ADA-FInfer: Inferring Face Representations from Adaptive Select Frames for High-Visual-Quality Deepfake Detection

Juan Hu, Jinwen Liang, *Member, IEEE*, Zheng Qin*, *Member, IEEE*, Xin Liao*, *Senior Member, IEEE*, Wenbo Zhou, *Member, IEEE*, and Xiaodong Lin, *Fellow, IEEE*

Abstract—Interpretable deepfake detection is gaining attention for providing explainable, trustworthy results, avoiding the limitations of 'black-box' models. Current interpretable methods focus on visible artifacts in low-visual-quality deepfakes, but these artifacts become less apparent in high-visual-quality deepfakes generated by advanced models. With advancements in deep generative models, producing high-visual-quality deepfakes has become a strategy to evade detection. To address this, we propose ADA-FInfer, an adaptive frame selection and interpretable face representation inference method for detecting high-visual-quality deepfakes. ADA-FInfer adaptively selects frames by analyzing optical flow to reveal manipulations. We also introduce an adaptive attack method that manipulates specific frames, and our adaptive selection strategy shows resistance to such attacks. ADA-FInfer uses an encoder to learn face representations from source and target faces, applying a representation-prediction loss to maximize the distinction between real and fake videos. To provide further insights, we employ the joint entropy, mutual information, and conditional entropy analyses to explain the method's effectiveness. Extensive experiments and ablation studies demonstrate that ADA-FInfer achieves promising performance in detecting high-visual-quality deepfakes.

Index Terms—Deepfake detection, high-visual-quality Deepfake videos, inferring face representations, adaptive frame selection.

1 Introduction

The rapid rise of synthetic media, particularly deepfake videos, has raised significant concerns about its impact on public trust and the integrity of information ecosystems [2], [3], [4], [5], [6], [7], [8], [9]. According to `sensity.ai` [10], deepfake-generated pornographic videos have accumulated over 135 million views online, frequently used for malicious purposes such as public shaming or extortion [11]. In 2021, a deepfake video falsely depicted Dr. Anthony Fauci stating that vaccines do not protect against COVID-19 [12]. The increasing sophistication of such videos presents a growing threat to individuals, political stability, and even national security [13], [14], [15], [16], [17]. To mitigate these risks, various deepfake detection methods have been developed [7], [8], [9], [18], [19], [20].

However, most existing deepfake detection frameworks rely on 'black-box' models, which often lack transparency and fail to provide explainable detection results [21]. The absence of clear, evidence-based explanations raises challenges in detection evaluation and reduces the utility of these models in high-stakes contexts such as legal forensics, where AI-based decisions must be backed by verifiable evidence [22], [23], [24]. Thus, there is a pressing need for deepfake detection methods that not only achieve accurate detection but also provide interpretable insights.

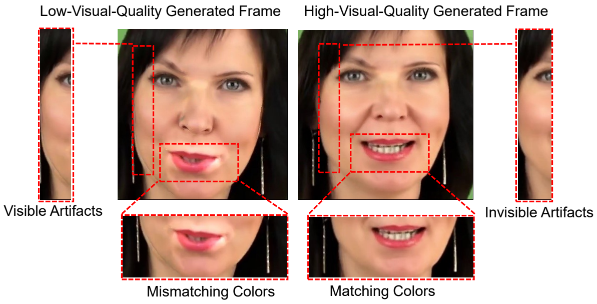

Low-Visual-Quality Generated Frame
High-Visual-Quality Generated Frame
Visible Artifacts
Invisible Artifacts
Mismatching Colors
Matching Colors

Fig. 1. A comparison between the DeepFake frames with low-visual-quality (left) and the DeepFake frames with high-visual-quality (right). The low-visual-quality frame is generated by the DeepFakes [25], and the high-visual-quality frame is generated by the NeuralTextures [26]. It shows that the high-visual-quality frame reduces the artifacts and color mismatch.

Current interpretable deepfake detection methods [2], [3], [4], [5], [6], [27], [28], [29] mainly focus on identifying visible features such as texture inconsistencies, temporal artifacts, and tampered traces. These approaches work well on low-visual-quality deepfake videos, where artifacts are more easily observed. However, as shown in Fig. 1, these artifacts become much less detectable in high-visual-quality deepfakes, which are generated using advanced techniques that minimize visible tampering by preserving color consistency, reducing temporal flicker, and producing high-resolution synthesized faces [30]. This evolution poses new challenges for detecting manipulated content in high-visual-

Juan Hu, Zheng Qin, and Xin Liao are with the College of Computer Science and Electronic Engineering, Hunan University, China. (e-mail: hujuan@hnu.edu.cn, zqin@hnu.edu.cn, xinliao@hnu.edu.cn).

Jinwen Liang is with the Department of Computing, the Hong Kong Polytechnic University, Hong Kong. (e-mail: jinwen.liang@polyu.edu.hk).

Wenbo Zhou is with CAS Key Laboratory of Electromagnetic Space Information, University of Science and Technology of China, China (e-mail: welbeckz@ustc.edu.cn).

Xiaodong Lin is with the School of Computer Science, University of Guelph, Guelph, ON N1G 2W1, Canada (e-mail: xlin08@uoguelph.ca).

**Corresponding authors: Zheng Qin and Xin Liao.*

Part of this research work was presented in AAAI 2022 [1].

quality videos.

Existing methods for detecting high-visual-quality deepfakes primarily rely on direct feature extraction [8], [9], [29], [31], [32], [33], whole-face predictions [34], [35], or 2D reconstructions [36]. These approaches tend to increase computational complexity and are less robust, particularly when faced with attacks that manipulate specific frames in a video. Moreover, they fail to incorporate adaptive frame selection, leading to performance degradation when partial frame manipulation occurs.

To address these challenges, a new method is needed—one that improves robustness and efficiency in detecting high-visual-quality deepfakes, provides clear explanations based on information theory, and defends against partial frame manipulation attacks.

Following the above motivations, we propose an adaptive frame selection strategy and an interpretable face representations inference-based method (ADA-FInfer) for detecting high-visual-quality deepfakes.

Our approach leverages information theory to explain the distinctions between real and fake videos. Specifically, for deepfake videos, unnatural facial expressions result in mismatches between the predicted and referenced target face representations. In contrast, real videos exhibit a strong match between predicted and referenced target faces.

This paper introduces several key innovations. First, we implement an adaptive frame selection strategy based on optical flow, selecting frames with minimal inter-frame motion to enhance detection performance. This ensures compatibility across different types of deepfake videos by adjusting the number of selected frames based on video content. Second, we apply the Gaussian-Laplacian pyramid for preprocessing, improving feature extraction from faces. Third, an encoder extracts high-dimensional information from both source and referenced target faces, reducing the overall dimensionality and improving computational efficiency. Fourth, an autoregressive model predicts long-term global facial features, helping to identify regularities across multiple time scales. The final step involves optimizing a representation-prediction loss, which maximizes the distinction between predicted and referenced target faces.

The major contributions of this paper are as follows:

(1) We transform deepfake detection into a video prediction regression task. Rather than directly extracting features from frames, ADA-FInfer predicts the target frame's face and compares it to the referenced target face, optimizing a representation-prediction loss for detecting high-visual-quality deepfakes.

(2) We introduce an adaptive attack method that manipulates specific frames within a video and develop a defense mechanism using our adaptive frame selection approach to counter partial frame manipulation.

(3) We provide interpretability analyses, demonstrating ADA-FInfer's effectiveness. Joint entropy analyses show that ADA-FInfer can effectively detect high-visual-quality deepfakes with low joint entropy. Mutual information analyses reveal that ADA-FInfer distinguishes real and fake videos, while conditional entropy analyses emphasize the importance of limiting the length of source and target faces.

(4) Extensive performance evaluations on high-visual-quality deepfake datasets demonstrate that ADA-FInfer achieves promising performance in extensive metrics.

The remainder of this paper is organized as follows. Section 2 illustrates the related work. Section 3 describes the proposed ADA-FInfer. Section 4 provides the interpretability analyses of ADA-FInfer. Section 5 demonstrates experimental evaluation results. Section 6 concludes this paper.

2 Related Work

In this section, we first describe several techniques for generating high-visual-quality Deepfake videos. Thereafter, the Deepfake video forensic methods are introduced as follows. Then, we provide the related work of the video prediction methods. Ultimately, we describe the differences between the initial version [1] and this paper.

2.1 High-Visual-Quality Deepfake Generation

The recent-developed DeepFake synthesis algorithms improve the visual quality of videos in several aspects, i.e., the resolution of frames, the color corrections, the smoothness mask, the temporal correlations, the training time, and the face ratio.

Specifically, the algorithm in [30] improved the resolution of frames by using encoder and decoder models with more layers and dimensions. The color transfer algorithm is applied to solve the color mismatch in low-visual-quality videos, and the training faces' colors are randomly perturbed. Furthermore, Li et al. [30] improved the mask generation step for reducing the boundary artifacts of the mask. Since the previous low-visual-quality videos exist the temporal flickering, the algorithm in [30] utilized the Kalman smoothing algorithm to improve the inter-frame continuity. Besides, the previous synthesis algorithms do not adapt to the lighting, the scene, and a set of other factors in generating the Deepfake videos. Consequently, the face area in low-visual-quality videos often has perceptible distortions, such as jitter, blur, or strange artifacts. Thereafter, the algorithm in [37] took these factors into full consideration and generated many high-visual-quality face images through a long training time. If the subject's face is too close to the camera, the forgery algorithm will not produce a clear and credible face transformation. When generating the datasets, the forgery algorithm filters out all videos with above-average faces. To tackle this issue, Dolhansky et al. [38] filtered out all faces with a high face ratio. Hereafter, the high-visual-quality Deepfake videos with a low face ratio are generated.

2.2 Deepfake Videos Detection Methods

With the proliferation of artificial intelligence and multimedia technologies [39], the threshold for generating Deepfake videos is getting lower. Detecting Deepfake videos becomes an important issue in multimedia forensics.

Lots of Deepfake detection methods are proposed to against inchoate Deepfakes, i.e., cue-inspired methods, data-driven methods, and multi-domain fusion methods. The cue-inspired methods include methods that rely on semantic-cues [2], [4] and signal-cues [6], [40], [41], [42]. With the recent progress of Deepfake technologies, purposely training during the generation of fake videos can make some cue-inspired methods invalid. The data-driven methods [18], [19], [20], [43], [44], [45], [46], [47] rely on mass

TABLE 1
Summaries of existing high-visual-quality DeepFake video detection works.

Category	Methods	Limitations	Capabilities
Direct-based	SPSL [8]	Lack integration of temporal features	Detect spatial and phase spectrum forgery traces
	NoiseDF [31]	Lack integration of temporal features	Detect digital forensic noise traces of face-background pairs
	DisGRL [32]	Lack robust feature extraction	Detect forgery-sensitive and genuine compact visual traces
	STN [33]	Lack robust feature extraction	Detect spatial and temporal inconsistencies
	FT-two-stream [9]	Lack robust feature extraction	Detect spatial artifacts and temporal inconsistencies
	Xia et al. [29]	Lack integration of temporal features	Detect facial textural disparities
Prediction-based	Oc-fakedect [35]	Lack robust feature extraction	Predict the whole faces
	RECCE [34]	High Multi-Adds Calculation in the Model	Predict the whole faces
	BRCNet [36]	Lack robust feature extraction	Predict 2D and 3D reconstruction

data and deep learning. However, these data-driven methods focus on a single domain and ignore the fusion of other domains, which promotes research on multi-domain fusion methods. The multi-domain fusion methods detect Deepfake videos by combining multi-domain features. Güera et al. [7] introduced an RNN model with InceptionV3 [48] to tackle Deepfake videos by extracting spatio-temporal features. Liu et al. [8], [49] combined the frequency domain and spatial domain to improve the transferability for Deepfake video detection. Masi et al. [50], [51] devised a two-branch method to isolate manipulated faces.

