Exposing the Deception: Uncovering More Forgery Clues for Deepfake Detection

Zhongjie Ba[1,2], Qingyu Liu[1,2], Zhenguang Liu[1,2] *, Shuang Wu[3], Feng Lin[1,2], Li Lu[1,2], Kui Ren[1,2]

[1]State Key Lab. of Blockchain and Data Security, Zhejiang University, Hangzhou, China
[2]ZJU-Hangzhou Global Scientific and Technological Innovation Center, Hangzhou, China
[3]Black Sesame Technologies, Singapore
{zhongjieba,qingyuliu}@zju.edu.cn, liuzhenguang2008@gmail.com, wushuang@outlook.sg, {flin,li.lu,kuiren}@zju.edu.cn,

Abstract

Deepfake technology has given rise to a spectrum of novel and compelling applications. Unfortunately, the widespread proliferation of high-fidelity fake videos has led to pervasive confusion and deception, shattering our faith that seeing is believing. One aspect that has been overlooked so far is that current deepfake detection approaches may easily fall into the trap of overfitting, focusing only on forgery clues within one or a few local regions. Moreover, existing works heavily rely on neural networks to extract forgery features, lacking theoretical constraints guaranteeing that sufficient forgery clues are extracted and superfluous features are eliminated. These deficiencies culminate in unsatisfactory accuracy and limited generalizability in real-life scenarios.

In this paper, we try to tackle these challenges through three designs: (1) We present a novel framework to capture broader forgery clues by extracting multiple non-overlapping local representations and fusing them into a global semantic-rich feature. (2) Based on the information bottleneck theory, we derive Local Information Loss to guarantee the orthogonality of local representations while preserving comprehensive task-relevant information. (3) Further, to fuse the local representations and remove task-irrelevant information, we arrive at a Global Information Loss through the theoretical analysis of mutual information. Empirically, our method achieves state-of-the-art performance on five benchmark datasets. Our code is available at https://github.com/QingyuLiu/Exposing-the-Deception, hoping to inspire researchers.

1 Introduction

Fueled by the accessibility of large-scale video datasets and the maturity of deepfake technologies (Nirkin, Keller, and Hassner 2019; Li et al. 2019), one may effortlessly create massive forgery videos beyond human discernibility. However, malicious usage of deepfake can have serious influences, ranging from identity theft and privacy violations to large-scale financial frauds and dissemination of misinformation. For instance, in March 2022, hackers created a fake video of the Ukrainian president Zelenskyy in which he stands at a podium and addresses Ukrainian soldiers to lay down their arms. Such events are far from isolated, and they highlight the risk of deepfake technology in misleading the public and undermining trust. Consequently, accurate and effective deepfake detection is essential for mitigating these risks.

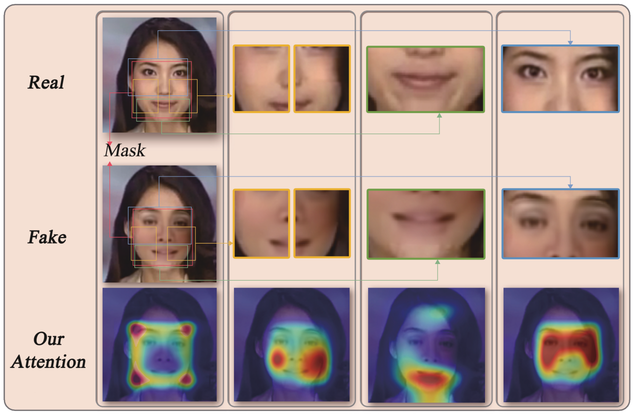


Figure 1: Example visualization of four local salient features obtained by our method. Each feature focuses on distinct forgery regions with little overlap. We zoom in to show the detailed differences for these regions between a real sample and a fake sample. Our method can grasp broader forgery clues including *blending ghosts, consistent and symmetrical skin tones, tooth details*, and *stitching seams*.

Fundamentally, deepfake detection amounts to recognizing forgery clues that distinguish real and synthetic images. There emerge many studies for deepfake detection, which can be roughly categorized into two main branches. One line of work (Afchar et al. 2018; Dolhansky et al. 2020) employs CNN networks to *automatically* learn the clues within manipulated images. Another line of work is dedicated to pondering and exploring the differences between fake and real images by incorporating human observation and understanding. These approaches hone in on high-level semantic imperfections of counterfeits (Haliassos et al. 2021), as well as underlying imperceptible patterns in artifacts (such as *blending ghost* (Shiohara and Yamasaki 2022) and *frequency domain anomalies* (Liu et al. 2021)).

