end-to-end optimization of latent semantic cues, our approach is better equipped to capture subtle irregularities introduced across a wide spectrum of manipulations. This is further evidenced by our model's balanced performance across all four manipulation types, highlighting its generalization capabilities compared to prior works.

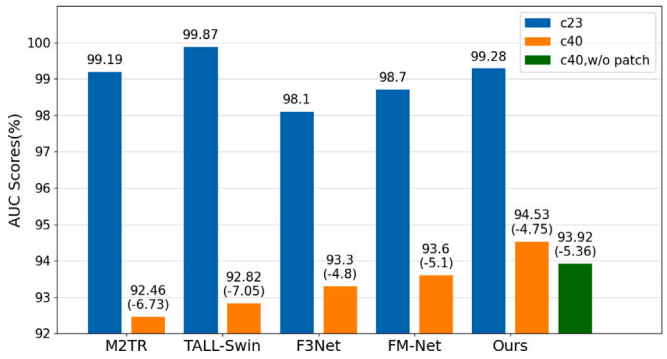


Fig. 4. AUC scores (%) of different methods under video compression. The orange and blue bars respectively denote the model's performance under c40 and c23 video compression levels. Additionally, the green bar represents an ablation study using the entire image input instead of patch-based input. Our model demonstrates the most robust performance, with the least degradation across these varying compression conditions.

Table 5
Quantitative results (ACC) on the FaceForensics++ (LQ) dataset with four manipulation methods: DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). The best results are boldfaced.

Method	Train on remaining three				
	DF	F2F	FS	NT	Avg
Xception [38]	98.56	90.17	63.44	69.98	80.54
F3-Net [2]	98.94	91.33	63.27	76.15	81.58
MASDT [77]	97.84	96.27	67.89	78.23	85.06
FD2Net [20]	98.51	89.01	68.60	71.11	81.81
FM-Net [19]	**99.45**	94.64	**83.67**	79.22	89.23
Ours	99.37	**96.38**	81.19	**80.31**	**89.31**

In contrast, methods like Xception and FD2Net show noticeable weaknesses on FS and NT, while MASDT achieves competitive performance on F2F but lags behind on other manipulations.

FENeRF augmentation and VPD analysis improve the model's ability to distinguish subtle differences by highlighting underlying inconsistencies. The inclusion of VPD enhances our model's robustness across diverse manipulations and ensures thorough analysis of augmented samples, contributing to superior performance. Table 4 clearly highlights the performance drop when VPD is excluded, reinforcing its contribution to improved detection capabilities, as shown in Table 1. While data augmentation enriches the training data, VPD ensures better identification of subtle irregularities.

In summary, the results reinforce that our method withstands diverse forgery techniques better than competitors. This robustness comes from its capacity to holistically comprehend intrinsic facial properties across multiple modalities and manipulation types, achieving superior performance consistently.

4.4. Ablation study

4.4.1. Effect of dual-branch network (DBNet)

We perform ablation studies to analyze our model. Table 6 reports results on FF++ c40 for variants using VPD classifier alone, RNN classifier alone, and the full Dual-BranchNet.

Interestingly, the VPD classifier achieves competitive 93.67% AUC and 92.58% accuracy, highlighting localized textures as potent cues. However, the RNN (93.99% AUC) surpasses VPD, validating 3D dynamics as superior. Crucially, fusing both modalities, our full model accomplishes balanced top results — 94.53% AUC and 93.20% accuracy. No isolated modality is comprehensive; only a joint representation achieves optimal generalizability. Notably, training time comparisons are illuminating — our VPD classifier completes in 10 h compared to 21 h for Xception [38], matching efficacy 50% faster through patch-focused efficiencies. Our RNN is even more economical at 3 h.

Table 6
Ablation study on the efficacy of our model variants on FF++ c40. We report the AUC (%) and ACC (%) to quantify the effectiveness of incorporating the RNN and VPD components in isolation and in tandem.

RNN	VPD	AUC	ACC
–	✓	93.67	92.58
✓	–	93.99	92.61
✓	✓	**94.53**	**93.20**

Table 7
AUC scores (%) of loss function variants (CSL, DSL, PCL) on FF++ c40.