The refined Deepfake technologies can reduce the artifacts, color mismatch, temporal flickering, and inaccurate face masks, which makes the forensics in high-visual-quality Deepfake videos even more challenging. Lots of methods are proposed for high-visual-quality Deepfake videos. Specifically, we broadly split the existing high-visual-quality DeepFake video detection works into direct-based methods and prediction-based methods. The majority of these methods detect DeepFake videos by directly extracting features from videos or frames, and are thus classified as direct-based methods. The limitations of these methods lie in the absence of temporal information [8], [29], [31] or robust features [9], [32], [33]. To improve the robustness and avoid relying on direct-based features, a limited number of prediction-based methods have emerged. These studies predict whole faces [34], [35] and 2D reconstruction [36] to detect DeepFake videos. We summarise these methods in Table 1.

2.3 Video Prediction Methods

Various methods are proposed for video prediction, and these methods achieve reasonable performance, i.e., direct pixel synthesis methods, explicit transformations methods, and representation-based prediction methods.

The conventional video prediction methods attempt to directly predict future pixels by implicitly modelling the scene dynamics and low-level details. Specifically, Byeon et al. [52] introduced a fully context-aware architecture to predict the future accurately. They utilize the past context, which is of crucial importance for video prediction. Since some information is not passed to the next layers during learning, Oliu et al. [53] proposed an architecture based on a recurrent convolution layer. However, the pixel space is high-dimensional, which makes it difficult to extract a robust representation from a raw pixel.

The explicit transformations methods model the transformations that take a frame at time *one* to the frame at time *two*. Typically, Lu et al. [54] described a modular framework for video frame prediction, allowing the prediction of an entire image sequence rather than just a single frame ahead. Chen et al. [55] found that different objects in the same scene often moved and deformed in different ways intuitively, which promotes them to explore a novel video prediction model that learns local motion transformation. However, the performance of explicit transformation methods drops when dealing with multiple objects.

The representation-based prediction methods transfer the frames into representations and predict future video representations. Video prediction remains a challenging task because of the high-dimensionality of video frames. Hsieh et al. [56] addressed this challenge by proposing a framework named DDPAE, which can learn both the latent decomposition and disentanglement. Thereafter, Oord et al. [57] proposed a framework for extracting compact latent representations to encode predictions. The representation-based prediction methods can reduce the representation dimensionality, which is a benefit for video prediction.

2.4 Differences from FInfer [1]

This paper is an extended version of the conference paper [1]. We briefly summarize the differences as follows.

(1) To enhance the detection performance, we devise an adaptive frame selection strategy based on the optical flow and propose ADA-FInfer. The devised strategy makes the model resistant to adaptive attacks in which parts of the frame are tampered with. Furthermore, the proposed strategy allows for flexible prediction ways that are not limited to predicting current faces from previous faces. Moreover, the proposed strategy adapts to different types of videos and further improves the cross-domain detection performance. By utilizing the selection strategy, ADA-FInfer improves the detection performance in multiple metrics.

(2) We provide additional analyses for the adaptive selection strategy and further analyses of the interpretation for detecting videos with high-visual-quality. By analyzing the entropy, joint entropy, and mutual information, we can provide interpretations for the distinction between high-visual-quality Deepfake videos and real videos.

(3) To evaluate the flexibility and compatibility of ADA-FInfer, we provide extended ablation studies in different prediction ways and different frame selection ways, ex-

This article has been accepted for publication in IEEE Transactions on Dependable and Secure Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2024.3523289

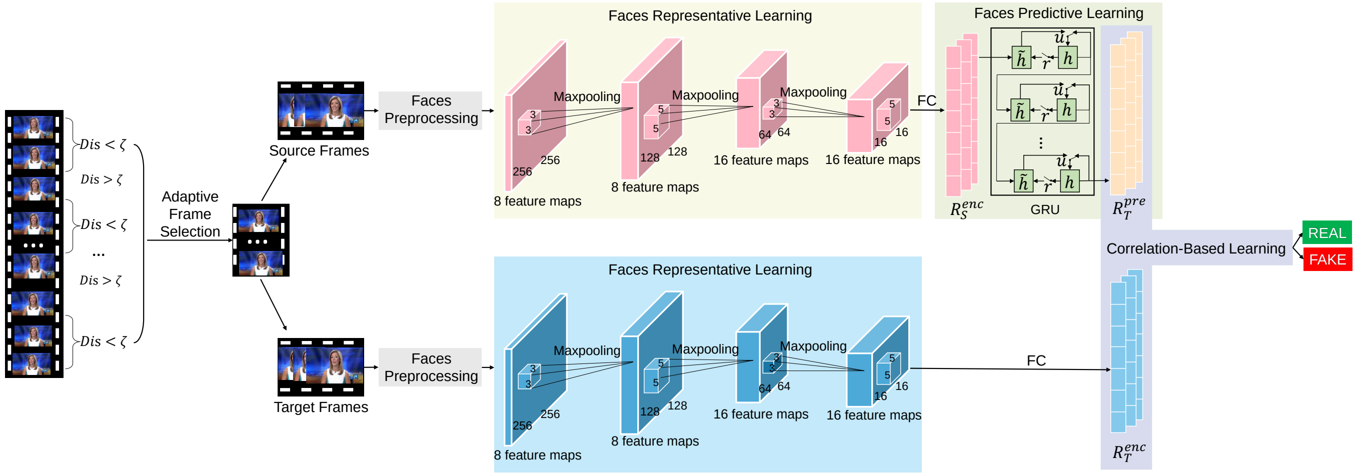


Fig. 2. The proposed ADA-FInfer for detecting high-visual-quality Deepfake videos. Source frames and target frames are selected from videos based on optical flow. Faces are preprocessed with the Gaussian-Laplacian pyramid to expose boundaries. The faces representative learning module encodes source faces and referenced target faces into vectors. The faces predictive learning module predicts the target face representations from the source face representations. The correlation-based learning module compares the predicted target face representations with the referenced target face representation. By optimizing a devised representation-prediction loss, ADA-FInfer can effectively detect the high-visual-quality Deepfake videos.

tended cross-domain detection in different scenarios, and extended comparison experiments.

(4) We observe that the high-visual-quality Deepfake videos and real videos are different in the correlation matrix between the predicted target representations and the referenced target representations, which demonstrates the effectiveness of ADA-FInfer.

3 The Proposed ADA-FInfer

As illustrated in Fig. 2, the proposed ADA-FInfer consists of five components: frames selection, faces preprocessing, faces representative learning, faces predictive learning, and correlation-based learning between referenced target faces and predicted target faces. These five parts are mutually reinforced with optimization. In this section, we first present all notations. Thereafter, we formulate an overview of the proposed ADA-FInfer. Ultimately, the five components of the proposed method are introduced.

3.1 Notations

Table 2 shows all notations and descriptions. In the adaptive frame selection process, *Opt* represents the optical flow matrices of consecutive frames. *Dis* is the inter-frame motion value.

Let $X = \{x_1, x_2, \dots, x_s, \dots, x_{s+t}\}$ be $s + t$ frames in a video. The s and t are the length of source frames and the length of target frames, respectively. The x_i is the i-th frame. All the frames are utilized to crop faces. Let $A = \{a_1, a_2, \dots, a_s, \dots, a_{s+t}\}$ be the faces extracted from X. These faces are divided into source faces and target faces. Let $A_S = \{a_1, a_2, \dots, a_s\}$ and $A_T = \{a_{s+1}, a_{s+2}, \dots, a_{s+t}\}$ be the source faces and target faces, respectively. The a_i is the i-th face. Let P and Q be the spatial image intensity changes and temporal image intensity changes. Let ζ be the threshold that determines the frame selection. These faces are preprocessed by Gaussian-Laplacian pyramid. Let $G_k(\alpha, \beta)$ and $L_k(\alpha, \beta)$ be Gaussian pyramid and Laplacian pyramid of layer k, respectively. Let f^{enc} be the encoder of the representative learning module. The source faces and target faces are converted to vectors by utilizing f^{enc}. The referenced source face representations are denoted as R_S^{enc}, and the referenced target face representations are denoted as R_T^{enc}. Let g^{pre} be the prediction model of the faces predictive learning module. The prediction process utilizes various gates and Time-Distributed layer. Let u, τ, h, r, and d be the update gate, the time step, the hidden states, the reset states, and Time-Distributed layer in the prediction process, respectively. The faces predictive learning module obtains the predicted face representations that are denoted as R_T^{pre}. Let *corr* be the correlation between R_T^{enc} and R_T^{pre}. The loss function of the ADA-FInfer is denoted as L_N.

In the information theory analyses process, $H(x_s)$, $H(x_s, x_{s+t})$, $MI(x_s, x_{s+t})$, *TE*, and $H(x_s|x_{s+t})$ can be represented by information entropy, joint entropy, mutual information, temporal entropy and conditional entropy, respectively.

In the Multi-Adds calculation process, the kernel size of a convolutional layer is represented by kc. The numbers of the input channel are represented by C_{in}. The number of the output channel is C_{out}. The size of the output feature map is represented by M_{out}. The Multi-Adds computation of the convolutional layer is represented by M_A.

3.2 Overview the Proposed ADA-FInfer

Fig. 2 shows the proposed ADA-FInfer for high-visual-quality Deepfake video detection. First, we implement the devised adaptive frame selection strategy to select frames from videos. Second, we preprocess faces by utilizing the Gaussian-Laplacian pyramid block. Third, we propose an encoder that utilizes representative learning to encode the referenced source faces and referenced target faces. Since the data dimensions of video frames are enormous, encoding frames to a low-dimensional is essential for effective prediction. Fourth, we use an autoregressive model to predict the representations of target frames. The prediction model can integrate the information of previous frames to predict

TABLE 2
Notations and Descriptions

Notations	Descriptions
Opt	The optical flow matrices of consecutive frames.
Dis	The inter-frame motion value.
X	$X = \{x_1, \ldots, x_s, \ldots, x_{s+t}\}$ is set of frames.
s	The length of source frames.
t	The length of target frames.
x_i	The i-th frame.
A	$A = \{a_1, \ldots, a_s, \ldots, a_{s+t}\}$ is set of faces.
A_S	$A_S = \{a_1, a_2, \ldots, a_s\}$ is source faces.
A_T	$A_T = \{a_s, a_{1+s}, \ldots, a_{s+t}\}$ is target faces.
a_i	The i-th face.
b	The size of the resized face is $b * b$.
$G_k(\alpha, \beta)$	The Gaussian pyramid of layer k at pixel (α, β).
$L_k(\alpha, \beta)$	The Laplacian pyramid of layer k at pixel (α, β).
f^{enc}	The encoder of the representative learning module.
g^{pre}	The prediction model.
R_S^{enc}	The referenced source face representations.
R_T^{enc}	The referenced target face representations.
u	The update gate.
τ	The time step.
h	The hidden states.
r	The reset states.
d	The Time-Distributed layer.
R_T^{pre}	The predicted target face representations.
corr	The correlation between R_T^{enc} and R_T^{pre}.
L_N	The loss function of the proposed method.
$H(x_s)$	The information entropy of x_s.
$H(x_s, x_{s+t})$	The joint entropy of x_s and x_{s+t}.
$MI(x_s, x_{s+t})$	The mutual information of x_s and x_{s+t}.
TE	The temporal entropy.
$H(x_s \mid x_{s+t})$	The conditional entropy of x_s and x_{s+t}.
kc	The kernel size of a convolutional layer.
C_{in}	The numbers of input channel.
C_{out}	The numbers of output channel.
M_{out}	The size of the output feature map.
M_A	The Multi-Adds of a convolutional layer.
P	The spatial image intensity changes.
Q	The temporal image intensity changes.
ζ	The threshold that determines the frame selection.

effectively. Finally, we carry out correlation-based learning, which contrasts referenced target face representations with predicted target face representations. A representation-prediction loss is devised to allow for the whole model to be trained end-to-end. The loss can be fed back to the representative learning module and predictive learning module, which would later prompt the model to encode face representations, predict the target representations, and detect the videos.

