FInfer: Frame Inference-Based Deepfake Detection for High-Visual-Quality Videos

Juan Hu[1], **Xin Liao**[1*], **Jinwen Liang**[1], **Wenbo Zhou**[2], **and Zheng Qin**[1]

[1]The College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China.
[2]CAS Key Laboratory of Electromagnetic Space Information, University of Science and Technology of China.

Abstract

Deepfake has ignited hot research interests in both academia and industry due to its potential security threats. Many countermeasures have been proposed to mitigate such risks. Current Deepfake detection methods achieve superior performances in dealing with low-visual-quality Deepfake media which can be distinguished by the obvious visual artifacts. However, with the development of deep generative models, the realism of Deepfake media has been significantly improved and becomes tough challenging to current detection models. In this paper, we propose a frame inference-based detection framework (FInfer) to solve the problem of high-visual-quality Deepfake detection. Specifically, we first learn the referenced representations of the current and future frames' faces. Then, the current frames' facial representations are utilized to predict the future frames' facial representations by using an autoregressive model. Finally, a representation-prediction loss is devised to maximize the discriminability of real videos and fake videos. We demonstrate the effectiveness of our FInfer framework through information theory analyses. The entropy and mutual information analyses indicate the correlation between the predicted representations and referenced representations in real videos is higher than that of high-visual-quality Deepfake videos. Extensive experiments demonstrate the performance of our method is promising in terms of in-dataset detection performance, detection efficiency, and cross-dataset detection performance in high-visual-quality Deepfake videos.

Introduction

Motivations. The proliferation of artificial intelligence has given rise to various human portrait video tampering technologies, such as DeepFakes (DeepFakes 2018), Face2Face (Thies et al. 2018), FaceSwap (FaceSwap 2018), and NeuralTextures (Thies, Zollhöfer, and Nießner 2019). Although these technologies facilitate entertainment and cultural exchanges, abusing Deepfake technologies also brings potential threats and concerns to everyone. For example, illegal information such as fake news and manipulated pornographic videos may cause a profound distrust in society, threaten national and political security, and violate individual rights and interests. (Whittaker et al. 2020).

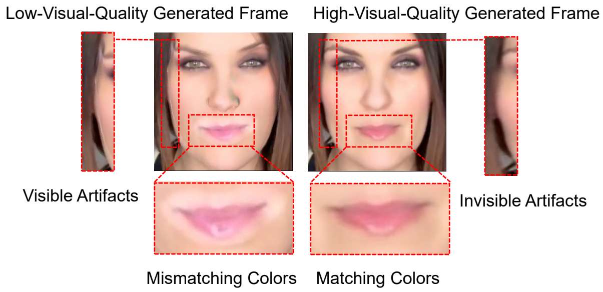
Low-Visual-Quality Generated Frame
High-Visual-Quality Generated Frame
Visible Artifacts
Invisible Artifacts
Mismatching Colors
Matching Colors

Figure 1: Comparisons of the Deepfake frames with different visual quality. The low-visual-quality frame (left) generated by the DeepFakes (DeepFakes 2018) leaves visible artifacts and color mismatches. The high-visual-quality frame (right) generated by the NeuralTextures (Thies, Zollhöfer, and Nießner 2019) reduces the artifacts and color mismatch.

Limited by technology and hardware resources, most of the fake videos in some inchoate datasets are with low-visual-quality, which have perceptible distortions, such as jitter, blur, and strange artifacts (Zi et al. 2020). Those low-visual-quality Deepfake videos can be easily distinguished by current detection models. However, with the rapid progress of deep generative models, the visual quality of fake videos has been improved from many aspects, such as frame resolution, color correction, smoothness mask, invisible tampered artifacts, and temporal correlations (Li et al. 2020). As shown in Fig. 1, the artifacts and color mismatches are occurred in fake video with low-visual-quality but are relieved in that with high-visual-quality. These significant improvements bring challenges to current detection models. Therefore, it is desirable to spend more effort developing a well-designed method for high-visual-quality Deepfake detection.

Related Work and Challenges. Recent Deepfake video detection methods can be broadly divided into three categories, i.e., cue-inspired methods, data-driven methods, and multi-domain fusion methods. Cue-inspired methods (Li, Chang, and Lyu 2018; Ciftci, Demir, and Yin 2020; Yang, Li, and Lyu 2019; Koopman, Rodriguez, and Geradts 2018; Li and Lyu 2019) expose observable features such as blinking inconsistencies, biological signals, and unrealistic de-

[*]Corresponding author: Xin Liao (xinliao@hnu.edu.cn).

tails to detect Deepfake videos. However, these detection methods may be circumvented by purposely training during the generation of fake videos. Data-driven methods (Afchar et al. 2018; Nguyen, Yamagishi, and Echizen 2019; Nguyen et al. 2019; Tan and Le 2019; Rossler et al. 2019; Zhao et al. 2021; Liu et al. 2021; Xu et al. 2021) extract the invisible features to detect these forgeries efficiently. These methods do not joint spatial information with other domain information, which may ignore crucial features of videos. Towards this end, multi-domain fusion methods (Güera and Delp 2018; Zhao, Wang, and Lu 2020; Qian et al. 2020; Masi et al. 2020; Hu et al. 2021; Sun et al. 2021) train the detection model across multiple domains such as spatial domain, temporal domain, and frequency domain making processes.