BCE	CSL	DSL	PCL	AUC
✓	–	–	–	93.32
✓	✓	–	–	94.03
✓	–	✓	–	94.10
✓	–	–	✓	94.21
✓	✓	✓	–	94.35
✓	✓	–	✓	94.45
✓	–	✓	✓	94.50
✓	✓	✓	✓	**94.53**

In summary, the DBNet design proves favorable, leveraging complementary strengths through effective fusion outperforming any single specialized modality. This validates benefits of multimodal modeling while improving training efficiency.

4.4.2. Effect of temporal consistency-aware loss

We conduct ablation experiments to evaluate our loss function's effectiveness. Table 7 reports AUC for models trained with different BCE+TCL sub-loss combinations on FF++ c40. Inclusion of any individual sub-loss (CSL: spatial, DSL: temporal, PCL: periodic) improves performance by ∼ 1%, validating their distinct contributions. Combining multiple losses achieves further gains, with the full BCE+TCL achieving top 94.53% AUC – a significant 1.21% relative increase over BCE alone. This clearly demonstrates TCL's ability to comprehensively supervise the RNN, enhancing learned representations. By jointly modeling consistency across frames, it guides the model to focus on subtle dynamics indicative of manipulations rather than isolated artifacts.

Recognizing the limitations of standard classification losses in addressing the unique challenges of face forgery detection, our framework introduces a specialized Temporal Consistency-aware Loss (TCL). Prior works have shown that naively training a spatio-temporal network with a binary cross-entropy (BCE) loss can lead the model to rely on "easy" but unreliable manipulation artifacts, failing to uncover the full scope of forgery clues [59]. Furthermore, optimizing each training sample equally, as in the traditional BCE framework, makes it difficult to effectively capture the underlying temporal complexities in facial expressions, which is crucial for robust generalization [78]. In contrast, the TCL is designed with a deeper understanding that facial forgery detection transcends conventional classification — it requires a profound comprehension of the subtle, time-evolving patterns in genuine facial dynamics. By explicitly incorporating multifaceted consistency measures to constrain the RNN training process, the TCL loss function empowers the model to better exploit the temporal information embedded in the 3D landmark sequences, leading to superior performance in detecting a diverse range of face manipulations.

4.4.3. Effect of the mixing ratio in the VPD

We investigate the impact of manipulating the mixing ratio r of real to total patches fed to the VPD classifier during training. Fig. 5 plots the AUC scores achieved on various datasets for ratios considered between 0.3–0.7. The results clearly show that extreme ratios, whether heavily biased towards real or fake faces, negatively impact performance. This validates that an imbalanced distribution of data

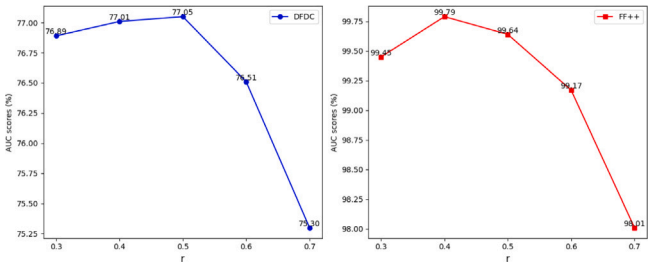


Fig. 5. AUC scores (%) achieved on various datasets for different ratios (r). The blue line delineates the performance on the DFDC dataset, whereas the red line corresponds to that on the FF++ dataset. The x-axis signifies the varying values of the ratio r, while the ordinate corresponds to the respective AUC scores.

impedes the classifier's ability to distinguish manipulations accurately. Interestingly, ratios closer to 0.5 produce superior results, with 0.4 achieving the best AUC in most cases. A slightly higher proportion of fake faces provides informative irregularity cues without overwhelming the classifier. This observation aligns with the objective of detecting fake samples. In summary, the quantitative evaluation proves that neither heavily skewed nor perfectly balanced mixing yields optimal outcomes. An intermediate ratio of 0.4–0.5 real to total patches elicits the most robust and discriminative learning within the VPD framework for the task. This finding provides useful guidance on configuring the training data presentation.

4.5. Limitations

While our approach enhances robustness and generalization, several limitations remain. The pre-processing strategies of FENeRF augmentation and 3D landmark extraction increase training overhead. Per-frame reconstruction and separate model passes during inference also induce latency issues. Moreover, interpreting the temporal patterns learned by the RNN and evaluating on more diverse data remain challenging. Addressing efficiency, transparency, broad evaluation and multimodal encoding through techniques like audio integration can help address such challenges and realize this framework's full potential for practical large-scale deployment. Nonetheless, our work contributes meaningful advances and provides promising directions for continued progress on this important task.