3.3 Adaptive Frame Selection Strategy

The adaptive frame selection strategy aims to overcome the prediction errors caused by high inter-frame motion, promote the prediction and further improve the detection performance. The strategy adaptively chooses the frames in a given video sequence for the prediction process based on the motion occurring between consecutive frames. Then, the inter-frame motions are quantified and the frames are adaptively selected according to the quantified value of movements.

This strategy significantly improves detection performance by making the model resistant to adaptive attacks where parts of the frame are tampered with, enabling flexible prediction methods that are not limited to predicting current faces from previous frames, and adapting to different types of videos, which further enhances cross-domain detection performance across multiple metrics.

We use optical flow to quantify the inter-frame motion. The optical flow between two successive frames is calculated using the Lucas Kanade algorithm [58]. According to Lucas Kanade algorithm, the optical flow matrices *Opt* can be represented as:

$$Opt = P^{-1}(-Q). \tag{1}$$

The P and Q can be represented as:

$$P = \begin{bmatrix} \sum\limits_{j=1}^{j=\eta} ((a_i^j)_\chi)^2 & \sum\limits_{j=1}^{j=\eta} (a_i^j)_\chi (a_i^j)_\gamma \\ \sum\limits_{j=1}^{j=\eta} (a_i^j)_\chi (a_i^j)_\gamma & \sum\limits_{j=1}^{j=\eta} ((a_i^j)_\gamma)^2 \end{bmatrix}, \tag{2}$$

$$Q = \begin{bmatrix} \sum\limits_{j=1}^{j=\eta} ((a_i^j)_\chi)((a_i^j)_\mu) \\ \sum\limits_{j=1}^{j=\eta} ((a_i^j)_\gamma)((a_i^j)_\mu) \end{bmatrix}, \tag{3}$$

where $(a_i^j)_\chi$, $(a_i^j)_\gamma$, $(a_i^j)_\mu$ represent the derivative of face a_i at pixel j along the abscissa axis, vertical axis, time axis directions, respectively, and η is the size of the calculation window.

All values in *Opt* are normalized to $[0, 1]$ by using the min-max rules. Then, we obtain the inter-frame motion value *Dis* by calculating the L2 norm of *Opt*.

$$Dis = ||Opt||_2. \tag{4}$$

We use a threshold ζ that was decided empirically to adaptively select frames. A high value of *Dis* is considered to have high inter-frame motion, and it could potentially represent huge face changes. Since slight face changes can facilitate the prediction process, we use ζ to select frames with small inter-frame motion values. Videos can be divided into multiple video clips. The first video clip begins with the first frame of a video and ends with the frame whose *Dis* is higher than ζ. Other video clips begin with the next frame of the previous video clip and end with the frame whose *Dis* is higher than ζ. These video clips are sorted in descending order according to the lengths of the clips. We select the video clip with a median length, and explanations of the selection strategy are provided in Section 4.3. For the selected frames, there are flexible prediction ways, such as predicting current faces based on the previous faces, predicting previous faces based on current faces, and predicting previous faces and future faces based on current faces. We discuss these prediction ways in Section 5.4.

3.4 Faces Preprocessing

In the faces preprocessing module, the Gaussian-Laplacian pyramid is utilized to improve the visibility of tamper traces that facilitate the face representations.

(1) The Gaussian pyramid is used to generate multiple sets of faces at different scales. The Gaussian pyramid of layer k is obtained from the layer $k-1$ by the Gaussian function, W convolution, and downsampling.

$$G_k(\alpha,\beta) = \sum_{b=-2}^{2}\sum_{b=-2}^{2} W(b,b)G_{k-1}(2\alpha+b, 2\beta+b). \quad (5)$$

By implementing the Gaussian pyramid, a face can be broken down into successively small groups of pixels and blurred. Faces are blurred in this way, making the edges of the face and tamper traces easier to detect.

(2) The Laplacian pyramid of layer k is obtained from $G_k(\alpha,\beta)$ and $G_{k+1}(\alpha,\beta)$.

$$L_k(\alpha,\beta) = G_k(\alpha,\beta) - PyrUp(G_{k+1}(\alpha,\beta)), \quad (6)$$

where $PyrUp$ is a function of upsampling. The Laplacian pyramid can not only provide representations of the face but also provide a set of face details at different scales. With detailed representations, the detection of tamper traces can be facilitated. We discuss the impact of the Gaussian-Laplacian pyramid in Section 5.6.

3.5 Faces Representative Learning

In the faces representative learning module, an encoder is utilized to extract source face information and target face information. When we predict the target frames, a lot of frames are used to train the model. One of the challenges of predicting high-dimensional data is to reconstruct every detail in the data, which is a time-consuming task. Furthermore, modelling complex relationships from frames is computationally intensive. To improve training efficiency, we propose an encoder with a lightweight structure to obtain the representations of the faces. The encoder consists of four different blocks, and the details are provided as follows.

In the first blocks, the convolutional layer uses 8 filters, and the kernel size is 3*3. Then, the ReLU activation functions and Batch Normalization [59] are utilized to introduce non-linearities and regularize the output. To reduce overfitting, the Max pooling layer with kernel size 2*2 is used to train the model. In the second block, in order to ensure that the model can learn effective features from frames, we use a combination of the small convolution kernel and the large convolution kernel. Thus, the kernel size of the second convolution module is 5*5. Other settings are the same as the first blocks. In the third block, the convolutional layer uses 16 filters, and the kernel size is 3*3. The kernel size of the Max pooling layer is 4*4. Other settings are the same as the first blocks. In the fourth block, the convolutional layer uses 16 filters, and the kernel size is 5*5. Other settings are the same as the third block.

The encoder learns features by encoding information shared at multiple time points while discarding short-term local information. It maps the source faces $A_S = \{a_1, a_2, \ldots, a_s\}$ and target faces $A_T = \{a_s, a_{s+1}, \ldots, a_{s+t}\}$ to a sequence of latent representations. The source face representations R_S^{enc} and target face representations R_T^{enc} can be expressed as follows.

$$R_S^{enc} = f^{enc}(A_S), \quad (7)$$

$$R_T^{enc} = f^{enc}(A_T). \quad (8)$$

3.6 Faces Predictive Learning

The faces predictive learning module utilizes the regressive prediction to infer the representations of target faces. When predicting further in the target faces, the amount of shared information is greatly reduced, and the model needs to infer more long-term global features. These inferred features across many time scales and are often more useful for detecting the regularities of faces. The details are provided as follows.

The update gate is utilized to update the information that the source faces are carried into the current faces. At the time step τ, the update gate u_τ is calculated by using the following equation.

$$u_\tau = \sigma(W_u \cdot [h_{\tau-1}, R_S^{enc}(\tau)]), \quad (9)$$

where $h_{\tau-1}$ represents the information of $\tau-1$, $R_S^{enc}(\tau)$ represents the source face representations of τ, and W_u is a trainable parameter.

The reset gate is utilized to remove redundant short-term local information from the source faces. At the time step τ, the reset gate r_τ is calculated by using the following equation.

$$r_\tau = \sigma(W_r \cdot [h_{\tau-1}, R_S^{enc}(\tau)]), \quad (10)$$

where W_r is a trainable parameter.

The candidate hidden states assist the calculation of the hidden states to facilitate the prediction. The candidate hidden states are the new memory content that will use the reset gate to store the relevant information from the source faces. At the time step τ, the candidate hidden states $\widetilde{h_\tau}$ are calculated as follows.

$$\widetilde{h_\tau} = tanh(W_{\widetilde{h}} \cdot [r_\tau, R_S^{enc}(\tau)]), \quad (11)$$

where $W_{\widetilde{h}}$ is a trainable parameter

The hidden states are utilized to hold information for the current faces information and pass it down to the network. The update gate u_τ determines what to collect from the current memory content $\widetilde{h_\tau}$ and the previous faces $h_{\tau-1}$. The hidden states h_τ are calculated as follows.

$$h_\tau = (1-u_\tau) \bigodot \widetilde{h_\tau} + u_\tau \bigodot h_{\tau-1}. \quad (12)$$

The prediction module transmits relevant information along a long chain sequence to make predictions. The regressive prediction can preserve long-term global features through various gate functions and remove irrelevant information. The output op can be calculated as follows.

$$op = \sigma(W_o \cdot h), \quad (13)$$

where W_o is a trainable parameter.

The Time-Distributed layer d uses the output op to perform a series of tensor operations. It can flatten the face representations so that the predicted representations become one-to-one correspondences with the time series. We use the following equation to distribute the face predictions.

$$R_T^{pre} = d(op) = \{r_{s+1}^{pre}, r_{s+2}^{pre}, \cdots, r_{s+t}^{pre}\}. \quad (14)$$

Algorithm 1: The algorithm process of the proposed ADA-FInfer

```
Input:
The source faces $A_S$. The target faces $A_T$. The initial learning rate $\alpha_{df} = 0.001$ decayed by the factor $0.2$ when the accuracy plateaus. The batch size $b = 8$. The number of iterations $num\_iter$.
Output:
Trained models: the detection model $\theta_{df}$, the encoder model $f^{enc}$, the predicted model $g^{pre}$.
Calculating the optical flow matrices Opt according to Eq. (1)(2)(3).
Normalizing values of Opt, and calculating the L2 norm of Opt to obtain the inter-frame motion Dis.
for $frames \in video$ do
    if $Dis > \zeta$ then
        Cutting the video clip. The clips end at the frame where Dis of consecutive frames is higher than the threshold $\zeta$.
end for
Selecting the video clip with a median length.
Preprocessing the video clips and getting faces $A_S$ and $A_T$
while $Step < max\_steps$ do
    for $i = 1 \to num\_iter$ do
        $R_S^{enc} = f^{enc}(A_S)$
        $R_T^{enc} = f^{enc}(A_T)$
        $R_T^{pre} = g^{pre}(R_S^{enc})$
        $g_{\theta_{df}} \leftarrow \nabla_{\theta_{df}}(\frac{1}{b}\sum_{i=1}^{b} L_N(R_T^{enc}, R_T^{pre}))$
        $\theta_{df} \leftarrow \theta_{df} + \alpha_{df} \cdot \text{Adam}(\theta_{df}, g_{\theta_{df}})$
    end for
end while
```

3.7 Correlation-Based Learning

The correlation-based learning module utilizes the representation-prediction loss to optimize the model. For the Deepfake videos, the faces lack expressiveness, which could affect the coherence of the expressions. The incoherent inter-frame expression has a crucial impact on prediction. Thereafter, the prediction of the target faces will be inaccurate. On the contrary, the faces of real videos are objective and coherent among frames. The consistency of frames makes it possible to predict the correct target face. Thus, the correlation between the predicted target face representations and target face representations of the real video is greater than that of fake videos. It is necessary to perform correlation-based learning. The details are provided as follows.