Although the aforementioned methods achieve good performances in detecting inchoate datasets, they still need improvement in recent-developed high-visual-quality Deepfake videos. Previous methods (Li, Chang, and Lyu 2018; Afchar et al. 2018; Yang, Li, and Lyu 2019; Hu et al. 2021) focus on specific features which are easily tracked in the low-visual-quality videos, while those features may be badly weakened in high-visual-quality ones and cause the detection performance reduction. Thus we need a more general manner to enlarge the tampered traces of fake videos. Besides, the artifacts dependence of aforementioned methods (Rossler et al. 2019; Zhao et al. 2021) may also cause severely overfitting when conducting the cross-dataset detection. An effective way to solve the overfitting problem is to extend the training data. However, current methods focus on the performance but not the computation efficiency, which brings undesirable time costs. Furthermore, most of the existing detection methods benefit from the powerful ability of CNN, but CNN-based methods lack theoretical interpretation, which is not conducive to the understanding of detection technology. In summary, there are three major challenges when detecting the high-visual-quality Deepfake videos, i.e., 1) it is challenging to enlarge the tampered traces in high-visual-quality Deepfake videos for better performance, 2) it is challenging to improve the robustness for cross-dataset detection and improve detection efficiency, 3) it is challenging to provide interpretable theoretical analysis.

Contributions. To address the challenges, we propose a frame inference-based detection framework (FInfer) to detect high-visual-quality Deepfake videos by inferring future frames' facial representations. To predict future frames' facial representations, FInfer discards the short-term local information between frames and infers more long-term global features. These long-term global features that span multiple time steps are useful for mining the regularities of faces and predicting the future frames' facial representations. The predicted representations will be influenced by Deepfake modification and result in a mismatch with referenced future frames' facial representations. Based on this idea, we first employ an encoder to extract useful representations from current frames and referenced future frames. Thereafter, an autoregressive model is utilized to predict the future frames' facial representations. Ultimately, the face representations of predicted future frames and referenced future frames are jointly optimized by a representation-prediction loss. In this manner, the model can accumulate information over time to predict the future frame's facial representations. Since the correlation of facial representations between predicted future frames and referenced future frames in the real video is higher than that in the fake video, optimizing the representation-prediction loss can improve the detection performance of high-visual-quality Deepfake videos. To the best of our knowledge, we are the first to consider Deepfake detection from the perspective of inferring the future frames' facial representations. The main contributions of this work are three folds:

1) We transform the Deepfake detection task to a video frame inference task, which brings a different viewpoint for Deepfake detection. Different from existing methods that extract features from frames directly, FInfer infers future frames' facial representations by using a video prediction regression model. Ultimately, FInfer obtains the correlation of facial representations between the predicted future frames and the referenced future frames for Deepfake videos detection.

2) We use information theory to analyze the effectiveness of FInfer, which theoretically shows the interpretability of our framework. The joint entropy analysis indicates that high-visual-quality Deepfake videos with low joint entropy can be ideally detected by inferring future frames. On the other hand, the mutual information of fake videos is lower than that of real videos. The mutual information analysis demonstrates that the distinction between real videos and fake videos can be detected by FInfer.

3) We conduct extensive experiments for evaluating FInfer. Experimental results illustrate that FInfer achieves promising performance in extensive metrics.

FInfer: Frame Inference-Based Detection Framework

Fig. 2 shows the proposed frame inference-based detection framework for high-visual-quality Deepfake videos. The current frames and future frames are considered as source frames and target frames for predictions, respectively. FInfer consists of four components: faces preprocessing, faces representative learning, faces predictive learning, and correlation-based learning. First, we extract frames from videos and detect faces from frames. The Gaussian-Laplace pyramid block is utilized to transform faces data. Second, since the data dimensions of video frames are enormous, we utilize the representative learning to construct an encoder to encode the source and the referenced target faces to low dimensional space. With such an encoder, the spatial features of faces are encoded into a compact latent embedding space, which ensures the effectiveness of predictions. Third, we use an autoregressive model to predict the representations of target faces. The prediction model integrates the information of source faces and predicts the representations of the target faces. Fourth, we leverage the correlation-based learning module, which optimizes the model with a devised representation-prediction loss. The representation-prediction loss allows the whole model to be trained end-to-end. FInfer can feedback the loss to the representative learn-