After scrutinizing and experimenting with the implementations of state-of-the-art approaches, we obtain two empiri-

*The corresponding author.

cal insights: (1) Despite achieving a high AUC on the training dataset, current methods usually experience a substantial decrease in AUC on unseen datasets. This may stem from the fact that current methods tend to unintentionally learn shortcuts for the training dataset, focusing only on one or a few forgery clues. (2) Current works heavily rely on neural networks to automatically extract forgery features, lacking rigorous theoretical guarantees to capture sufficient label-relevant information and to eliminate superfluous information. Consequently, the extracted features may converge to insufficient representations or trivial features, compromising the accuracy of such methods.

Motivated by these, we advocate extracting broader forgery clues for deepfake detection and seek to lay the mathematical foundation for sufficient forgery feature extraction. Specifically, we first adaptively extract multiple disentangled local features focused on non-overlapping aspects of the suspect image (as in Fig. 1). To ensure the orthogonality of these local features while preserving comprehensive task-relevant information, we utilize mutual information to derive an information bottleneck objective, *i.e.,* Local Information Loss. Secondly, we fuse local features into a global representation guided by Global Information Loss that serves to eliminate task-irrelevant information.

To evaluate the effectiveness of our method, we conduct extensive experiments on five widely used benchmark datasets, *i.e.*, FaceForensics++ (Rossler et al. 2019), two versions of Celeb-DF (Li et al. 2020b), and two versions of DFDC (Dolhansky et al. 2020). We also conduct ablation studies to assess the efficacy of each key component in our method. Our method achieves state-of-the-art performance for both in-dataset (the training and test datasets are sampled from the same domain) and cross-dataset (the training and test datasets are two different datasets) settings. In summary, our contributions are as follows:

- We propose a novel framework for deepfake detection that aims to obtain broader forgery clues.
- We mathematically formulate a mutual information objective to effectively extract disentangled task-relevant local features. Additionally, we introduce another objective for aggregating the local features and eliminating superfluous information. We provide a rigorous theoretical analysis to show how these mutual information objectives can be optimized.
- Empirically, our method achieves state-of-the-art performance on five benchmark datasets. Interesting and new insights are also presented (*e.g.,* most deepfake detection approaches tend to focus on only a few specific regions around the face swap boundaries).

2 Related Work

Fueled by the maturity of deep learning models and large-scale labeled datasets, deep learning has found its applications in various fields (Zhang et al. 2023; Wei et al. 2023; Liu et al. 2022; Chiou et al. 2020; Liu et al. 2023; Song, Chen, and Jiang 2023), especially for deepfake. The ease of access and misuse of deepfake technology has led to the materialization of severe risks, and developing deepfake detection to counteract such threats is all the more pertinent and urgent. Deepfake detection (Ying et al. 2023; Ba et al. 2023; Hua et al. 2023; Wu et al. 2023; Pan et al. 2023; Shuai et al. 2023) faces a significant challenge posed by the sophistication of deepfake technology that can create highly realistic content that is barely distinguishable from real ones.

A large body of literature (Dong et al. 2022b; Li et al. 2020a; Shiohara and Yamasaki 2022; Chen et al. 2022a; Haliassos et al. 2021; Zhao et al. 2021a) focuses on semantic facial feature clues of forgeries. ICT (Dong et al. 2022b) models identity differences in the inner and outer facial regions. Face X-ray (Li et al. 2020a) and SBIs (Shiohara and Yamasaki 2022) find the blending boundaries of face swap as evidence for forged images and build private augmented datasets. Chen.*et al.* (Chen et al. 2022a) further expands upon blending-based forgeries, considering the eyes, nose, mouth, and blending ratios. LipForensics (Haliassos et al. 2021) observes the irregularities of mouth movements in forgery videos. However, such methods only apply to the detection of face swaps and semantic-guided forgery detection cannot be exhaustive vis-à-vis the rapid development of deepfake techniques.