Correlation calculation. R_T^{enc} and R_T^{pre} are integrated into the correlation-based learning module. R_T^{enc} is the output of the faces representative learning module. R_T^{pre} is the output of the faces predictive learning module. The correlation of the e-th video between the predicted target face representations R_T^{pre} and the referenced target face representations R_T^{enc} is calculated as follows.

$$corr_e = sigmoid\left(\frac{\langle R_T^{pre}, R_T^{enc}\rangle}{t}\right). \tag{15}$$

Loss function. In the process of backpropagation, the representation-prediction loss is employed to train the model end-to-end. The equation of the representation-prediction loss L_N is shown in Eq. (16).

$$L_N = \frac{-1}{N}\sum_e ((y_e ln(corr_e)) + (1 - y_e) ln(1 - corr_e)), \tag{16}$$

where y_e represents the labels of the e-th video.

Model optimization. In order to optimize the proposed method, we update the representative learning module and prediction model iteratively and minimize the sum of the loss L_N. A detailed description of the ADA-FInfer is provided in Algorithm 1 for training the models. First, we select frames and preprocess faces to obtain the input data A_S and A_T. Second, the faces representative learning module obtains the representations R_S^{enc} and R_T^{enc}. Third, the faces predictive learning module obtains the representations of predicted target faces R_T^{pre}. Fourth, the correlation-based learning module minimizes the loss L_N, which contrasts R_T^{enc} with R_T^{pre}. After the training, the original videos have a higher relevance *corr* because of the natural expression. The Deepfake videos exit stiff facial expressions, which causes some impact on the prediction. Thus, the Deepfake videos have a lower *corr* value. Then, the model can find the difference between the original videos and Deepfake videos. Finally, the model outputs the *corr* and calculates the accuracy by utilizing the binary accuracy algorithm.

4 Interpretation Analyses

In this section, we first provide analyses for the distinction between low-visual-quality videos and high-visual-quality videos. Then, the analyses for detecting the difference between real videos and fake videos are conducted. Finally, we provide analyses for the proposed selection strategy.

4.1 Interpretation for the Difference between High-Visual-Quality Videos and Low-Visual-Quality Videos

In this subsection, we analyze the joint entropy to show that the proposed method can ideally detect high-visual-quality Deepfake videos with low joint entropy.

Firstly, we express the joint entropy of the transition from frame x_s to frame x_{s+t} by following the idea of [60], [61]. Three matrices $C_{s,s+t}^R$, $C_{s,s+t}^G$, $C_{s,s+t}^B$ are created carrying information on the grey level transitions between frames x_s and x_{s+t}. Suppose x_s and x_{s+t} with gray levels ${x_s}^\omega$ and ${x_{s+t}}^\psi$, respectively. The joint entropy of the R component is expressed by:

$$H(x_s, x_{s+t})^R = -\sum_{{x_s}^\omega}\sum_{{x_{s+t}}^\psi} C_{s,s+t}^R({x_s}^\omega, {x_{s+t}}^\psi) log C_{s,s+t}^R({x_s}^\omega, {x_{s+t}}^\psi), \tag{17}$$

This article has been accepted for publication in IEEE Transactions on Dependable and Secure Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2024.3523289

where $C^R_{s,s+t}(x_s{}^\omega, x_{s+t}{}^\psi)$ corresponds to the probability: a pixel with grey level $x_s{}^\omega$ in frame x_s has grey level $x_{s+t}{}^\psi$ in frame x_{s+t}. The total joint entropy is given by:

$$H(x_s, x_{s+t}) = H(x_s, x_{s+t})^R + H(x_s, x_{s+t})^G + H(x_s, x_{s+t})^B. \tag{18}$$

Secondly, we illustrate the distinction between high-visual-quality videos and low-visual-quality videos in terms of definition analyses. The joint entropy $H(x_s, x_{s+t})$ is introduced to measure the uncertainty between x_s and x_{s+t}. The high uncertainty between frames means that there is less information carrying between frames, which could make matrices $C^R_{s,s+t}$, $C^G_{s,s+t}$, $C^B_{s,s+t}$ low. According to Eq. (17), low $C^R_{s,s+t}$, $C^G_{s,s+t}$, $C^B_{s,s+t}$ make the joint entropy high. Therefore, the frame with high uncertainty means that the joint entropy between the frames is high. For low-visual-quality Deepfake videos, the generated faces contain some visible artifacts. These artifacts are produced by the generation process with uncertainties. For example, the artifacts will appear in one part of the current frame but appear in the next frame. Therefore, the frames' artifacts can not be predicted by other frames' artifacts, which enlarges the uncertainty among frames. For high-visual-quality Deepfake videos, the visible artifacts are relieved, and the frames can be predicted without the disturbance of visible artifacts. Therefore, the uncertainty of low-visual-quality fake videos is higher than that of high-visual-quality fake videos. Then, the joint entropy of low-visual-quality Deepfake videos is higher than that of high-visual-quality videos. That is,

$$H(x_s, x_{s+t})^h < H(x_s, x_{s+t})^l, \tag{19}$$

where $H(x_s, x_{s+t})^h$ represents the $H(x_s, x_{s+t})$ of high-visual-quality videos, $H(x_s, x_{s+t})^l$ represents the $H(x_s, x_{s+t})$ of low-visual-quality videos.

Thirdly, we utilize statistical results to explain the joint entropy. Specifically, we randomly selected 200 high-visual-quality videos generated by NeuralTextures [26] and 200 low-visual-quality videos generated by DeepFakes [25] to calculate the inter-frame joint entropy. The face content of these two kinds of videos is about the same, which can avoid the influence of face content on joint entropy. The results in Fig. 3 demonstrate that the joint entropy of low-visual-quality videos is higher than that of high-visual-quality videos.

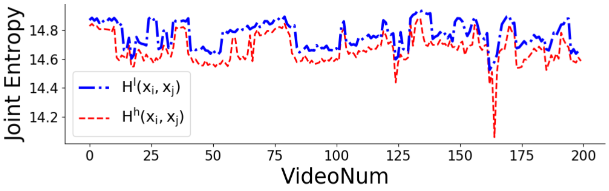


Fig. 3. A joint entropy comparison between the low-visual-quality videos and high-visual-quality videos.

The proposed ADA-FInfer detects high-visual-quality videos by predicting the target frames. The low-visual-quality videos with high $H(x_s, x_{s+t})^l$ and high uncertainty between x_s and x_{s+t} can cause inaccurate predictions. For high-visual-quality videos, the joint entropy $H(x_s, x_{s+t})^l$ between x_{s+t} and x_s is low, which is advantageous to the prediction process. Since ADA-FInfer is a detection method that is based on the prediction, the proposed ADA-FInfer is efficient for detecting videos with high-visual-quality.

4.2 Interpretation for Detecting the Difference between Real Videos and Fake Videos

We analyze the mutual information to illustrate that the proposed ADA-FInfer is effective in distinguishing real videos from Deepfake videos.

For a communication system that takes x_s as the source and x_{s+t} as the receiver, the amount of information that the source x_s obtained from the receiver x_{s+t} can be expressed by the mutual information.

$$MI(x_s, x_{s+t}) = H(x_s) + H(x_{s+t}) - H(x_s, x_{s+t}). \tag{20}$$

If the video is generated by Deepfakes techniques, Eq. (20) can be represented as follows.

$$MI(x_s, x_{s+t})^f = H(x_s)^f + H(x_{s+t})^f - H(x_s, x_{s+t})^f. \tag{21}$$

If the video is a real video, Eq. (20) can be represented as follows.

$$MI(x_s, x_{s+t})^r = H(x_s)^r + H(x_{s+t})^r - H(x_s, x_{s+t})^r. \tag{22}$$

Real faces have rich details, while fake faces lose some of the details [62], [63]. According to Chang et al. [64], the value of the entropy correlates positively with the details. Hence, real frames with rich details get higher entropy than that of fake frames. That is,

$$H(x_i)^r + H(x_j)^r > H(x_i)^f + H(x_j)^f. \tag{23}$$

The generated faces in fake videos are unnatural, which may bring in the inconsistency between x_i and x_j. Then, the uncertainty between x_i and x_j gets larger. According to the definition of entropy, high uncertainty of x_i and x_j means that the value of $H^f(x_i, x_j)$ gets larger. The faces in real videos keep spatio-temporal consistency, and the consistency decreases the uncertainty between x_i and x_j. Thus, the value of $H^r(x_i, x_j)$ gets smaller. That is,

$$H(x_i, x_j)^r < H(x_i, x_j)^f. \tag{24}$$

We combined Eqs. (22, 23, 24). Then, we have,

$$MI(x_s, x_{s+t})^r > MI(x_s, x_{s+t})^f. \tag{25}$$

The mutual information $MI(x_s, x_{s+t})$ is introduced to measure the amount of the information that x_s obtained from x_{s+t}. High $MI(x_s, x_{s+t})$ indicates more information that the source x_s can be obtained from the receiver x_{s+t}, which means there is high feasibility to predict x_{s+t} based on x_s. Eq. (25) reveals that the real videos' $MI(x_s, x_{s+t})^r$ is larger than that of fake videos. It illustrates that the real videos with large mutual information are beneficial for predicting, and the predicted x_{s+t} will be closer to the referenced x_{s+t}. Conversely, for the fake videos, the predicted x_{s+t} will mismatch the referenced x_{s+t}. Therefore, the proposed ADA-FInfer can detect the differences between the real videos and fake videos.

4.3 The Analysis for the Adaptive Frame Selection Strategy

Section 3.3 illustrates that we divide the video based on the optical flow and select frames from the divided video clips. We analyze the proposed adaptive frame selection strategy from two aspects: (1) Explanations for the division way: Dividing the video based on optical flow to select frames with minimal motion, leading to lower temporal entropy and reduced uncertainty, which enhances the prediction process. and (2) Explanations for the length of selected frames: Choosing frames from clips with a median length, balancing correlation between frames to avoid high uncertainty caused by overly long intervals, thereby improving prediction accuracy.

(1) According to Liang et al. [65], the temporal entropy can be expressed as:

$$TE = -\sum_{\tau=s}^{\tau=s+t}(THP \times log_2(THP)). \tag{26}$$

The *THP* can be represented as:

$$THP = \frac{TH(\tau)}{\sum\limits_{\tau=s}^{\tau=s+t} TH(\tau)}, \tag{27}$$

where *TH* represents the total number of moving pixels, and τ represents the time step. Based on Liang et al. [65], a slight change in the movement of the pixels represents a high value of *THP*. Since the Eq. (26) shows that *TE* decreases monotonically with *THP*, a high value of *THP* makes the value of *TE* low. According to the definition of temporal entropy, the low value of temporal entropy represents low uncertainty. Frames with low uncertainty can facilitate the prediction process. Therefore, slight movement in faces should be selected to perform the face representative learning. We implement optical flow to measure the movements. By using the quantified value of movements to divide the video into multiple clips, the movement of clips is guaranteed to be slight, which facilitates the prediction process.

(2) Since we select the video clip with a median length, it is essential to analyze the impact of the length. The proposed ADA-FInfer utilizes source faces to predict target faces. The length of frame selection depends on the length of source faces s and target faces t. Therefore, we analyze the length of the selected frames by analyzing s and t. We express conditional entropy of the R component:

$$\begin{aligned} &H(x_s|x_{s+t})^R = \\ &-\sum_{x_s{}^\omega}\sum_{x_{s+t}{}^\psi} C^R_{s,s+t}(x_s{}^\omega, x_{s+t}{}^\psi) log C^R_{s,s+t}(x_s{}^\omega|x_{s+t}{}^\psi), \end{aligned} \tag{28}$$

where $C^R_{s,s+t}(x_s{}^\omega|x_{s+t}{}^\psi)$ corresponds to the conditional probability: the likelihood of a pixel with grey level $x_s{}^\omega$ in frame x_s occurring, based on the occurrence of a pixel with grey level $x_{s+t}{}^\psi$ in frame x_{s+t}. The total joint entropy is given by:

$$H(x_s|x_{s+t}) = H(x_s|x_{s+t})^R + H(x_s|x_{s+t})^G + H(x_s|x_{s+t})^B. \tag{29}$$

When t is too large, the correlation between x_s and x_{s+t} can be small. Based on the aforementioned definition of conditional probability, a low correlation value makes conditional probabilities $C^R_{s,s+t}(x_s{}^\omega|x_{s+t}{}^\psi)$, $C^G_{s,s+t}(x_s{}^\omega|x_{s+t}{}^\psi)$, $C^B_{s,s+t}(x_s{}^\omega|x_{s+t}{}^\psi)$ low. According to Eq. (28), low conditional probabilities make the conditional entropy high. The high conditional entropy represents high uncertainty. However, high uncertainty can lead to a poor prediction process of ADA-FInfer.