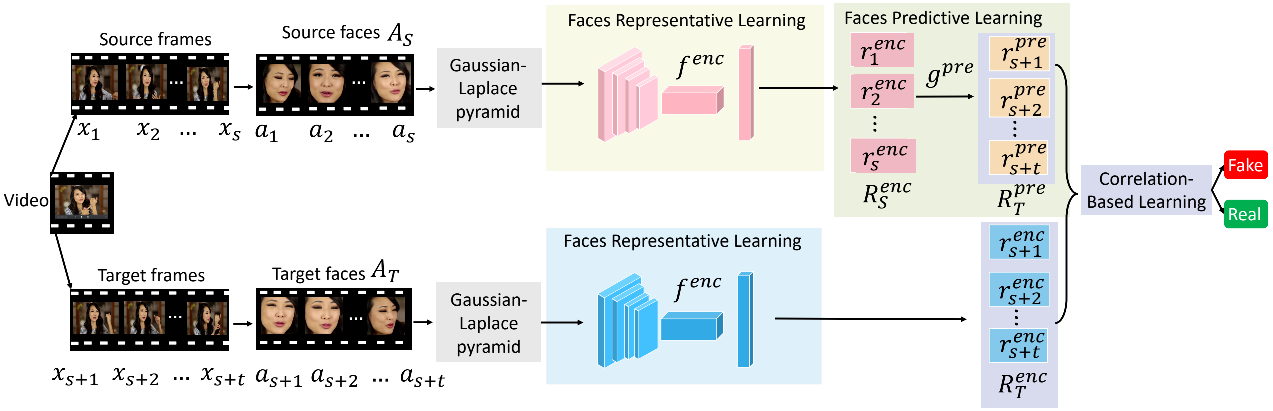

Figure 2: An overview of FInfer. The faces are extracted and transformed from videos. The faces representative learning module encodes both the source faces and target faces. The faces predictive learning module predicts the target face representations from the source face representations. The correlation-based learning module utilizes a representation-prediction loss to train FInfer. By optimizing the model, FInfer can effectively detect the high-visual-quality Deepfake videos.

ing module and predictive learning module, which would later help the model to encode face representations, predict the target representations, and detect the videos.

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. $A = \{a_1, a_2, \dots, a_s, \dots, a_{s+t}\}$ be the faces extracted from X. 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.

Faces Preprocessing

In the faces preprocessing module, operations such as face detection, Gaussian-Laplace pyramid are utilized to improve the visibility of tamper traces. The details are provided as follows.

1. We sequentially extract the consecutive frames X from videos. Since the tampered part of the Deepfake videos is the face area, we focus on the face regions and extract faces A from frames to detect videos.
2. Gaussian-Laplace pyramid block is utilized to expose the boundary artifacts of faces. The Gaussian pyramid is used to generate multiple sets of faces at different scales, which can show the details of faces and the outline information of faces. The Laplace pyramid is utilized to minimize the face information loss caused by the Gaussian pyramid. Finally, the boundaries of faces are exposed.

Faces Representative Learning

In the faces representative learning module, an encoder is utilized to extract source faces information and target faces information as vectors. With the faces representative learning module, the curse of dimensionality is avoided, and the features can be represented. The details are provided as follows.

An encoder is proposed to obtain the representations of source faces and target faces, which can extract useful representations from high-dimensional frames data and improve the training efficiency. The encoder f^{enc} consists of four convolutional layers and a full-connected layer. The first convolutional layer uses 8 filters with a kernel size of $3 * 3$. The second convolutional layer uses 8 filters with a kernel size of $5 * 5$. The third convolutional layer uses 16 filters with a kernel size of $3 * 3$. The fourth convolutional layer uses 16 filters with a kernel size of $5 * 5$. The dimension of the full-connected layer is 128. f^{enc} maps the source faces and target faces to a sequence of latent representations R^{enc}.

$$R^{enc} = f^{enc}(A) = \{r_1^{enc}, r_2^{enc}, \dots, r_{s+t}^{enc}\}. \quad (1)$$

Let $R_S^{enc} = \{r_1^{enc}, r_2^{enc}, \dots, r_s^{enc}\}$ be the representations of source faces. Let $R_T^{enc} = \{r_{s+1}^{enc}, r_{s+2}^{enc}, \dots, r_{s+t}^{enc}\}$ be the representations of target faces.

Faces Predictive Learning

The faces predictive learning module utilizes the regressive prediction to infer the representations of target faces. The details are provided as follows.

1. Source faces representations R_S^{enc} are utilized as the input of the regressive prediction. Since the GRU (Oord, Li, and Vinyals 2018) solves the problem of gradient disappearance and brings high efficiency in the case of extensive training data, we adopt GRU for the regressive part that utilizes various gate functions and states.
2. The update gate is utilized to update the information that the previous faces are carried into the current faces.
3. The reset gate is utilized to infer what is removed from the previous faces.
4. The candidate hidden states use reset gates to store the relevant information from the previous faces.
5. The hidden states are utilized to hold information for the current faces and pass it down to the network.
6. The Time-distributed layer uses facial representations series to perform a series of tensor operations.

7. The prediction module g^{pre} transmits relevant information along a sequence of facial representations to make predictions. g^{pre} preserves important features through various gate functions and infers representations of target faces. The output predictions can be calculated as follows,

$$R_T^{pre} = g^{pre}(R_S^{enc}) = \{r_{s+1}^{pre}, r_{s+2}^{pre}, \ldots, r_{s+t}^{pre}\}. \quad (2)$$

The predicted representations are made up of consecutive vectors with the dimension of $128 * 1$.

Correlation-Based Learning

The correlation-based learning module utilizes the representation-prediction loss to optimize the model. The details are provided as follows.