Another class of works (Frank et al. 2020; Liu et al. 2021; Luo et al. 2021; Qian et al. 2020) proposes to take into further consideration human understanding differences in the frequency domain. Qian.*et al.* (Qian et al. 2020) employ frequency as complementary evidence for detecting forgeries, which can reveal either subtle forgery clues or compression errors. Frank.*et al.* (Frank et al. 2020) and SPSL (Liu et al. 2021) search for ghost artifacts resulting from up-sampling operations in generative networks. While such works include more features than pure semantically visual clues, the additional features modelled tend to be domain-specific, thereby failing to generalize well to cross-dataset scenarios.

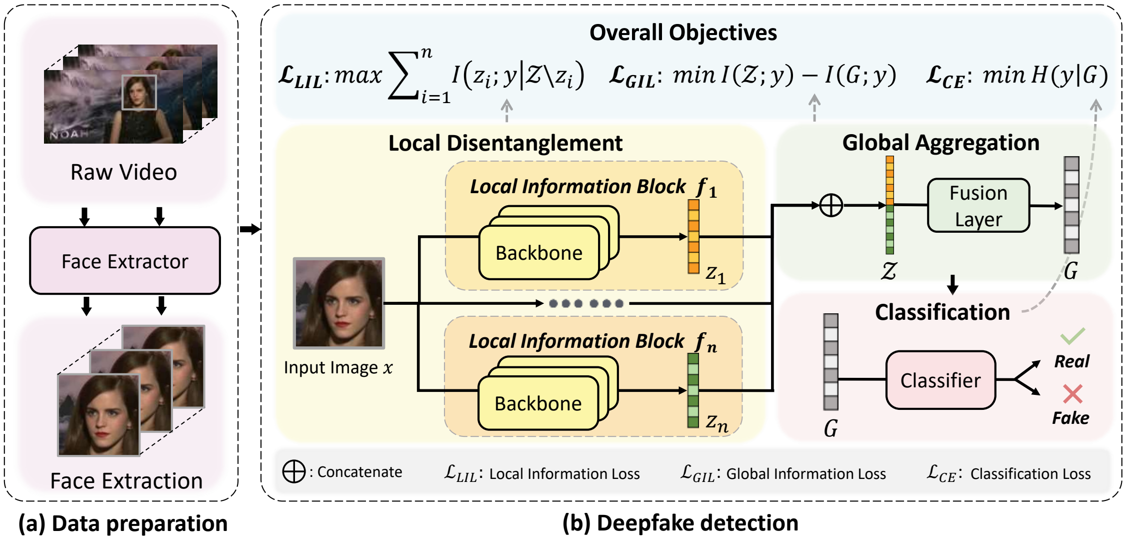
Researchers have also engaged in multi-headed attention modules to correlate the low-level textural features and high-level semantics at different regions for deepfake detection (Zhao et al. 2021a). Nevertheless, a challenge persists, as there exists no concrete theoretical assurance that these attention regions segmented based on the paradigm of human vision remain entirely task-relevant and independent. Furthermore, the performance of such attention-based models is greatly affected by data scarcity.

3 Methodology

3.1 Overview

Presented with a suspect image, we aim to judge its authenticity by extracting forgery clues that could distinguish between genuine and synthetic images. Technically, deepfake detection can be viewed as a binary classification problem.

Early methods (Afchar et al. 2018; Dolhansky et al. 2020) directly utilize deep neural networks to automatically learn differences between genuine and synthetic images. Recent works (Shiohara and Yamasaki 2022; Haliassos et al. 2021; Liu et al. 2021) try to draw inspiration from human understanding and explore human-perceivable forgery clues. Unfortunately, current approaches tend to unintentionally learn


Raw Video
Face Extractor
Face Extraction
Overall Objectives
$\mathcal{L}_{LIL}$: $max \sum_{i=1}^{n} I(z_i; y|\mathcal{Z}\backslash z_i)$
$\mathcal{L}_{GIL}$: $min\, I(\mathcal{Z}; y) - I(G; y)$
$\mathcal{L}_{CE}$: $min\, H(y|G)$
Local Disentanglement
Local Information Block f_1
Backbone
z_1
Input Image x
Local Information Block f_n
Backbone
z_n
Global Aggregation
Fusion Layer
$\mathcal{Z}$
G
Classification
G
Classifier
Real
Fake
⊕: Concatenate
$\mathcal{L}_{LIL}$: Local Information Loss
$\mathcal{L}_{GIL}$: Global Information Loss
$\mathcal{L}_{CE}$: Classification Loss