Therefore, the length of t can not be set too large. To reduce the data dimensionality, the length of source frames s shall be limited. In this way, we select frames by choosing the clip with a median length.

5 Experimental Evaluations

In this section, we first provide the experimental settings and datasets. Thereafter, we evaluate the impacts of the s and t and the impacts of the prediction ways. Furthermore, we implement the ablation study. Then, the detection efficiency comparisons, in-dataset detection performance comparisons, and cross-domain detection performance comparisons are provided. Thereafter, we visualize the representations and the *corr* and discuss the threat model. Finally, we provide discussions on several aspects.

5.1 Experimental Settings

For the data preparation of the method, we select frames by utilizing the devised strategy, and we extract frames from each video by using `FFmpeg` [66]. Then, we utilize `dlib` [67] to do the face detection, landmark extraction, and face alignment. Then, the Gaussian-Laplacian pyramid block preprocesses the faces. Thereafter, these faces are input to the faces representative learning module. By constructing faces representative learning module, we obtain the faces representations required for prediction. The learning rate is set as 0.001. When the accuracy plateaus, the learning rate will be divided by 5. We utilize `Adam` [68] optimizer to optimize the model. The default threshold of the binary accuracy algorithm is equal to 0.5. We carry out experiments in `Keras` on NVIDIA Titan Xp. Furthermore, there are both interpretable methods and uninterpretable methods in the comparison experiments.

5.2 Datasets

Four high-visual-quality Deepfake videos datasets, i.e., Celeb-DF dataset [30], WildDeepfake dataset [37], DFDC dataset [38], and Kodf dataset [69] are utilized to show the detection performance of ADA-FInfer. The Celeb-DF dataset contains 5639 high-visual-quality DeepFake videos. These videos are generated by constructing an improved synthesis process. The WildDeepfake dataset is well-made due to a longer training time with lots of high-visual-quality face images. It consists of 7314 face sequences. These faces are extracted from 707 Deepfake videos that are gathered from the Internet. The DFDC dataset is a large-scale Deepfake detection dataset published by Facebook. DFDC dataset achieves high-visual-quality of manipulations by choosing pairs of similar appearances. Kodf Datasets generate Deepfake videos by using six different models [69]. FaceForensics++ [43] dataset is utilized to train the cross-dataset model. FaceForensics++ dataset comprises 1000 original videos and 4 types of fake videos, namely: DeepFakes [25], Face2Face [70], FaceSwap [71], and NeuralTextures [26]. We note that the FaceForensics++ dataset contains videos with low-visual-quality [25], and the proposed method focuses on detecting Deepfake videos with high-visual-quality.

5.3 Adaptive Attack

We explore an adaptive attack method, which strategically forges selective frames through adaptive choice, thereby challenging our detection model. Specifically, the adaptive attack method involves selecting a subset of frames from the original video, denoted as OV, and potentially applying forgeries to these frames. This process can be represented as a function $Att(OV, \phi)$ that maps the video OV and parameters ϕ (which defines the selection and forgery strategy) to a set of frames. The attacker's goal can then be formulated as an optimization problem:

$$\max_{\phi} L^{Att}(m^{para}, Att(OV, \phi), Op^{Att}), \tag{30}$$

where m^{para} represents the model parameters, Op^{Att} represents the prediction output. This expression seeks to find the parameter set ϕ that maximizes the objective function L^{Att}, which in turn depends on how the frames are selected and altered by $Att(OV, \phi)$.

We conduct experiments involving the tampering of partial video frames to demonstrate how the adversary bypasses the detection and show the differences between the conference version named FInfer [1] and the extended version.

To the best of our knowledge, there are currently 2 publicly partially manipulated video datasets: TVIL [72] and Undercover [73] datasets. We conduct a performance comparison between our extended version of the paper, the conference version, and existing methods, using TVIL [72] and Undercover [73] datasets subjected to partial frame tampering. If a model uses a single frame as input, we randomly sample 30 frames for each video to calculate the video-level AUC scores. We also use IoU to evaluate the fit of fake frame detection [73]. The results in Table 3 show that the proposed ADA-FInfer performs better than other methods.

Unlike the conference version of our paper, which employs random frame selection for detection, our extended version introduces a method of adaptive frame selection for this purpose. This enhancement is crucial when attackers employ adaptive attacks, selecting only a subset of frames for forgery instead of tampering with all frames in a video. The random frame selection approach is less effective against such attacks, as it may miss the altered frames. In contrast, our adaptive frame selection method initially screens the frames before proceeding with detection. This preliminary selection increases the likelihood of identifying frames that contain forgeries, especially in videos where only a portion has been tampered with, thereby offering a more robust defence against these types of attacks.

TABLE 3
Comparisons of the in-dataset evaluations between ADA-FInfer and baseline methods on partial fake TVIL [72] and Undercover [73] datasets.

Method	TVIL		Undercover	
	IoU ↑	AUC↑	IoU ↑	AUC↑
SPSL [8]	86.9	85.3	86.2	87.7
NoiseDF [31]	85.4	82.5	83.6	85.2
DisGRL [32]	82.0	83.7	83.9	84.1
STN [33]	84.6	82.2	84.8	84.5
FT-two-stream [9]	85.0	80.6	82.3	83.5
Xia et al. [29]	82.3	80.4	81.7	83.8
Oc-fakedect [35]	79.6	80.2	79.5	82.1
RECCE [34]	84.7	85.1	86.1	85.3
BRCNet [36]	85.4	85.3	86.8	85.8
FInfer [1]	86.9	86.4	87.6	86.7
ADA-FInfer	**92.1**	**92.4**	**92.6**	**90.7**

TABLE 4
The forgery detection ACC (%) for the impacts of prediction ways. CBP represents the way of predicting current faces based on the previous faces. PBC represents the way of predicting previous faces based on the current faces. PFBC represents the way of predicting previous faces and future faces based on current faces.

Prediction Way	Celeb-DF	WildDeepfake	DFDC	Kodf
CBP	**76.1**	**72.4**	**72.6**	**70.7**
PBC	75.8	71.2	71.4	70.3
PFBC	75.7	71.5	71.0	70.1

5.4 Impacts of the Prediction Ways

In the proposed method, the referenced source faces are utilized to predict the representations of predicted target faces. The prediction ways can be divided into 3 categories: predicting **c**urrent faces **b**ased on the **p**revious faces (CBP), predicting **p**revious faces **b**ased on **c**urrent faces (PBC), and predicting **p**revious faces and **f**uture faces **b**ased on **c**urrent faces (PFBC). We provide experiments to evaluate the impacts of prediction ways. Specifically, we apply different prediction ways on Celeb-DF, WildDeepfake, DFDC datasets. The detection performance of the proposed ADA-FInfer is given in Table 4. It illustrates that the prediction ways have a slight impact on the high-visual-quality Deepfake video detection. The slight impact means that ADA-FInfer is flexible in the prediction ways. Since the prediction way of predicting current faces based on the previous faces achieves slightly better than other ways, we use the prediction way of CBP in the following experiments.

5.5 Impacts of the Frame Selection Strategy

In this subsection, we conduct experiments to illustrate the effectiveness of the proposed adaptive selection strategy. There are amount of frames in a video, and the selection of the frames has an effect on the experimental results. The strategy of the conference version [1] randomly selects frames from videos. Nguyen et al. [19] select the first 10 frames of videos for detection. To evaluate the effectiveness of the adaptive selection strategy, we compare the performance of the proposed strategy with the aforementioned selection strategies proposed in Ref. [1], [19], including the specific strategy and random strategy. The specific strategy means that we utilize the first set of frames to evaluate

TABLE 5
The forgery detection ACC (%) for the impacts of choosing different frame selection strategies.

Selection Strategy	Celeb-DF	WildDeepfake	DFDC	Kodf
Specific Strategy	75.5	70.9	70.4	68.2
Random Strategy	75.8	71.7	71.6	68.3
Proposed Strategy	**76.1**	**72.4**	**72.6**	**70.7**

TABLE 6
The forgery detection ACC (%) with and without the Gaussian-Laplacian pyramid block.

Datasets	without pyramid block	with pyramid block
Celeb-DF	72.4	**76.1**
WildDeepfake	68.8	**72.4**
DFDC	69.0	**72.6**
Kodf	68.2	**70.7**

the model. The random strategy means that we randomly extract frames and input them into the network. The results in Table 5 illustrate that the proposed strategy performs better than that of the specific strategy and random strategy.

The specific strategy, with its fixed selection of the first set of frames, does not cope well with the variability of the video. The random strategy is not optimized for selection, and it would have an impact on the detection results when unfavourable video clips are selected. The proposed adaptive selection strategy selects frames according to the video content, which makes the method applicable to various videos. By implementing the devised selection strategy, the faces predictive learning module is able to predict the predicted target faces in a better way and ultimately improve the detection performance. Therefore, we use the adaptive selection strategy to detect Deepfake videos.

5.6 Impacts of the Gaussian-Laplacian Pyramid Block

The Gaussian-Laplacian pyramid block is utilized for face preprocessing. It is important to evaluate the effect of the Gaussian-Laplacian pyramid block on the high-visual-quality Deepfake video detection. Thus, we perform the ablation study of Gaussian-Laplacian pyramid block on the ADA-FInfer. Specifically, we conduct the experiment of the ADA-FInfer without the Gaussian-Laplacian pyramid block, and the results are shown in the first column of Table 6. Thereafter, the second column of Table 6 shows the performance of the ADA-FInfer with the Gaussian-Laplacian pyramid block. It can be observed that the performance of the ADA-FInfer without the Gaussian-Laplacian pyramid block is decreased. That may be because the Gaussian-Laplacian pyramid block can expose the manipulation traces. The ADA-FInfer with the Gaussian-Laplacian pyramid block can suppress high-level face content and thereby amplify artifacts, which is beneficial for representing faces and predicting the faces.

5.7 Detection Efficiency Comparisons

We provide the Multi-Adds analysis of the proposed ADA-FInfer. The calculation process of Multi-Adds is shown as follows.

The multiplication computation $Mult$ of a convolutional layer is shown as follows.

$$Mult = kc \times C_{in} \times M_{out} \times C_{out}. \tag{31}$$

The addition computation Add is calculated as follows.

$$Add = kc \times (C_{in} - 1) \times M_{out} \times C_{out}. \tag{32}$$

The bias computation is Bia and is calculated as follows.

$$Bia = M_{out} \times C_{out}. \tag{33}$$

The Multi-Adds computation M_A of the convolutional layer is obtained by adding up the Eqs. (31, 32, 33), that is,

$$M_A = 2 * (kc \times C_{in}) \times M_{out} \times C_{out}. \tag{34}$$

According to the above equations, we calculate the Multi-Adds of the proposed ADA-FInfer to be 96.8×10^6. Besides, the Multi-Adds of the state-of-the-art methods are shown in Table 7. It shows that the Multi-Adds of ADA-FInfer are lower than that of other methods. The architecture of ADA-FInfer is lightweight, which promotes high detection efficiency. We note that the detection efficiency of ADA-FInfer is slightly lower than that of FInfer. The reason is that the adaptive frame selection strategy of ADA-FInfer increases the complexity of the algorithm and affects the detection efficiency. Compared to the Multi-Adds of 96.8×10^6, the impact of adaptive frame selection on efficiency is relatively minor.