1. The faces representative learning module outputs the representations of target faces R_T^{enc}. The output of the face predictive learning module is R_T^{pre}. R_T^{enc} and R_T^{pre} are integrated into the correlation-based learning module.
2. The correlation $corr$ between the predicted target face representations R_T^{pre} and the referenced target face representations R_T^{enc} is calculated as follows,

$$corr = sigmoid\left(\frac{\sum_{i=s+1}^{i=s+t} \langle R_T^{pre}, R_T^{enc} \rangle}{t}\right). \quad (3)$$

3. In the process of backpropagation, the representation-prediction loss is employed to train the model end-to-end. The formula of the representation-prediction loss L_N is shown in Eq. (4),

$$L_N = -\frac{1}{N}\sum_z((y_z \times ln(corr)) + (1-y_z) \times ln(1-corr)), \quad (4)$$

where y_z represents the labels of the z-th video.

4. In order to optimize the proposed method, we update the faces representative learning module and faces predictive learning module iteratively and minimize the sum of the representation-prediction loss L_N. Such procedure is shown in Algorithm 1. The input data is A_S and A_T. The faces representative learning module obtains the representations of source faces R_S^{enc} and the representations of target faces R_T^{enc}. The face predictive learning module obtains the representations of predicted target faces R_T^{pre}. The correlation-based learning module minimizes the representation-prediction loss L_N.
5. After the training, the real videos have a higher correlation $corr$ because of the natural expression. The Deepfake videos exist stiff facial expressions, which causes some impact on the prediction. Thus, the Deepfake videos have a lower $corr$ value. Then, the model can detect the difference in facial variations between real videos and Deepfake videos.
6. Finally, we get the $corr$ and calculate the accuracy by the binary accuracy algorithm.

Algorithm 1: The algorithm process of the proposed frame inference-based detection framework.

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 network θ_{df}, f^{enc}, g^{pre}

```
while $\theta_{df}$ have not converged do
    for $i = 1 \rightarrow num\_iter$ do
        $R_S^{enc} = f^{enc}(A_S)$
        $R_T^{enc} = f^{enc}(A_T)$
        $R_T^{pre} = g^{pre}(R_S^{enc}))$
        $g_{\theta_{df}} \leftarrow \nabla_{\theta_{df}}(\frac{1}{b}\sum_{i=1}^{i=b} L_N(R_T^{pre}, R_T^{enc}))$
        $\theta_{df} \leftarrow \theta_{df} + \alpha_{df} \cdot \text{Adam}(\theta_{df}, g_{\theta_{df}})$
    end for
end while
```

Information Theory Analyses

We provide the information theory analyses to show the interpretability of FInfer. Let x_i and x_j be frames of X, namely $x_i, x_j \in X$.

Interpretation for Detecting Videos with High-Visual-Quality

Let $P(x_i, x_j)$ be the joint probability of x_i and x_j. $P(x_i, x_j)$ is obtained by the joint probability distribution of the frames. The joint entropy of frame x_i and x_j can be calculated as follows,

$$H(x_i, x_j) = -\sum_{x_i}\sum_{x_j} P(x_i, x_j) \log_2 P(x_i, x_j). \quad (5)$$

Since the predictive learning module of FInfer utilizes current frames' facial representations to predict future frames' representations, the joint entropy, which demonstrates the uncertainty between two frames, shows the feasibility of prediction. Namely, the x_i and x_j with high uncertainty mean the joint entropy between x_i and x_j is high, which may incur low feasibility to predict x_i from x_j. Fake frames with low-visual-quality may contain visible tampered traces, such as artifacts and mismatching colors in Fig. 1, which increases the uncertainty between x_i and x_j. The uncertainty of low-visual-quality fake videos is higher than that of high-visual-quality fake videos. Therefore, it can be seen from the definition of joint entropy that the inter-frame joint entropy of low-visual-quality videos is higher than that of high-visual-quality videos.

In addition, statistical analysis is performed. Let $H^h(x_i, x_j)$ and $H^l(x_i, x_j)$ are $H(x_i, x_j)$ with high-visual-quality videos and low-visual-quality videos, respectively. According to Eq. (5), we calculate $H^h(x_i, x_j)$ and $H^l(x_i, x_j)$. Specifically, we randomly selected 200 high-visual-quality videos generated by NeuralTextures (Thies,

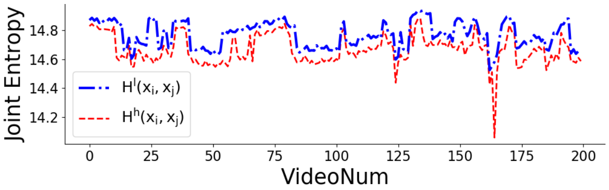

Figure 3: A joint entropy comparison between the low-visual-quality videos and high-visual-quality videos.

Zollhöfer, and Nießner 2019) and 200 low-visual-quality videos generated by DeepFakes (DeepFakes 2018) to calculate the inter-frame joint entropy. These faces have the same face ID and face attribute. Therefore, the face content is almost 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, i.e.,

$$H^h(x_i, x_j) < H^l(x_i, x_j). \tag{6}$$

Thus, the inter-frame joint entropy of fake videos with high-visual-quality is lower than that with low-visual-quality. Low joint entropy of two frames indicates low uncertainty of two frames, which is beneficial for prediction. Therefore, FInfer has significant advantages for high-visual-quality Deepfake videos detection by predicting future frames' faces from current frames' faces.