5. Conclusion

We propose DBNet, a multimodal framework for face forgery detection that leverages both spatial and temporal artifacts. It integrates an RNN and VPD classifier operating on 3D landmarks and local textures respectively, and employs a novel temporal consistency loss to supervise the RNN. Additionally, FENeRF-based 3D data augmentation reinforces representation learning. Extensive evaluations demonstrate DBNet achieves state-of-the-art robustness and generalization abilities. While pre-processing overhead and latency remain open challenges, our work establishes promising directions towards large-scale, real-time deployment through continued optimization of computational efficiency and multisensory encoding. Overall, DBNet makes meaningful advances on the critical task of face manipulation detection with room for exciting future work.

CRediT authorship contribution statement

Changshuang Zhou: Writing – original draft, Visualization, Software, Resources, Investigation, Data curation, Conceptualization. **Frederick W.B. Li:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Chao Song:** Writing – review & editing, Supervision. **Dong Zheng:** Visualization, Validation. **Bailin Yang:** Writing – review & editing, Validation, Funding acquisition, Conceptualization.

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.

Acknowledgments

This work was supported in part by the National Natural Science Foundation (Grant No. 62172366) and the Zhejiang Province Natural Science Foundation (Grant No. LD24F020003).

Data availability

The authors do not have permission to share data.

References

[1] T. Dzanic, K. Shah, F. Witherden, Fourier spectrum discrepancies in deep network generated images, Adv. Neural Inf. Process. Syst. 33 (2020) 3022–3032.
[2] Y. Qian, G. Yin, L. Sheng, Z. Chen, J. Shao, Thinking in frequency: Face forgery detection by mining frequency-aware clues, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII, Springer, 2020, pp. 86–103.
[3] J. Li, H. Xie, J. Li, Z. Wang, Y. Zhang, Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6458–6467.
[4] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, N. Yu, F 2 trans: High-frequency fine-grained transformer for face forgery detection, IEEE Trans. Inf. Forensics Secur. 18 (2023) 1039–1051.
[5] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, S.-N. Li, M2tr: Multi-modal multi-scale transformers for deepfake detection, in: Proceedings of the 2022 International Conference on Multimedia Retrieval, 2022, pp. 615–623.
[6] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, B. Guo, Face x-ray for more general face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5001–5010.
[7] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, W. Xia, Learning self-consistency for deepfake detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15023–15033.
[8] H. Dang, F. Liu, J. Stehouwer, X. Liu, A.K. Jain, On the detection of digital face manipulation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5781–5790.
[9] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, R. Ji, Local relation learning for face forgery detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 2, 2021, pp. 1081–1088.
[10] N. Larue, N.-S. Vu, V. Struc, P. Peer, V. Christophides, Seeable: Soft discrepancies and bounded contrastive learning for exposing deepfakes, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 21011–21021.
[11] B.M. Le, S.S. Woo, Quality-agnostic deepfake detection with intra-model collaborative learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22378–22389.
[12] D. Liu, Z. Dang, C. Peng, Y. Zheng, S. Li, N. Wang, X. Gao, FedForgery: generalized face forgery detection with residual federated learning, IEEE Trans. Inf. Forensics Secur. (2023).
[13] I. Amerini, L. Galteri, R. Caldelli, A. Del Bimbo, Deepfake video detection through optical flow based cnn, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
[14] O. De Lima, S. Franklin, S. Basu, B. Karwoski, A. George, Deepfake detection using spatiotemporal convolutional networks, 2020, arXiv preprint arXiv:2006.14749.
[15] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, L. Ma, Delving into the local: Dynamic inconsistency learning for deepfake video detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, No. 1, 2022, pp. 744–752.
[16] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, P. Natarajan, Recurrent convolutional strategies for face manipulation detection in videos, Interfaces (GUI) 3 (1) (2019) 80–87.
[17] Z. Yang, J. Liang, Y. Xu, X.-Y. Zhang, R. He, Masked relation learning for deepfake detection, IEEE Trans. Inf. Forensics Secur. 18 (2023) 1696–1708.
[18] Z. Guo, L. Wang, W. Yang, G. Yang, K. Li, LDFnet: Lightweight dynamic fusion network for face forgery detection by integrating local artifacts and global texture information, IEEE Trans. Circuits Syst. Video Technol. (2023).
[19] X. Zhu, H. Fei, B. Zhang, T. Zhang, X. Zhang, S.Z. Li, Z. Lei, Face forgery detection by 3D decomposition and composition search, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[20] X. Zhu, H. Wang, H. Fei, Z. Lei, S.Z. Li, Face forgery detection by 3d decomposition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2929–2939.