5.8 In-Dataset Detection Performance Comparisons

We compare ADA-FInfer with the baseline methods on the high-visual-quality Deepfake videos datasets: Celeb-DF dataset [30], WildDeepfake dataset [37], DFDC, and Kodf dataset [38]. We sample 20 frames for each video to calculate the frame-level detection performance. To overall evaluate the ADA-FInfer, we use AUC to measure performance. The maximum value of the AUC score is 1. A high AUC score corresponds to excellent classification performance. The advantage of AUC is that it is insensitive to sample distribution and can describe the overall performance of the classification algorithm [74]. Table 8 shows the AUC results of ADA-FInfer and other baseline methods for high-visual-quality Deepfake video detection. ADA-FInfer achieves slight advantages in terms of average.

ADA-FInfer builds upon and extends the technical strategy of FInfer. The results in Table 8 show that ADA-FInfer achieves better in-dataset detection performance compared to FInfer, demonstrating the effectiveness of the extended technical strategy in enhancing detection accuracy.

5.9 Cross-Domain Detection Performance Comparisons

To evaluate the transferability of ADA-FInfer, we simulate cross-domain Deepfake detection in different scenarios: cross-dataset detection and cross-manipulation detection.

First, we conduct the cross-dataset experiments on high-visual-quality datasets, i.e., Celeb-DF, WildDeepfake, DFDC, and Kodf. We follow the setting of Ref. [8], [20] and compare ADA-FInfer with baseline methods. The cross-dataset model is trained on Faceforensics++, but tested on Celeb-DF, WildDeepfake, DFDC, and Kodf datasets. Results of some methods are evaluated in [20]. The cross-dataset results in Table 9 illustrate that the detection performance of most methods decreases significantly in the cross-dataset experiments, but ADA-FInfer shows satisfied transferability on detecting high-visual-quality Deepfake videos. For example, Xie et al. [29] achieve the best performance in the in-dataset detection of WildDeepfake, but the cross-dataset detection performance of WildDeepfake is decreased.

SPSL [8] integrates phase maps with spatial features and reaches the best performance on Celeb-DF, and ADA-FInfer achieves the best performance on average. SPSL [8] and

TABLE 7
Comparisons of the number of Multi-Adds $(\times 10^6)$ of the network between ADA-FInfer and baseline methods.

	SPSL [8]	NoiseDF [31]	DisGRL [32]	STN [33]	FT-two-stream [9]	Xia et al. [29]	Oc-fakedect [35]	RECCE [34]	BRCNet [36]	ADA-FInfer
Multi-Adds	408.8	150.1	552.1	1651.2	362.8	784.3	5731.7	6207.9	1450.3	**96.8**

TABLE 8
Comparisons of the in-dataset evaluation (AUC (%)) between ADA-FInfer and baseline methods on Celeb-DF, WildDeepfake, DFDC, and Kodf datasets.

Method	Celeb-DF	WildDeepfake	DFDC	Kodf	Avg
Meso4 [18]	66.2	66.5	76.5	78.9	72.0
FWA [6]	60.2	57.9	73.0	76.1	66.8
Xception [43]	89.8	86.8	81.6	90.2	87.1
ADDNet [37]	95.2	86.2	79.7	85.3	86.6
FT-two-stream [9]	86.7	68.1	64.0	82.3	75.3
Xia et al. [29]	95.8	**86.9**	85.1	91.4	89.8
FInfer [1]	93.3	81.4	82.9	92.8	87.6
ADA-FInfer	**96.5**	85.6	**86.0**	**93.7**	**90.5**

TABLE 9
Comparisons of the cross-dataset evaluation (AUC (%)) between ADA-FInfer and baseline methods on Celeb-DF, WildDeepfake, DFDC, and Kodf datasets.

Method	Celeb-DF	WildDeepfake	DFDC	Kodf	Avg
SPSL [8]	**76.9**	70.3	66.2	69.7	70.8
NoiseDF [31]	75.9	62.5	63.9	67.2	67.4
DisGRL [32]	70.0	66.7	70.9	69.1	69.2
STN [33]	67.6	62.1	64.8	64.5	64.8
FT-two-stream [9]	65.6	59.8	59.1	65.1	62.4
Xia et al. [29]	52.2	68.7	66.3	66.4	63.4
Oc-fakedect [35]	66.3	62.2	68.0	65.4	65.5
RECCE [34]	68.7	64.3	69.1	67.3	67.4
BRCNet [36]	70.9	68.3	69.8	66.8	69.0
Ours	76.1	**72.4**	**72.6**	**70.7**	**73.0**

TABLE 10
Comparisons of the cross-manipulation evaluation (AUC (%)) between ADA-FInfer and baseline methods on each forgery type of FaceForensics++ when trained on one type.

Method	Train	F2F	FS	NT	Ave
Freq-SCL [42]	DF	58.9	66.9	63.6	63.1
MultiAtt [20]		66.4	67.3	66.0	66.6
RECCE [34]		**70.7**	74.3	**67.3**	**70.8**
FInfer [1]		69.4	74.2	66.8	70.1
ADA-FInfer		70.6	**74.4**	67.0	70.7
Method	Train	DF	FS	NT	Ave
Freq-SCL [42]	F2F	67.6	55.4	66.7	63.2
MultiAtt [20]		73.0	65.1	71.9	70.0
RECCE [34]		76.0	64.5	**72.3**	70.9
FInfer [1]		76.3	64.4	69.8	70.2
ADA-FInfer		**77.9**	**65.4**	70.1	**71.1**
Method	Train	DF	F2F	NT	Ave
Freq-SCL [42]	FS	75.9	54.6	49.7	60.1
MultiAtt [20]		82.3	61.7	54.8	66.3
RECCE [34]		82.4	**64.4**	**56.7**	**67.8**
FInfer [1]		82.0	62.4	55.6	66.7
ADA-FInfer		**83.1**	62.9	56.3	67.4
Method	Train	DF	F2F	FS	Ave
Freq-SCL [42]	NT	79.1	74.2	54.0	69.1
MultiAtt [20]		74.6	80.6	60.9	72.0
RECCE [34]		78.8	**80.9**	63.7	74.5
FInfer [1]		79.6	75.8	64.7	73.4
ADA-FInfer		**80.6**	78.4	**65.4**	**74.8**

NoiseDF [31] ignore the temporal features, which results in performance degradation in detecting fake videos with temporal forgery traces. The Celeb-DF dataset predominantly features spatial forgery traces, whereas WildDeepfake and DFDC contain lots of temporal forgery traces. This distinction leads to SPSL [8] and NoiseDF [31] performing well in detecting Celeb-DF, but losing their advantages in detecting WildDeepfake and DFDC.

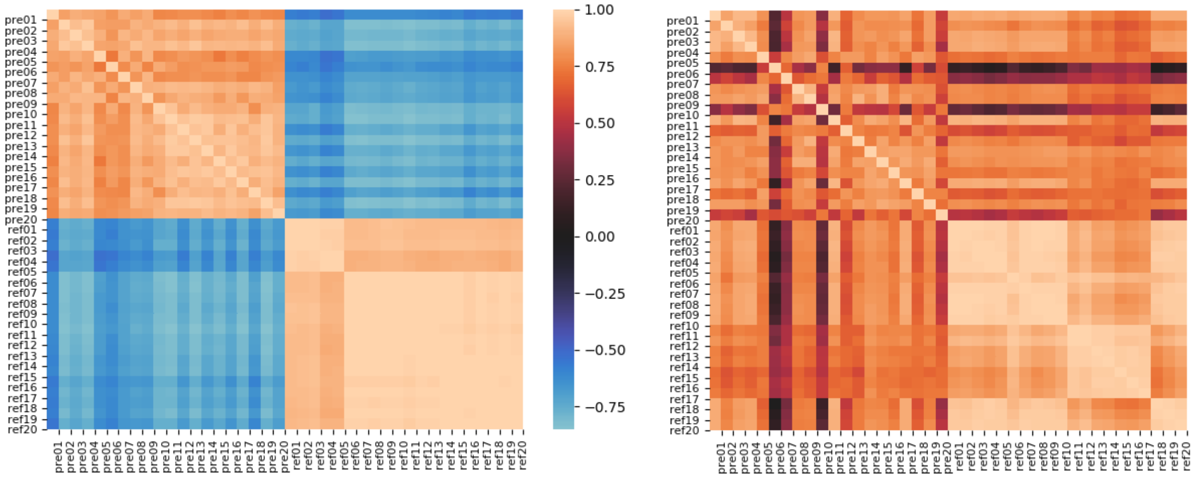
ADA-FInfer detects the high-visual-quality Deepfake videos by incorporating predictions into the training process. When detecting high-visual-quality videos, ADA-FInfer compares the predicted target representations with the referenced target representations instead of extracting features directly from the frames, which significantly benefits the model's robustness. Besides, the adaptive frame selection strategy selects frames based on video content, which facilitates the model to adapt to different types of videos. Furthermore, by utilizing the adaptive frame selection strategy, ADA-FInfer achieves better cross-domain detection performance than that of FInfer.

Secondly, we carry out cross-manipulation experiments to assess the transferability. There are 4 types of videos in FaceForensics++, which are generated from 4 forgery technologies: DeepFakes (DF), Face2Face (F2F), FaceSwap (FS), and NeuralTextures (NT). Following RECCE [34], we select each type of video for training and the remaining three types of videos for testing. Results in Table 10 illustrate that ADA-FInfer outperforms current methods in many scenarios. When the model is trained on DeepFakes and tested on the other three types of videos, ADA-FInfer achieves the best detection performance. When training on Face2Face and NeuralTextures, ADA-FInfer shows the best transferability on average. However, our method has limitations in the scenario of training on DeepFakes and FaceSwap. When generating videos, DeepFakes and FaceSwap alter the whole face area leaving swap artifacts, but Face2Face and NeuralTextures alter parts of the face area leaving invisible artifacts and exhibiting a high visual quality. As the analyses in Section 4.1, ADA-FInfer is efficient for detecting videos with high-visual-quality. Since the visual quality of videos generated by Face2Face and NeuralTextures is higher than that of videos generated by DeepFakes and FaceSwap, ADA-FInfer can efficiently build the detection model on Face2Face and NeuralTextures. Therefore, the performance of training on Face2Face and NeuralTextures is better than that of training on DeepFakes and FaceSwap.

5.10 Visualizations of Representations and corr

To illustrate the importance of exploring the correlation between predicted target face representations and referenced target face representations, the correlation matrix is visualized through heatmaps. We input high-visual-quality Deepfake videos and real videos into the model. Then,

This article has been accepted for publication in IEEE Transactions on Dependable and Secure Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TDSC.2024.3523289



(a) Fake

(b) Real

Fig. 4. Heatmaps for the correlation matrix between the predicted target representations and the referenced target representations. The predicted representations and referenced representations are outputted by the second last layer and third last layer of ADA-FInfer, respectively. The different colors between the fake video (a) and the real video (b) represent the different correlation values.