(a) Data preparation **(b) Deepfake detection**

Figure 2: Method overview. In the data preparation phase, we first extract frame-level facial bounding boxes from raw videos. For deepfake detection, our method consists of three modules. i) We first employ local information blocks f_i to extract multiple disentangled local features z_i corresponding to different forgery regions. We introduce local information loss to ensure that z_i has comprehensive forgery-related information and is orthogonal to z_j. ii) We fuse all z_i into a global feature G under the guidance of a Global Information Loss. iii) Finally, G is passed to the classification module to output the prediction result. We design our Local and Global Information Loss based on information bottleneck theory.

shortcuts[1], which actually makes approaches focus only on one or a few forgery regions. This overfitting issue manifests as the limited generalizability of state-of-the-art methods, namely significant accuracy decreases when applied to unseen datasets. Furthermore, due to the lack of rigorous theoretical constraints, neural networks of current methods may converge to trivial features or insufficient representations.

Motivated by these, we propose to broaden our extraction of forgery clues by adaptively extracting multiple non-overlapping local features. We also establish a theoretical foundation to ensure the orthogonality and sufficiency of extracted features. Overall, our approach consists of two key components: the *local disentanglement module* and the *global aggregation module*. The local disentanglement module serves to extract the non-overlapping local features while the global aggregation module is designed to aggregate them into a global representation.

The pipeline of our proposed approach is shown in Fig. 2, which can be summarized as follows. (1) For image preprocessing, we extract face regions using a popular pre-trained backbone network. (2) Given the preprocessed image as an input, we design the local disentanglement module to extract multiple local features. The local disentanglement module comprises n local information blocks $\{f_i\}_{i=1}^{n}$, each extracting a local feature z_i. f_i consists of feature extraction backbone networks such as ResNet (He et al. 2016). To ensure that local features contain comprehensive information related to the task and z_i is orthogonal to $z_j (i \neq j)$, we derive the Local Information Loss $\mathcal{L}_{LIL}$. (3) Thereafter, we design the global aggregation module to fuse local features. Specifically, we first concatenate all local features to a joint local representation $\mathcal{Z} = \bigoplus_i^n z_i$. Then, a fusion layer f_g serves to fuse and compress $\mathcal{Z}$ to obtain our final global representation G for classification. To guide this global representation extraction, we design a Global Information Loss $\mathcal{L}_{GIL}$ that facilitates the retaining of sufficient task-related information and the elimination of superfluous information in $\mathcal{Z}$.

In what follows, we elaborate on the details of the *local disentanglement* and *global aggregation* modules one by one.

3.2 Local Disentanglement Module

In this section, we provide the key derivation of Local Information Loss within the local disentanglement module.

Given an input image x with n ($n \geq 2$) associated local feature representations z_i, our Local Information Loss objective seeks to ensure two fundamental properties within the joint local representation $\mathcal{Z} = \bigoplus_i^n z_i$, *i.e.,* **comprehensiveness and orthogonality**. Comprehensiveness mandates the inclusion of maximal task-relevant information within $\mathcal{Z}$, while orthogonality necessitates that the individual local features z_i remain non-overlapping. To facilitate understanding, Fig. 3 shows the information relationship when $n = 2$.

In the terminology of mutual information theory, the relationship between labels y and $\mathcal{Z}$ is expressed as:

$$I(y; \mathcal{Z}) = I(y; z_1, \cdots, z_n) = H(y) - H(y \mid \mathcal{Z}), \qquad (1)$$

where $I(*)$ is mutual information and $H(*)$ is entropy. $I(y; \mathcal{Z})$ expresses the amount of predictive information (*i.e.*,

[1] Shortcuts are decision rules optimized for benchmark performance but incapable of transferring to more challenging testing conditions due to a domain gap.

current task-related information) contained in $\mathcal{Z}$. $H(y \mid \mathcal{Z})$ and $H(y)$ represent the required and whole information related to the task, respectively. The comprehensiveness objective of information in $\mathcal{Z}$ is given by:

$$\max I(y; \mathcal{Z}). \tag{2}$$

The orthogonality condition between two probability distributions is equivalent to them having zero mutual information. As such, we can disentangle local feature representations by minimizing the mutual information between them, *i.e.,* $\min \sum_{i \neq j}^{n} I(z_i; z_j)$. According to the definition of interaction information (McGill 1954), $I(z_i; z_j)$ can be further decomposed into:

$$I(z_i; z_j) = \underbrace{I(z_i; z_j; y)}_{\text{target}} + \underbrace{I(z_i; z_j \mid y)}_{\text{superfluous}}, \tag{3}$$

where $I(z_i; z_j; y)$ represents the amount of label information retained within both z_i and z_j, while $I(z_i; z_j \mid y)$ is extraneous (superfluous) information encoded within both z_i and z_j, which is irrelevant to the task. For the orthogonality of local features, we are primarily concerned with label-related (target) information. As for the elimination of superfluous information, we formulate an objective inspired by the information bottleneck, namely Global Information Loss (which will be discussed later in the following section). We first focus on the target term in Eq. 3, *i.e.,*:

$$\min \sum_{i \neq j}^{n} I(z_i; z_j; y). \tag{4}$$

By applying the chain rule for mutual information, $I(y; \mathcal{Z}) = \sum_{i=1}^{n} I(z_i; y \mid z_1, \cdots, z_{i-1})$, we can rewrite Eq. 2 as:

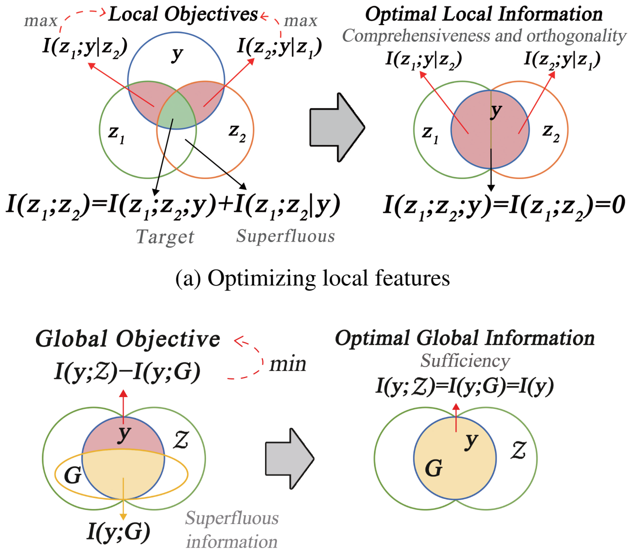
$$\max I(y; \mathcal{Z}) \leq \max \sum_{i \neq j}^{n} I(z_i; y \mid \mathcal{Z} \setminus z_i) + I(z_i; z_j; y), \tag{5}$$

where $\mathcal{Z} \setminus z_i \equiv z_1 \oplus \cdots \oplus z_{i-1} \oplus z_{i+1} \oplus \cdots \oplus z_n$. Overall, the comprehensiveness and orthogonality constraints for local feature extraction can be achieved by simultaneously optimizing Eq. 5 and Eq. 4. It is worth noting that these optimization objectives are in conflict with $I(z_i; z_j; y)$. After resolving these conflicting constraints, the local objective is eventually:

$$\max \sum_{i=1}^{n} I(z_i; y \mid \mathcal{Z} \setminus z_i). \tag{6}$$

Intuitively, the local objective corresponds to the red regions illustrated in Fig. 3a. By optimizing Eq. 6, our goal is to ideally cover all task-relevant information with disentangled local features.

However, directly estimating Eq. 6 is intractable in general. Earlier works (Poole et al. 2019) have pointed out major difficulties in mutual information estimation, primarily due to the curse of dimensionality (the amount of samples for accurately estimating mutual information scales exponentially with the embedding dimension). In light of this, we optimize Eq. 6 via a variational approach instead of explicitly estimating the mutual information. We have the following theorem (detailed proof is in supplementary files):



(a) Optimizing local features

(b) Optimizing the global feature

Figure 3: Information content of feature representations

Theorem: *Eq. 6 has a lower bound due to:*

$$\sum_{i=1}^{n} I(z_i; y \mid \mathcal{Z} \setminus z_i) \geq \sum_{i=1}^{n} D_{KL}\left[\mathbb{P}_z \| \mathbb{P}_{z \setminus z_i}\right], \tag{7}$$

where $\mathbb{P}_{\mathcal{Z} \setminus z_i} = p\left(y \mid \mathcal{Z} \setminus z_i\right), \mathbb{P}_{\mathcal{Z}} = p\left(y \mid \mathcal{Z}\right)$ represent the predicted distributions. D_{KL} denotes the Kullback-Leibler (KL) divergence.