[21] Z. Sun, Y. Han, Z. Hua, N. Ruan, W. Jia, 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, 2021, pp. 3609–3618.

[22] A. Haliassos, K. Vougioukas, S. Petridis, M. Pantic, 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, 2021, pp. 5039–5049.

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

[24] W. Bai, Y. Liu, Z. Zhang, B. Li, W. Hu, AUNet: Learning relations between action units for face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24709–24719.

[25] A. Luo, R. Cai, C. Kong, X. Kang, J. Huang, A.C. Kot, Forgery-aware adaptive vision transformer for face forgery detection, 2023, arXiv preprint arXiv:2309.11092.

[26] C. Yang, S.-N. Lim, One-shot domain adaptation for face generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5921–5930.

[27] L. Chen, Y. Zhang, Y. Song, L. Liu, J. Wang, 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, 2022, pp. 18710–18719.

[28] J. Sun, X. Wang, Y. Zhang, X. Li, Q. Zhang, Y. Liu, J. Wang, Fenerf: Face editing in neural radiance fields, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 7672–7682.

[29] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, S.K. Nayar, Face swapping: automatically replacing faces in photographs, in: ACM SIGGRAPH 2008 Papers, 2008, pp. 1–8.

[30] C. Cao, Y. Weng, S. Zhou, Y. Tong, K. Zhou, Facewarehouse: A 3d facial expression database for visual computing, IEEE Trans. Vis. Comput. Graphics 20 (3) (2013) 413–425.

[31] K. Dale, K. Sunkavalli, M.K. Johnson, D. Vlasic, W. Matusik, H. Pfister, Video face replacement, in: Proceedings of the 2011 SIGGRAPH Asia Conference, 2011, pp. 1–10.

[32] J. Thies, M. Elgharib, A. Tewari, C. Theobalt, M. Nießner, Neural voice puppetry: Audio-driven facial reenactment, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16, Springer, 2020, pp. 716–731.

[33] S. Suwajanakorn, S.M. Seitz, I. Kemelmacher-Shlizerman, Synthesizing obama: learning lip sync from audio, ACM Trans. Graph. (ToG) 36 (4) (2017) 1–13.

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

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

[36] Y. Nirkin, Y. Keller, T. Hassner, Fsgan: Subject agnostic face swapping and reenactment, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7184–7193.

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

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

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

[40] K. Sun, H. Liu, T. Yao, X. Sun, S. Chen, S. Ding, R. Ji, An information theoretic approach for attention-driven face forgery detection, in: European Conference on Computer Vision, Springer, 2022, pp. 111–127.

[41] X. Dong, J. Bao, D. Chen, T. Zhang, W. Zhang, N. Yu, D. Chen, F. Wen, B. Guo, Protecting celebrities from deepfake with identity consistency transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 9468–9478.

[42] S. Hu, Y. Li, S. Lyu, Exposing GAN-generated faces using inconsistent corneal specular highlights, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 2500–2504.

[43] I. Demir, U.A. Ciftci, Where do deep fakes look? synthetic face detection via gaze tracking, in: ACM Symposium on Eye Tracking Research and Applications, 2021, pp. 1–11.

[44] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, L. Ma, Spatiotemporal inconsistency learning for deepfake video detection, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 3473–3481.

[45] Y. Zheng, J. Bao, D. Chen, M. Zeng, F. Wen, Exploring temporal coherence for more general video face forgery detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15044–15054.

[46] J. Guan, H. Zhou, Z. Hong, E. Ding, J. Wang, C. Quan, Y. Zhao, Delving into sequential patches for deepfake detection, Adv. Neural Inf. Process. Syst. 35 (2022) 4517–4530.

[47] W. Guan, W. Wang, J. Dong, B. Peng, T. Tan, Robust face-swap detection based on 3d facial shape information, in: CAAI International Conference on Artificial Intelligence, Springer, 2022, pp. 404–415.