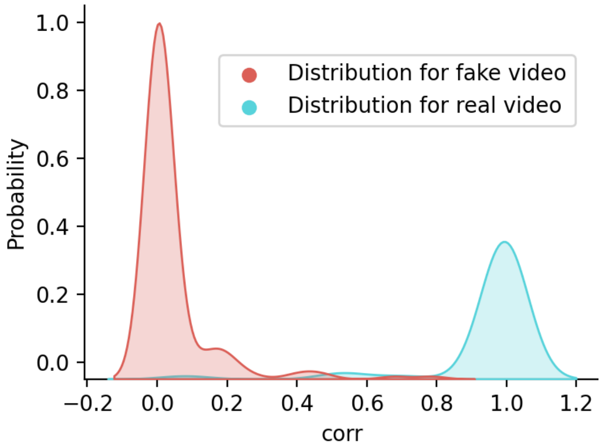


Fig. 5. The probability distribution of *corr* between fake (red) and real (green) videos. The *corr* scores are outputted by the softmax layer of ADA-FInfer. It shows that real videos and fake videos are distributed in different areas.

we output the second last layer and the third last layer of ADA-FInfer. Specifically, the output of the second last layer is the predicted target face representations with the dimensionality of $128 * 1$. The output of the third last layer is the referenced target face representations with the dimensionality of $128 * 1$. As shown in Fig. 4, the first 20 of the coordinates are the predicted target face representations, and the last 20 of the coordinates are the referenced target face representations. The lighter cell represents the higher correlation. For the fake videos, the blue cells represent that the correlation between the predicted target face representations and referenced target face representations is low. For the real videos, the orange cells represent that there is a high correlation between the predicted target face representations and referenced target face representations. Therefore, it can be concluded that ADA-FInfer can effectively detect high-visual-quality Deepfake videos by inferring frames and comparing the correlation between the predicted target face representations and referenced target face representations.

The proposed method learns the facial variation rules of facial expressions by making predictions and comparing the *corr*. We plot the probability distribution of *corr*. As shown in Fig. 5, X-axis represents the continuous desirable value of *corr*. The value of *corr* is between 0 and 1. The minimum value of the X-axis is slightly smaller than the minimum value of *corr*, and the maximum value of the X-axis is slightly larger than the maximum value of *corr*. The Y-axis represents the probability of *corr*. Fig. 5 illustrates that the distribution of fake videos is different from real videos. Most of the fake videos are with low *corr*, but most real videos are with high *corr*. That is, ADA-FInfer can detect videos by comparing the *corr*.

5.11 Threat model.

To counter advanced Deepfake attacks, we devise an adaptive frame selection strategy specifically designed to address attacks involving partial frame manipulation. This strategy enhances the robustness of our detection method against sophisticated Deepfake techniques.

Our threat model assumes that attackers leverage advanced deep learning methods to generate high-visual-quality Deepfake videos for purposes such as misinformation, defamation, or bypassing facial recognition systems. We account for the sophistication of various Deepfake generation techniques, including DeepFakes, Face2Face, FaceSwap, NeuralTextures, and improved DeepFakes found in datasets like FF++, Celeb-DF, DFDC, and WildDeepfake, as well as the diverse sources they may exploit, such as publicly available videos in the WildDeepfake dataset.

Defenders require advanced detection techniques capable of extracting robust features to identify unseen patterns produced by emerging Deepfake technologies. To evaluate this, we simulate detection in unseen datasets through cross-dataset and cross-manipulation experiments. As demonstrated in Tables 9 and 10, our proposed method surpasses existing approaches in detecting previously unseen datasets.

Furthermore, defenders need access to diverse datasets that cover a variety of Deepfake videos to train models capable of identifying a broad range of manipulation techniques. Our experiments involve five publicly available datasets to address this need. Finally, defenders must detect fake videos swiftly and efficiently, minimizing computational overhead.

TABLE 11
The detection ACC (%) of ADA-FInfer and baseline methods on FaceForensics++ dataset.

Datasets	Mesonet [18]	FWA [6]	Xception [43]	ADDNet [37]	FT-two-stream [9]	Xia et al. [29]	FInfer [1]	ADA-FInfer
FaceForensics++	89.0	80.1	97.3	94.1	90.6	96.2	94.4	**97.4**

Our method, ADA-FInfer, achieves this with fewer Multi-Adds than current alternatives.

5.12 Discussions

Low-Visual-Quality Deepfake detection. Limited by technology and hardware resources, the inchoate datasets named FF++ are with low visual quality, which has perceptible distortions, such as jitter, blur, and strange artifact [37]. As shown in Table 11, when detecting low-visual-quality FF++ datasets, Xception [43] achieved a suboptimal detection performance of 97.3%. ADA-FInfer achieved 97.4% on the FF++ dataset, reflecting a 0.1% improvement.

The influence of ζ. We use $\zeta = 0.3, 0.4, 0.5, 0.6, 0.7$ to conduct the adaptive frame selection strategy and detect videos. As shown in Table 12, when setting $\zeta = 0.3, 0.4, 0.6, 0.7$, there is a slight drop in performance. One possible reason is that, when considering these values of ζ, optimal video frame segments were not selected, thereby affecting the video detection process. Overall, ζ has an impact on performance, so we use the default value 0.5.

TABLE 12
The cross-dataset forgery detection AUC (%) in different values of ζ.

Datasets	$\zeta = 0.3$	$\zeta = 0.4$	$\zeta = 0.5$	$\zeta = 0.6$	$\zeta = 0.7$
Celeb-DF	75.4	75.8	**76.1**	75.7	75.3
WildDeepfake	71.8	72.1	**72.4**	72.0	71.6
DFDC	71.6	72.1	**72.6**	71.9	71.1
Kodf	69.8	70.4	**70.7**	70.1	69.4
Avg	72.2	72.6	**73.0**	72.4	71.9

TABLE 13
The processing time of each phase of ADA-FInfer (on FF++). The h represents the hour. The s represents the second.

Phase	Processing Time
Frame Extraction	$5h$
Adaptive Frame Selection	$2h$
End-to-End Training Time	$0.2h$
Testing Time	$4s$

The processing time of each phase. We conduct experiments to measure the processing time for each phase of our ADA-FInfer model. The results have been included in the revised manuscript and are shown in Table 13. The FF++ dataset contains 4000 videos, with the frame extraction phase taking 5 hours and the adaptive frame selection phase taking 2 hours. When the model converges, the training phase takes 0.2 hours. The testing phase of each video takes 4 seconds. These results indicate that our proposed ADA-FInfer achieves acceptable processing times.

6 Conclusions

The increasing sophistication of high-visual-quality deepfakes necessitates the development of interpretable detection methods. In this paper, we introduce ADA-FInfer, an approach that infers face representations from adaptively selected frames, enhancing the interpretability of high-visual-quality deepfake detection. By employing an adaptive frame selection strategy, our method improves detection performance in unknown domains and offers resilience against adaptive attacks targeting partial frame manipulation. We further provided interpretability analyses using joint entropy, mutual information, and conditional entropy to explain the differences between real and fake videos, as well as the rationale behind the adaptive selection strategy. These analyses demonstrate the effectiveness of ADA-FInfer. Experimental results on high-visual-quality datasets confirm the method's generalizability and efficiency. Despite its strengths, ADA-FInfer shows limitations when trained on datasets like DeepFakes and FaceSwap but tested on Face2Face and NeuralTextures. In future work, we aim to further explore the representation and prediction processes, enabling self-supervised learning to improve detection performance and address the limitations.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Nos. U22A2030, U20A20174, 61972142), the National Key Research and Development Program of China (No. 2022YFB3103500), the Hunan Provincial Funds for Distinguished Young Scholars (No. 2024JJ2025). Juan Hu thanks the support of the Ministry of Education, Singapore, under its MOE AcRF TIER 3 (Grant MOE-MOET32022-0001).

References

[1] J. Hu, X. Liao, J. Liang, W. Zhou, and Z. Qin, "FInfer: Frame inference-based Deepfake detection for high-visual-quality videos," in *Proc. of AAAI*, 2022, pp. 951–959.

[2] Y. Li, M. Chang, and S. Lyu, "In ictu oculi: Exposing AI created fake videos by detecting eye blinking," in *Proc. of IEEE WIFS*, 2018, pp. 1–7.

[3] U. A. Ciftci, I. Demir, and L. Yin, "Fakecatcher: Detection of synthetic portrait videos using biological signals," *IEEE TPAMI*, 2020, DOI: 10.1109/TPAMI.2020.3009287.

[4] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in *Proc. of IEEE ICASSP*, 2019, pp. 8261–8265.

[5] M. Koopman, A. M. Rodriguez, and Z. Geradts, "Detection of Deepfake video manipulation," in *Proc. of IMVIP*, 2018, pp. 133–136.

[6] Y. Li and S. Lyu, "Exposing Deepfake videos by detecting face warping artifacts," in *Proc. of IEEE/CVF CVPR Workshops.*, 2019, pp. 46–52.

[7] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. of IEEE AVSS*, 2018, pp. 1–6.

[8] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, “Spatial-phase shallow learning: rethinking face forgery detection in frequency domain,” in *Proc. of IEEE/CVF CVPR*, 2021, pp. 772–781.

[9] J. Hu, X. Liao, W. Wang, and Z. Qin, “Detecting compressed deepfake videos in social networks using frame-temporality two-stream convolutional network,” *IEEE TCSVT*, vol. 32, no. 3, pp. 1089–1102, 2021.

[10] https://www.thehindu.com/sci-tech/technology/the-danger-of-deepfakes/article66327991.ece.

[11] H. Etienne, “The future of online trust (and why deepfake is advancing it),” *AI and Ethics*, vol. 1, no. 4, pp. 553–562, 2021.

[12] https://www.sfchronicle.com/opinion/openforum/article/medical-deepfakes-cyberattacks-17649573.php.

[13] G. Shan, B. Zhao, J. R. Clavin, H. Zhang, and S. Duan, “Poligraph: Intrusion-tolerant and distributed fake news detection system,” *IEEE TIFS*, vol. 17, pp. 28–41, 2022.

[14] S. Samtani, H. Chen, M. Kantarcioglu, and B. Thuraisingham, “Explainable artificial intelligence for cyber threat intelligence (XAI-CTI),” *IEEE TDSC*, vol. 19, no. 04, pp. 2149–2150, 2022.

[15] C. Wang and H. Zhu, “Representing fine-grained co-occurrences for behavior-based fraud detection in online payment services,” *IEEE TDSC*, vol. 19, no. 1, pp. 301–315, 2022.

[16] D. Cole, S. Newman, and D. Lin, “A new facial authentication pitfall and remedy in web services,” *IEEE TDSC*, vol. 19, no. 4, pp. 2635–2647, 2022.

[17] S. Hosseini Moghaddam and M. Abbaspour, “Friendship preference: Scalable and robust category of features for social bot detection,” *IEEE TDSC*, vol. 20, no. 2, pp. 1516–1528, 2023.

[18] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “Mesonet: a compact facial video forgery detection network,” in *Proc. of IEEE WIFS*, 2018, pp. 1–7.

[19] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” in *Proc. of IEEE ICASSP*, 2019, pp. 2307–2311.

[20] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, “Multi-attentional deepfake detection,” in *Proc. of IEEE/CVF CVPR*, 2021, pp. 2185–2194.

[21] K. Jayakumar and N. Skandhakumar, “A visually interpretable forensic deepfake detection tool using anchors,” in *IEEE ICITR*, 2022, DOI: 10.1109/ICITR57877.2022.9993294.

[22] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” *arXiv preprint arXiv:1702.08608*, 2017.

[23] S. Rajaraman, S. Candemir, I. Kim, G. Thoma, and S. Antani, “Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs,” *Applied Sciences*, vol. 8, no. 10, p. 1715, 2018.

[24] https://nij.ojp.gov/topics/articles/slow-steady-march-towards-more-reliable-forensic-science.