Given the above analytical derivations, we can thus formulate the Local Information Loss as:

$$\mathcal{L}_{LIL} = \min_{\theta} \exp\left(-\sum_{i=1}^{n} D_{KL}\left[\mathbb{P}_z \| \mathbb{P}_{z \setminus z_i}\right]\right), \tag{8}$$

where θ denotes the model parameters in our local disentanglement module. Here, since the KL-divergence is not bounded above, *i.e.* $D_{KL} \in [0, \infty)$, we take the exponential of its negative value to transform the objective from maximization to minimization. The transformed objective is bounded within $(0, 1]$ which is numerically advantageous. Upon optimizing for this objective, local features are constrained to be mutually orthogonal while simultaneously approaching the maximal covering of all task-related information. In this way, our method uncovers more forgery clues and disentangles forgery regions adaptively, thus obtaining feature representations with richer task-related information.

3.3 Global Aggregation Module

Following the local disentanglement module, the concatenated local features $\mathcal{Z}$ encompass comprehensive but not purified information related to the task. Thus, we pass $\mathcal{Z}$ through our global aggregation module, which plays the role of an information bottleneck[2] to eliminate the superfluous

[2] The concept of information bottlenecks is proposed in (Tishby, Pereira, and Bialek 2000) which attributes the robustness of a machine learning model to its ability to distill superfluous noises while retaining only useful information.

information and obtain a global representation G. The information bottleneck objective can be formulated as:

$$\mathcal{L}_{IB} = H(G) - I(y; G), \tag{9}$$

where $H(G)$ denotes the total information content in G. Once again, estimating $\mathcal{L}_{IB}$ is an intractable problem in practice due to the curse of dimensionality. Minimizing superfluous information will therefore be delegated to the network operations and is not explicitly supervised.

Therefore, to ensure that G has sufficient label information, we employ a variational approach once again. Since G is a representation learnt from $\mathcal{Z}$, the task-relevant information in G is upper-bounded by that in $\mathcal{Z}$, denoted as $I(y; G) \leq I(y; \mathcal{Z})$. By minimizing the label information difference between local features and the global feature, we optimize G for the **sufficiency** of label information:

$$\min I(y; \mathcal{Z}) - I(y; G). \tag{10}$$

We make use of the following theorem from (Tian et al. 2021):

$$\min I(y; \mathcal{Z}) - I(y; G) \iff \min[D_{KL}[\mathbb{P}_{\mathcal{Z}} \| \mathbb{P}_G]]]. \tag{11}$$

Finally, we arrive at the Global Information Loss:

$$\mathcal{L}_{GIL} = \min_{\phi} \mathbb{E}_{G \sim E_{\phi}(G|\mathcal{Z})} \left[D_{KL}[\mathbb{P}_{\mathcal{Z}} \| \mathbb{P}_G]\right], \tag{12}$$

where ϕ denotes the model parameters of the global aggregation module.

Overall Objective The overall objective for our framework consists of a cross-entropy classification loss, Local Information Loss, and Global Information Loss:

$$\mathcal{L} = \mathcal{L}_{CE} + \alpha \mathcal{L}_{LIL} + \beta \mathcal{L}_{GIL}, \tag{13}$$

where α and β are hyperparameters for the model.

4 Evaluation

In this section, we conduct extensive experiments on five large-scale deepfake datasets to evaluate the proposed method, including setup, comparison with state-of-the-art methods, ablation study, and visualization results. See the supplementary file for more experimental results.

4.1 Experimental Setup

Datasets. Following existing deepfake detection approaches (Chen et al. 2022a; Bai et al. 2023), we evaluate our model on five public datasets, namely FaceForensics++ (FF++) (Rossler et al. 2019), two versions of Celeb-DF (Li et al. 2020b) and two versions of DeepFake Detection Challenge (DFDC) (Dolhansky et al. 2020) datasets. **FF++** dataset, which is the most widely used dataset, utilizes four forgery-generation methods for producing 4,000 forgery videos, *i.e.,* DeepFakes (DF), Face2Face (FF), FaceSwap (FS), and NeuralTextures (NT). FF++ has three compression versions and we use the high-quality level (C23) one for training. Celeb-DF dataset contains two versions, termed **Celeb-DF-V1** (CD1) and **Celeb-DF-V2** (CD2). CD1 consists of 408 pristine videos and 795 manipulated videos, while CD2 contains 590 real videos and 5,639 DeepFake videos. DFDC dataset includes **DFDC-Preview** (DFDC-P) and **DFDC**. DFDC-P as the preview of DFDC consists of 5,214 videos. DFDC, as one of the most large-scale face swap datasets, contains more than 110,000 videos sourced from 3,426 actors.