Interpretation for Detecting the Difference between Real Videos and Fake Videos

Let $RGB_i = \{R_i, G_i, B_i\}$ be a set of color value of RGB channel of frame $x_i \in X$, where R_i, G_i, B_i are red, green, and blue values of x_i, respectively. Let $L = 256$ be the discrete series of RGB channel color value.

The information entropy of x_i can be calculated as follows.

$$H(x_i) = -\sum_{v \in RGB_i} \sum_{v=0}^{L-1} P_v(x_i) \log_2 P_v(x_i), \tag{7}$$

where $P_v(x_i)$ represents the probability of the existence of the color value RGB_i.

The mutual information can be calculated as follows.

$$MI(x_i, x_j) = H(x_i) + H(x_j) - H(x_i, x_j). \tag{8}$$

If the video is a fake video, Eq. (8) can be represented as follows.

$$MI^f(x_i, x_j) = H^f(x_i) + H^f(x_j) - H^f(x_i, x_j). \tag{9}$$

If the video is a real video, Eq. (8) can be represented as follows.

$$MI^r(x_i, x_j) = H^r(x_i) + H^r(x_j) - H^r(x_i, x_j). \tag{10}$$

Since low entropy of frames indicates few details of frames (Golestaneh and Karam 2016), real frames with rich details indicate a higher entropy value than that of fake frames. That is,

$$H(x_i)^r + H^r(x_j) > H(x_i)^f + H^f(x_j). \tag{11}$$

The generated faces in fake videos lack of expressiveness which may bring in the inconsistency between x_i and x_j. Then, the uncertainty between x_i and x_j gets larger. That is, the value of $H^f(x_i, x_j)$ gets larger. The faces in real videos is naturally coherent, which decreases the uncertainty between x_i and x_j. Thus, the value of $H^r(x_i, x_j)$ gets smaller. That is,

$$H^r(x_i, x_j) < H^f(x_i, x_j). \tag{12}$$

According to Eqs. (8-12),

$$MI^r(x_i, x_j) > MI^f(x_i, x_j). \tag{13}$$

Mutual information between x_i and x_j demonstrates the amount of information that x_i obtained from x_j. Thus, high mutual information indicates high feasibility to predict x_j based on x_i, which introduces a high correlation between the predicted x_j and the referenced x_j. According to Eq. (13), the real videos with lager $MI(x_i, x_j)$ will get higher correlation between the predicted x_j and the referenced x_j than that of fake videos. Therefore, the proposed FInfer can detect the difference between real videos and fake videos.

Experimental Evaluations

Experimental Settings

Implement Details. We utilize FFmpeg (Cheng et al. 2012) to extract frames sequentially from videos for data preprocessing. The dlib (King 2009) is adopted to detect face regions from frames. We discard videos if the dlib does not recognize the correct face regions. The extracted face regions are input to the faces representative learning module, which produces the face representations for experiments. The batch size is set as 8. In the training phase, we set the learning rate as 0.001, which will be divided by 5 when the accuracy plateaus. The Adam optimizer (Kingma and Ba 2014) is utilized to optimize the model. We set the default threshold, whose value is 0.5, to calculate binary accuracy. All experiments are conducted in Keras on NVIDIA Titan Xp. The accuracy (ACC) and area under Receiver Operating Characteristic Curve (AUC) are utilized to denote the evaluation metrics for extensive experiments.

Datasets. The FaceForensics++ (FF++) (Rossler et al. 2019) dataset, the Celeb-DF dataset (Li et al. 2020), the WildDeepfake dataset (Zi et al. 2020), and the DFDC-preview dataset(Dolhansky et al. 2019) are utilized to show the performance of FInfer. The FF++ dataset contains 1000 original videos and four types of forgery videos, i.e., DeepFakes, Face2Face, FaceSwap, and NeuralTextures. The Celeb-DF dataset contains 5639 high-visual-quality Deepfake videos with celebrities generated by improved synthesis processes. The WildDeepfake dataset contains 7314 high-visual-quality face sequences extracted from 707 Deepfake videos. The DFDC preview dataset contains about 5000

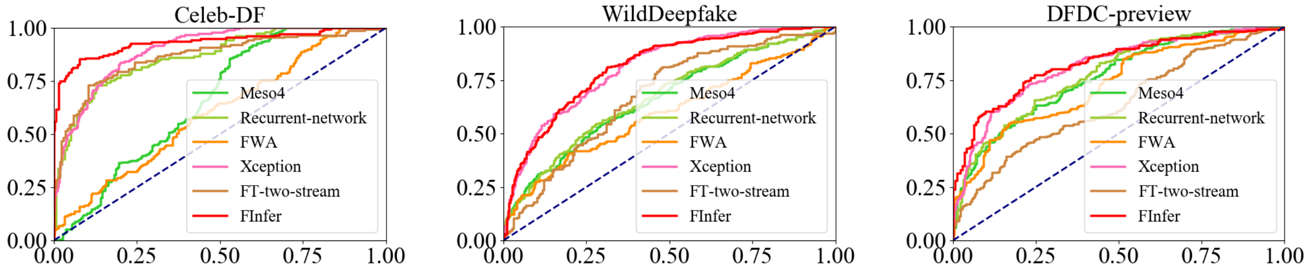

Figure 4: In-dataset ROC curves of FInfer and baseline methods on Celeb-DF, WildDeepfake, and DFDC-preview datasets.