[25] DeepFakes, “Accessed october 10, 2018,” https://github.com/deepfakes/faceswap, 2018.

[26] J. Thies, M. Zollhöfer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures,” *ACM TOG*, vol. 38, no. 4, pp. 1–12, 2019.

[27] B. Malolan, A. Parekh, and F. Kazi, “Explainable deep-fake detection using visual interpretability methods,” in *Prof. of ICICT*. IEEE, 2020, pp. 289–293.

[28] L. Trinh, M. Tsang, S. Rambhatla, and Y. Liu, “Interpretable and trustworthy deepfake detection via dynamic prototypes,” in *Prof. of WACV*, 2021, pp. 1973–1983.

[29] Z. Xia, T. Qiao, M. Xu, N. Zheng, and S. Xie, “Towards deepfake video forensics based on facial textural disparities in multi-color channels,” *INS*, vol. 607, pp. 654–669, 2022.

[30] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-DF: A large-scale challenging dataset for deepfake forensics,” in *Proc. of IEEE/CVF CVPR*, 2020, pp. 3207–3216.

[31] T. Wang and K. P. Chow, “Noise based deepfake detection via multi-head relative-interaction,” in *Proc. of AAAI*, vol. 37, no. 12, 2023, pp. 14 548–14 556.

[32] Z. Shi, H. Chen, L. Chen, and D. Zhang, “Discrepancy-guided reconstruction learning for image forgery detection,” in *Proc. of IJCAI*, 2023.

[33] K. Lin, W. Han, S. Li, Z. Gu, H. Zhao, and Y. Mei, “Detecting deepfake videos using spatiotemporal trident network,” *ACM TMCCA*, 2023.

[34] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, “End-to-end reconstruction-classification learning for face forgery detection,” in *Proc. of IEEE/CVF CVPR*, 2022, pp. 4113–4122.

[35] H. Khalid and S. S. Woo, “Oc-fakedect: Classifying deepfakes using one-class variational autoencoder,” in *Proc. of CVPR Workshops*, 2020, pp. 656–657.

[36] D. Zhang, C. Fu, D. Lu, J. Li, and Y. Zhang, “Bi-source reconstruction based classification network for face forgery video detection,” *IEEE TCSVT*, 2023.

[37] B. Zi, M. Chang, J. Chen, X. Ma, and Y.-G. Jiang, “WildDeepfake: A challenging real-world dataset for deepfake detection,” in *Proc. of ACM MM*, 2020, pp. 2382–2390.

[38] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, “The deepfake detection challenge (DFDC) dataset,” *arXiv preprint arXiv:2006.07397*, 2020.

[39] J. Liang, Z. Qin, S. Xiao, L. Ou, and X. Lin, “Efficient and secure decision tree classification for cloud-assisted online diagnosis services,” *IEEE TDSC*, vol. 18, no. 4, pp. 1632–1644, 2021.

[40] J. Hernandez-Ortega, R. Tolosana, J. Fiérrez, and A. Morales, “Deepfakeson-Phys: Deepfakes detection based on heart rate estimation,” in *Proc. of AAAI*, 2021, pp. 2638–2646.

[41] M. Chen, X. Liao, and M. Wu, “PulseEdit: Editing physiological signals in facial videos for privacy protection,” *IEEE TIFS*, vol. 17, pp. 457–471, 2022.

[42] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang, “Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection,” in *Proc. of IEEE/CVF CVPR*, 2021, pp. 6458–6467.

[43] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *Proc. of IEEE ICCV*, 2019, pp. 1–11.

[44] Z. Xu, J. Liu, W. Lu, B. Xu, X. Zhao, B. Li, and J. Huang, “Detecting facial manipulated videos based on set convolutional neural networks,” *Journal of Visual Communication and Image Representation*, vol. 77, pp. 103–119, 2021.

[45] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, “Local relation learning for face forgery detection,” in *Proc. of AAAI*, 2021, pp. 1081–1088.

[46] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, “A machine learning based approach for deepfake detection in social media through key video frame extraction,” *SN Computer Science*, vol. 2, no. 2, pp. 1–18, 2021.

[47] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proc. of ICML*, 2019, pp. 6105–6114.

[48] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proc. of IEEE/CVF CVPR*, 2016, pp. 2818–2826.

[49] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao, “Thinking in frequency: Face forgery detection by mining frequency-aware clues,” in *Proc. of ECCV*, 2020, pp. 86–103.

[50] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, “Two-branch recurrent network for isolating deepfakes in videos,” in *Proc. of ECCV*, 2020, pp. 667–684.

[51] Z. Zhao, P. Wang, and W. Lu, “Detecting deepfake video by learning two-level features with two-stream convolutional neural network,” in *Proc. of ICCAI*, 2020, pp. 291–297.

[52] W. Byeon, Q. Wang, R. K. Srivastava, and P. Koumoutsakos, “ContextVP: Fully context-aware video prediction,” in *Proc. of ECCV*, 2018, pp. 753–769.

[53] M. Oliu, J. Selva, and S. Escalera, “Folded recurrent neural networks for future video prediction,” in *Proc. of ECCV*, 2018, pp. 716–731.

[54] C. Lu, M. Hirsch, and B. Scholkopf, “Flexible spatio-temporal networks for video prediction,” in *Proc. of IEEE/CVF CVPR*, 2017, pp. 6523–6531.

[55] X. Chen, W. Wang, J. Wang, and W. Li, “Learning object-centric transformation for video prediction,” in *Proc. of ACM MM*, 2017, pp. 1503–1512.

[56] J.-T. Hsieh, B. Liu, D.-A. Huang, L. Fei-Fei, and J. C. Niebles, “Learning to decompose and disentangle representations for video prediction,” in *Proc. of NeurIPS*, 2018, pp. 515–524.

[57] A. v. d. Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.

[58] B. D. Lucas, T. Kanade *et al.*, "An iterative image registration technique with an application to stereo vision," in *Proc. of IJCAI*, 1981, pp. 674–679.

[59] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. of ICML*, 2015, pp. 448–456.

[60] N. R. Pal and S. K. Pal, "Entropy: A new definition and its applications," *IEEE TSMC*, vol. 21, no. 5, pp. 1260–1270, 1991.

[61] Z. Cernekova, C. Nikou, and I. Pitas, "Shot detection in video sequences using entropy based metrics," in *Proc. of. ICIP*, 2002, pp. 421–424.

[62] Z. Liu, X. Qi, and P. H. Torr, "Global texture enhancement for fake face detection in the wild," in *Proc. of IEEE/CVF CVPR*, 2020, pp. 8060–8069.

[63] H. Chen, G. Hu, Z. Lei, Y. Chen, N. M. Robertson, and S. Z. Li, "Attention-based two-stream convolutional networks for face spoofing detection," *IEEE TIFS*, vol. 15, pp. 578–593, 2019.

[64] C.-I. Chang, K. Chen, J. Wang, and M. L. Althouse, "A relative entropy-based approach to image thresholding," *Pattern recognition*, vol. 27, no. 9, pp. 1275–1289, 1994.

[65] C.-W. Liang and C.-F. Juang, "Moving object classification using a combination of static appearance features and spatial and temporal entropy values of optical flows," *IEEE TITS*, vol. 16, no. 6, pp. 3453–3464, 2015.

[66] X. Lei, X. Jiang, and C. Wang, "Design and implementation of a real-time video stream analysis system based on FFMPEG," in *IEEE Fourth World Congress on Software Engineering*, 2013, pp. 212–216.

[67] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, no. 60, pp. 1755–1758, 2009.

[68] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[69] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, "Kodf: A large-scale korean deepfake detection dataset," in *Proc. of CVPR*, 2021, pp. 10 744–10 753.

[70] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in *Proc. of IEEE/CVF CVPR*, 2018, pp. 2387–2395.

[71] FaceSwap, "Accessed october 29, 2018," https://github.com/MarekKowalski/FaceSwap/, 2018.

[72] R. Zhang, H. Wang, M. Du, H. Liu, Y. Zhou, and Q. Zeng, "Ummaformer: A universal multimodal-adaptive transformer framework for temporal forgery localization," in *Proc. of ACM MM*, 2023, pp. 8749–8759.

[73] S. Saha, R. Perera, S. Seneviratne, T. Malepathirana, S. Rasnayaka, D. Geethika, T. Sim, and S. Halgamuge, "Undercover deepfakes: Detecting fake segments in videos," in *Proc. of ICCV Workshops*, 2023, pp. 415–425.

[74] S. Rosset, "Model selection via the AUC," in *Proc. of ICML*, 2004, pp. 89–96.

Juan Hu is a Ph.D. candidate in the College of Computer Science and Electronic Engineering, Hunan University, China, where she received her M.S. degree in 2019. She received the B.S. degree in the College of Electronic Information Science and Technology from Nanjing Agricultural University, China, in 2017. She is a visiting PhD student at the National University of Singapore from March 2022 to March 2023. Her current research interests include multimedia forensics and artificial intelligence.

Jinwen Liang is a Postdoctoral fellow in the Department of Computing of the Hong Kong Polytechnic University. He received his Ph.D. degree and B.S. degree from Hunan University, China, in 2021 and 2015, respectively. From 2018 to 2020, he was a visiting Ph.D. student at BBCR Lab, University of Waterloo, Canada. His research interests include applied cryptography, AI security, blockchain, and database security. He served as the Technical Program Committee Chair of the 1st international workshop on Future Mobile Computing and Networking for Internet of Things (IEEE FMobile 2022), Publicity Chair of the 6th International Workshop on Cyberspace Security (IWCSS 2022), TPC Member of IEEE VTC' 19 Fall. He is a member of the IEEE.

Zheng Qin received the Ph.D. degree in computer software and theory from Chongqing University, China, in 2001. From 2010 to 2011, he served as a Visiting Scholar at the Department of Computer Science, Michigan University. He is a professor in the College of Computer Science and Electronic Engineering, Hunan University, where he serves as the vice dean. He also serves as the director of Hunan Key Laboratory of Big Data Research and Application, the vice director of Hunan Engineering Laboratory of Authentication and Data Security. He is a member of China Computer Federation (CCF) and IEEE, respectively. His main interests are network and data security, privacy, data analytics and applications, machine learning, and applied cryptography.

Xin Liao received the B.S. and Ph.D. degrees in information security from the Beijing University of Posts and Telecommunications, in 2007 and 2012, respectively. He worked as a postdoctoral fellow at the Institute of Software, Chinese Academy of Sciences, and also a research associate at The University of Hong Kong. From 2016 to 2017, he was a visiting scholar at the University of Maryland, College Park, USA. He is currently a professor in the College of Computer Science and Electronic Engineering, Hunan University. He is serving as the Associate Editor for IEEE Signal Processing Magazine. He is also a member of Technical Committee (TC) on Multimedia Security and Forensics of Asia Pacific Signal and Information Processing Association, TC on Computer Forensics of Chinese Institute of Electronics, and TC on Digital Forensics and Security of China Society of Image and Graphics. His current research interests include multimedia forensics, computer vision, and image/video analysis. He is a senior member of IEEE.

Wenbo Zhou received his B.S. degree in 2014 from Nanjing University of Aeronautics and Astronautics, China, and Ph. D degree in 2019 from University of Science and Technology of China, where he is currently an Assistant Professor. His research interests are Deepfake Generation and Detection, AI Security, and Information Hiding.

Xiaodong Lin received his Ph.D. degree in information engineering from Beijing University of Posts and Telecommunications, China, and his Ph.D. degree in electrical and computer engineering from the University of Waterloo, Canada. He is currently a professor in the School of Computer Science at the University of Guelph, Canada. His research interests include wireless network security, applied cryptography, computer forensics, and software security. He is a fellow of the IEEE.