Implementation details. In data pre-processing, we use the state-of-the-art face extractor RetinaFace (Deng et al. 2020) and oversample pristine videos to balance training datasets. For the model architecture, we employ four local information blocks (LIBs) using pre-trained ResNet-34 (He et al. 2016) as the backbone. For training, we use the method in (Liebel and Körner 2018) to determine α and β in Eq. 13, to automatically balance the weights for these loss terms.

Evaluation metrics. We utilize the Accuracy (ACC), Area Under Receiver Operating Characteristic Curve (AUC), and log-loss score for empirical evaluation. (1) **ACC**. We employ the accuracy rate as one of the metrics in our evaluations, which is commonly used in deepfake detection tasks. (2) **AUC**. Considering the imbalance of pristine and manipulated videos in the datasets, we use AUC as the predominant evaluation metric. (3) **Logloss**. This is the evaluation metric designated for the Deepfake Detection Challenge. We evaluate the log-loss score to benchmark our method against winning teams. By default, we use frame-level metrics. Since our method uses a single frame as input, we also compute video-level AUC (as in (Haliassos et al. 2021)) for a more comprehensive comparison with video-level detection methods.

4.2 Comparison with Existing Methods

In this section, we benchmark our method against state-of-the-art deepfake detection methods for *in-dataset* and *cross-dataset* settings.

In-dataset performance In in-dataset evaluations, we train and test methods on FF++ (C23), CD2, and DFDC, respectively. Tab. 1 presents in-dataset comparison results. Considering that most current deepfake detection methods have not yet released their codes, we directly cited their results in the corresponding original papers. From Tab. 1, we observe that our method is capable of consistently outperforming existing methods on all three benchmarks. For example, the AUC of our method is 0.939 on DFDC while the state-of-the-art detection method Chugh.*et al.* (Chugh et al. 2020) is 0.907. The Logloss of our method is also outperformed by the champion team in DFDC.

Cross-dataset Performance and Model Generalizability The cross-dataset setting is more challenging than the in-dataset setting for deepfake detection. To evaluate the generalization abilities of the methods on unseen datasets, we train the models on the FF++ (C23) dataset and test them on CD1, CD2, DFDC-P, and DFDC datasets. Since our method uses a single frame as input, in addition to frame-level comparisons, we also compute the average AUC over frames in a video for comparison with video-level methods. Tab. 2 and Tab. 3 demonstrate cross-dataset comparison results in terms of frame-level and video-level AUC, respectively. Our first insight is that state-of-the-art deepfake detection methods still suffer from relatively low AUC on unseen datasets, which reveals that such methods are prone to overfitting the training dataset. The second insight

FF++(C23)		Celeb-DF-V2		DFDC		
Method	AUC↑	Method	AUC↑	Method	AUC↑	LogLoss↓
Xception (Rossler et al. 2019)	0.963	DeepfakeUCL (Fung et al. 2021)	0.905	Selim Seferbekov*	0.882	0.4279
Xception-ELA	0.948	SBIs (Shiohara and Yamasaki 2022)	0.937	NTechLab*	0.880	0.4345
SPSL (Masi et al. 2020)	0.943	Agarwal et al. 2020	0.990	Eighteen Years Old*	0.886	0.4347
Face X-ray (Li et al. 2020a)	0.874	Wu et al. 2023	0.998	WM*	0.883	0.4284
TD-3DCNN (Zhang et al. 2021)	0.722	TD-3DCNN	0.888	TD-3DCNN	0.790	0.3670
F^3-Net (Qian et al. 2020)	0.981	Xception	0.985	Chugh et al. 2020	0.907	-
FInfer (Hu et al. 2022)	0.957	FInfer	0.933	FInfer	0.829	-
Ours (ResNet34)	**0.983**	**Ours (ResNet34)**	**0.999**	**Ours (ResNet34)**	**0.939**	**0.3379**

Table 1: In-dataset comparison results on FF++, Celeb-DF-V2, and DFDC. We train and test models on the same dataset, reporting the frame-level AUC and LogLoss. * is the method of winning the top four teams in DFDC. The bold and underline mark the best and second performances, respectively.