	$s = 10$	$s = 20$	$s = 30$	$s = 40$
$t = 10$	87.30	88.67	89.26	89.06
$t = 20$	N/A	89.49	**90.47**	88.48
$t = 30$	N/A	N/A	86.72	89.45
$t = 40$	N/A	N/A	N/A	88.87

Table 1: ACC (%) of FInfer when choosing different s and t. We do not conduct the experiments on $s < t$.

Dataset	without pyramid block	with pyramid block
Celeb-DF	86.68	**90.47**
WildDeepfake	75.74	**75.88**
DFDC-preview	77.17	**80.39**

Table 2: Ablation study - The forgery detection ACC (%) with and without the Gaussian-Laplace pyramid block.

high-visual-quality videos that manipulated by choosing pairs of similar appearances.

Baselines. We compare FInfer with the baseline methods. The **FWA** (Li and Lyu 2019) is representative of the aforementioned cue-inspired methods. The **Meso4** (Afchar et al. 2018), **MesoInception4** (Afchar et al. 2018), **Xception** (Rossler et al. 2019), **Multi-task** (Nguyen et al. 2019), **Capsule** (Nguyen, Yamagishi, and Echizen 2019), **EfficientNet-B4** (Tan and Le 2019), **SPSL** (Liu et al. 2021), and **Multi-attention** (Zhao et al. 2021) are representative of the aforementioned data-driven methods. The **Recurrent-network** (Güera and Delp 2018), **LTW** (Sun et al. 2021), **FT-two-stream** (Hu et al. 2021), **Two-branch** (Masi et al. 2020), and F^3**-Net** (Qian et al. 2020) are representative of the aforementioned multi-domain fusion methods.

Impacts of s and t

In FInfer, the representations of source faces R_S^{enc} are used for predicting the representations of target faces R_T^{pre}. According to Eq. (5), the joint entropy of x_s and x_{s+t} is $H(x_s, x_{s+t})$, which demonstrates the uncertainty between x_s and x_{s+t}. When t grows large, the uncertainty between x_s and x_{s+t} increases, and the prediction feasibility of predicting x_{s+t} from x_s decreases. Therefore, the length of t shall be limited to a certain extent. If $s < t$, the predictions can be inaccurate, and we only conduct the experiments on $s > t$. To reduce the data dimensionality, the length of s shall be limited to a certain extent. We analyze the impact of the s and t as follows.

To improve the detection accuracy, we vary $s \in \{10, 20, 30, 40\}$ and $t \in \{10, 20, 30, 40\}$ and test the appropriate s and t for FInfer on the Celeb-DF dataset. Table 1 shows the detection performance of FInfer, which demonstrates that FInfer achieves the highest detection accuracy when choosing $t = 20$ and $s = 30$. Therefore, in the following experiments, we set t and s as 20 and 30, respectively.

Ablation Study: Impacts of the Gaussian-Laplace Pyramid Block

We perform ablation studies on FInfer to evaluate the effect of the Gaussian-Laplace pyramid block, which is utilized for faces preprocessing. Specifically, we evaluate FInfer without and with the Gaussian-Laplace pyramid block and show the results on the first and second line of Table 2, respectively. The experimental results demonstrate that the detection accuracy of FInfer with Gaussian-Laplace pyramid block is better than that without Gaussian-Laplace pyramid block. That may be because the Gaussian-Laplace pyramid block can expose the manipulation traces and amplify artifacts. As a result, FInfer with the Gaussian-Laplace pyramid block is beneficial for representing the faces and predicting the future frames' facial representations.

In-Dataset Detection Performance Comparisons

We use ACC and AUC to measure the detection performance on high-visual-quality Deepfake datasets, i.e., Celeb-DF, WildDeepfake, DFDC-preview. We sample 20 frames for each video to calculate the frame-level ACC and AUC scores. The comparison results are listed in Table 3. Compared with baseline methods, FInfer achieves comparable detection performance.

We also show the receiver operating characteristic (ROC) curve in Fig. 4. The abscissa shows the false positive rate (FPR), and the ordinate shows the true positive rate (TPR). The closer the curve to the top left corner, the better the detection performance. Fig. 4 demonstrates that FInfer achieves the satisfying detection performance compared with baseline methods in detecting high-visual-quality Deepfake videos.

FInfer learns the facial variation rules of facial expressions by inferring the future frames' facial representations and comparing the correlation $corr$. We plot $corr$ of the

	Meso4	**Recurrent-network**	**FWA**	**Xception**	**FT-two-stream**	**Two-branch**	**Multi-attention**	**SPSL**	FInfer
Mult-Adds	114.34	5003.34	8220.21	913.46	362.81	574.00	1994.76	408.80	**96.75**

Table 4: Comparisons of the number of Mult-Adds ($\times 10^6$) between FInfer and baseline methods.