[48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[49] Q. Fan, R. Panda, et al., Can an image classifier suffice for action recognition?, 2021, arXiv preprint arXiv:2106.14104.

[50] G.-P. Ji, M. Zhuge, D. Gao, D.-P. Fan, C. Sakaridis, L.V. Gool, Masked vision-language transformer in fashion, Mach. Intell. Res. 20 (3) (2023) 421–434.

[51] D. Wodajo, S. Atnafu, Deepfake video detection using convolutional vision transformer, 2021, arXiv preprint arXiv:2102.11126.

[52] H. Zhao, W. Zhou, D. Chen, W. Zhang, N. Yu, Self-supervised transformer for deepfake detection, 2022, arXiv preprint arXiv:2203.01265.

[53] S.A. Khan, H. Dai, Video transformer for deepfake detection with incremental learning, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 1821–1828.

[54] A. Khormali, J.-S. Yuan, DFDT: an end-to-end deepfake detection framework using vision transformer, Appl. Sci. 12 (6) (2022) 2953.

[55] W. Zhuang, Q. Chu, Z. Tan, Q. Liu, H. Yuan, C. Miao, Z. Luo, N. Yu, UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection, in: European Conference on Computer Vision, Springer, 2022, pp. 391–407.

[56] Y. Xu, J. Liang, G. Jia, Z. Yang, Y. Zhang, R. He, TALL: Thumbnail layout for deepfake video detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 22658–22668.

[57] Y. Li, S. Lyu, Exposing deepfake videos by detecting face warping artifacts, 2018, arXiv preprint arXiv:1811.00656.

[58] S. Das, S. Seferbekov, A. Datta, M. Islam, M. Amin, et al., Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3776–3785.

[59] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Li, AltFreezing for more general video face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4129–4138.

[60] Z. Zhong, L. Zheng, G. Kang, S. Li, Y. Yang, Random erasing data augmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, No. 07, 2020, pp. 13001–13008.

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

[62] Y. Li, L. Liu, H. Qin, S. Deng, M.A. El-Yacoubi, G. Zhou, Adaptive deep feature fusion for continuous authentication with data augmentation, IEEE Trans. Mob. Comput. (2022).

[63] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. Vetter, A 3D face model for pose and illumination invariant face recognition, in: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, IEEE, 2009, pp. 296–301.

[64] B. Mildenhall, P.P. Srinivasan, M. Tancik, J.T. Barron, R. Ramamoorthi, R. Ng, Nerf: Representing scenes as neural radiance fields for view synthesis, Commun. ACM 65 (1) (2021) 99–106.

[65] R.E. Kalman, A new approach to linear filtering and prediction problems, J. Basic Eng. 82 (1) (1960).

[66] C. Sagonas, E. Antonakos, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 faces in-the-wild challenge: Database and results, Image Vis. Comput. 47 (2016) 3–18.

[67] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, S.Z. Li, Towards fast, accurate and stable 3d dense face alignment, in: European Conference on Computer Vision, Springer, 2020, pp. 152–168.

[68] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, C.C. Ferrer, The deepfake detection challenge (dfdc) dataset, 2020, arXiv preprint arXiv:2006.07397.

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

[70] D.E. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.

[71] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, X. Yang, End-to-end reconstruction-classification learning for face forgery detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4113–4122.

[72] K. Yao, J. Wang, B. Diao, C. Li, Towards understanding the generalization of deepfake detectors from a game-theoretical view, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 2031–2041.

[73] B. Liu, B. Liu, M. Ding, T. Zhu, X. Yu, TI2Net: Temporal identity inconsistency network for deepfake detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 4691–4700.

[74] T. Baltrušaitis, P. Robinson, L.-P. Morency, Openface: an open source facial behavior analysis toolkit, in: 2016 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2016, pp. 1–10.

[75] J. Shang, T. Shen, S. Li, L. Zhou, M. Zhen, T. Fang, L. Quan, Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency, in: European Conference on Computer Vision, Springer, 2020, pp. 53–70.

[76] Z. Wang, X. Zhu, T. Zhang, B. Wang, Z. Lei, 3D face reconstruction with the geometric guidance of facial part segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 1672–1682.

[77] S. Das, M. Kolahdouzi, L. Özparlak, W. Hickie, A. Etemad, Unmasking deepfakes: Masked autoencoding spatiotemporal transformers for enhanced video forgery detection, 2023, arXiv preprint arXiv:2306.06881.

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