Method	Celeb-DF		WildDeepfake		DFDC-preview	
	ACC	AUC	ACC	AUC	ACC	AUC
Meso4	67.53	66.17	64.47	66.50	75.39	76.47
Recurrent-network	71.20	86.52	66.87	67.35	75.02	77.48
FWA	64.73	60.16	55.46	57.92	73.25	72.97
Xception	90.34	89.75	75.26	80.89	79.32	81.58
FT-two-stream	80.74	86.67	68.78	68.09	63.85	64.03
FInfer	**90.47**	**93.30**	**75.88**	**81.38**	**80.39**	**82.88**

Table 3: Comparisons of the in-dataset evaluation (ACC (%) and AUC (%)) between FInfer and baseline methods on Celeb-DF, WildDeepfke, and DFDC-preview datasets.

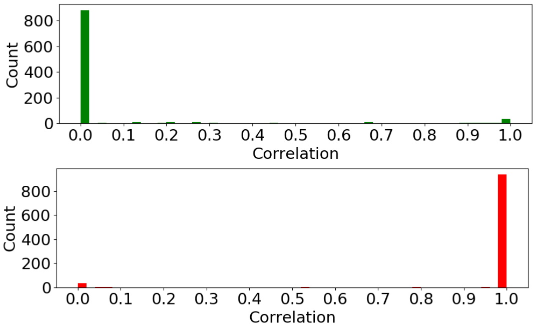

Figure 5: Histogram of correlation scores of fake (green) and real (red) videos.

videos from the testset. As shown in Fig. 5, with few exceptions, the *corr* of real videos is higher than the Deepfake videos.

Detection Efficiency Comparisons

We evaluate the number of Multiplication-Addition operations (Mult-Adds) to show the efficiency advantages of FInfer. We utilize k, C_{in}, C_{out}, and M_{out} to denote the kernel size of a convolutional layer, the number of the input channel, the number of the output channel, and the size of the output feature map, respectively. The number of Mult-Adds M_A of the convolutional layer is calculated by adding up the multiplication computation, addition computation, and bias computation, i.e.,

$$M_A = 2 \times k \times C_{in} \times C_{out} \times M_{out}. \tag{14}$$

We compare the number of Mult-Adds of FInfer with that of baseline methods, which are calculated according to Eq. (14). Table 4 shows that the number of Mult-Adds of FInfer is 96.75×10^6, which is smaller than that of baseline methods. Therefore, FInfer boosts the detection efficiency compared with baseline methods.

Method	FF++	Celeb
Meso4	84.70	54.80
MesoInception4	83.00	53.60
Recurrent-network	90.13	63.56
FWA	80.10	56.90
Xception	99.70	65.30
Multi-task	76.30	54.30
Capsule	96.60	57.50
EfficientNet-B4	99.70	64.29
Multi-attention	**99.80**	67.44
LTW	98.50	64.10
FT-two-stream	92.47	65.56
SPSL	96.91	**76.88**
Two-branch	93.18	73.41
F^3**-Net**	98.10	65.17
FInfer	95.67	70.60

Table 5: Cross-dataset evaluation (AUC(%)) on Celeb-DF by training on FF++. Results of some other methods are cited directly from (Zhao et al. 2021).

Method	WildDeepfake		DFDC-preview	
	ACC	AUC	ACC	AUC
Meso4	59.60	59.74	60.35	59.37
Recurrent-network	63.65	67.03	62.42	66.90
FWA	66.87	67.35	56.65	59.49
Xception	59.32	60.54	63.30	64.29
FT-two-stream	54.69	59.82	60.18	59.09
FInfer	**70.82**	**69.46**	**69.45**	**70.39**

Table 6: Comparisons of the cross-dataset evaluation (ACC (%) and AUC (%)) between FInfer and baseline methods on WildDeepfake, and DFDC-preview datasets.

Cross-Dataset Detection Performance Comparisons

To evaluate the robustness of FInfer, we conduct the cross-dataset experiments that are trained on FF++ with multiple forgery methods but tested on Celeb-DF. The cross-dataset results are shown in Table 5. The first column is the in-dataset detection performance of FF++, and the second column is the cross-dataset detection performance of Celeb-DF. **Multi-attention** achieves the state-of-the-art performance in FF++, however, its cross-dataset AUC is behind ours. Though **SPSL** and **Two-branch** achieve better cross-dataset detection performance than FInfer, the results in Table 4 illustrate that the detection efficiency of FInfer is better than **SPSL** and **Two-branch**. In addition, FInfer achieves

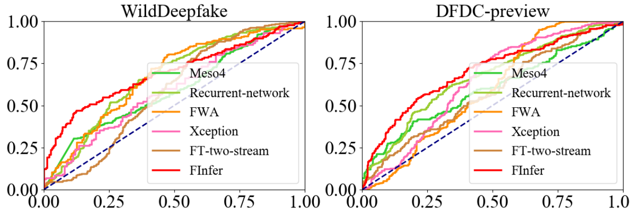

Figure 6: Cross-dataset ROC curves of FInfer and baseline methods on WildDeepfake, and DFDC-preview datasets.

comparable robustness performance than most of the baseline methods. Furthermore, we test other high-visual-quality Deepfake videos datasets, i.e., WildDeepfake and DFDC-preview. The results in Table 6 and Fig. 6 demonstrate that FInfer achieves competitive cross-dataset detection performance on WildDeepfake and DFDC-preview datasets.

The aforementioned comparison results show that the performance of FInfer is better than most baseline methods in detecting high-visual-quality Deepfake videos. That may be because that most baseline methods capture features that are weakened in the high-visual-quality videos, which causes an impact on the detection. Furthermore, FInfer detect the high-visual-quality Deepfake videos by incorporating frame inferences into the training process. When detecting high-visual-quality videos, FInfer infers the target frames' facial representations and compares the predicted target frames' facial representations with the referenced target frames' facial representations rather than extracting features directly from the frames, which benefits the detection model.

Conclusions

In this paper, we propose FInfer for a high-visual-quality Deepfake videos detection case. We formulates the Deepfake detection task as a future frames' facial representations inference task, which presents a different perspective for Deepfake videos detection. Besides, we adopt information theory analyses for FInfer, which demonstrates the effectiveness of FInfer theoretically. Extensive experiments show that FInfer achieves competitive detection performance and detection efficiency in different detection cases.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant Nos. 61972142, 62002334, U20A20174, 61772191), Hunan Provincial Natural Science Foundation of China (Grant No. 2020JJ4212).

References

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

Cheng, Y.; Liu, Q.; Zhao, C.; Zhu, X.; and Zhang, G. 2012. Design and implementation of mediaplayer based on FFmpeg. In *Software Engineering and Knowledge Engineering: Theory and Practice*, 867–874.

Ciftci, U. A.; Demir, I.; and Yin, L. 2020. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: 10.1109/TPAMI.2020.3009287.

DeepFakes. 2018. Accessed October 10, 2018. https://github.com/deepfakes/faceswap.

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

FaceSwap. 2018. Accessed October 29, 2018. https://github.com/MarekKowalski/FaceSwap/.

Golestaneh, S. A.; and Karam, L. J. 2016. Reduced-Reference Quality Assessment Based on the Entropy of DWT Coefficients of Locally Weighted Gradient Magnitudes. *IEEE Transaction on Image Process.*, 25(11): 5293–5303.

Güera, D.; and Delp, E. J. 2018. Deepfake video detection using recurrent neural networks. In *Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance*, 1–6.

Hu, J.; Liao, X.; Wang, W.; and Qin, Z. 2021. Detecting compressed Deepfake videos in social networks using frame-temporality two-stream convolutional network. *IEEE Transactions on Circuits and Systems for Video Technology*. DOI: 10.1109/TCSVT.2021.3074259.

King, D. E. 2009. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(60): 1755–1758.

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

Koopman, M.; Rodriguez, A. M.; and Geradts, Z. 2018. Detection of deepfake video manipulation. In *Proceedings of the Irish Machine Vision and Image Processing Conference*, 133–136.

Li, Y.; Chang, M.; and Lyu, S. 2018. In Ictu Oculi: Exposing AI created fake videos by detecting eye blinking. In *Proceedings of IEEE International Workshop on Information Forensics and Security*, 1–7.

Li, Y.; and Lyu, S. 2019. Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops.*, 46–52.

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

Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; and Yu, N. 2021. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 772–781.

Masi, I.; Killekar, A.; Mascarenhas, R. M.; Gurudatt, S. P.; and AbdAlmageed, W. 2020. Two-branch recurrent network for isolating deepfakes in videos. In *Proceedings of European Conference on Computer Vision*, 667–684.

Nguyen, H. H.; Fang, F.; Yamagishi, J.; and Echizen, I. 2019. Multi-task learning for detecting and segmenting manipulated facial images and videos. In *Proceedings of IEEE International Conference on Biometrics Theory, Applications and Systems*, 1–8.

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

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of European Conference on Computer Vision*, 86–103.

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

Sun, K.; Liu, H.; Ye, Q.; Liu, J.; Gao, Y.; Shao, L.; and Ji, R. 2021. Domain general face forgery detection by learning to weight. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2638–2646.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of International Conference on Machine Learning*, 6105–6114.

Thies, J.; Zollhöfer, M.; and Nießner, M. 2019. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 38(4): 1–12.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2018. Face2Face: Real-time face capture and reenactment of RGB videos. In *Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2387–2395.

Whittaker, L.; Kietzmann, T. C.; Kietzmann, J.; and Dabirian, A. 2020. All around me are synthetic faces: The mad world of AI-generated media. *IT Professional*, 22(5): 90–99.

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

Yang, X.; Li, Y.; and Lyu, S. 2019. Exposing deep fakes using inconsistent head poses. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 8261–8265.

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

Zhao, Z.; Wang, P.; and Lu, W. 2020. Detecting Deepfake video by learning two-level features with two-stream convolutional neural network. In *Proceedings of the International Conference on Computing and Artificial Intelligence*, 291–297.

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