Method	Training dataset	CD1	CD2	DFDC-P	DFDC
Xception (Rossler et al. 2019)	FF++	0.750*	0.778*	0.698*	0.636*
DSP-FWA (Li and Lyu 2018)	FF++	0.785*	0.814*	0.595*	-
Meso4 (Afchar et al. 2018)	FF++	0.422*	0.536*	0.594*	-
F^3-Net (Qian et al. 2020)	FF++	-	0.712*	0.729*	0.646*
Face X-ray (Li et al. 2020a)	PD	0.806	0.742*	0.809	-
Multi-Attention (Zhao et al. 2021a)	FF++	-	0.674	-	0.680*
OST (Chen et al. 2022b)	FF++	0.748	-	0.833	-
HCIL (Gu et al. 2022a)	FF++	-	0.790	0.692	-
LiSiam (Wang, Sun, and Tang 2022)	FF++	0.811	0.782	-	-
RECCE (Cao et al. 2022)	FF++	-	0.687	-	0.691
ICT (Dong et al. 2022b)	PD	0.814	0.857	-	-
DCL (Sun et al. 2022)	FF++	-	0.823	0.767	-
IID (Huang et al. 2023)	FF++	-	0.838	0.812	-
Ours (ResNet-34)	FF++	**0.818**	**0.864**	**0.851**	**0.721**

Table 2: Cross-dataset comparison results (frame-level AUC) on Celeb-DF-V1 (CD1), Celeb-DF-V2 (CD2), DFDC-Preview (DFDC-P), and DFDC. We train our method on FF++ (C23) and test it on other benchmark datasets. The 'PD' means private data. * is collected from (Dong et al. 2022b; Cao et al. 2022; Sun et al. 2022), and other results are directly cited from the corresponding original paper. The bold and underline mark the best and second performances, respectively.

is that our method is more robust, with significant improvement when tested on unseen datasets. This reflects that our model has a better capability for uncovering forgery clues. The improvements in generalizability can be attributed to the information bottleneck in our framework design, where our model demonstrates a better capacity for identifying different forms of deepfake artifacts instead of merely the instances specific to the training dataset. Overall, our method achieves state-of-the-art frame-level and video-level generalization performance. For frame-level comparisons, our method attains 0.818 and 0.857 AUCs on CD1 and CD2 respectively, outperforming the current state-of-the-art method ICT. Our method also improves the AUC on DFDC-P from 0.833 (OST) to 0.851, and on DFDC from 0.691 (RECCE) to 0.721. In contrast with video-level methods, our method surpasses the current state-of-the-art technique, AUNet, in terms of AUC with scores of 0.936, 0.902, and 0.754 on CD2, DFDC-P, and DFDC, respectively. Remarkably, despite employing solely traditional data augmentation techniques, our approach attains state-of-the-art performance across all four benchmarks, surpassing models (such as AUNet, SBIs, and ICT) trained on private augmented datasets.

4.3 Ablation Study

In this section, we first study the effectiveness of our two information losses, *i.e.,* Local Information Loss $\mathcal{L}_{\mathrm{LIL}}$ and Global Information Loss $\mathcal{L}_{\mathrm{GIL}}$. We then explore the impact of local feature quantity within the local information block.

Ablation Study on Information Losses We study the effects of removing $\mathcal{L}_{LIL}$ and $\mathcal{L}_{GIL}$ in our method. We train models on FF++ (C23) and test them on CD2. Tab. 4 demonstrates the results of the ablation study on the proposed two information losses. Clearly, we see that $\mathcal{L}_{LIL}$ and $\mathcal{L}_{GIL}$ play key roles in performance improvement over in-dataset and cross-dataset settings. The AUC improvement of using the proposed losses is more critical in *cross-dataset* than *in-dataset* settings. This empirical evidence suggests that incorporating the proposed losses may lead to extracting broader clues. Quantitatively, $\mathcal{L}_{LIL}$ and $\mathcal{L}_{GIL}$ have a dominant con-

Method	Training dataset	CD2	DFDC-P	DFDC
Xception (Rossler et al. 2019)	FF++	0.737*	0.679*	0.709*
F^3-Net (Qian et al. 2020)	FF++	0.757*	-	0.709*
PCL+I2G (Zhao et al. 2021b)	PD	0.900	0.744	0.675
FST-Matching (Dong et al. 2022a)	FF++	0.894	-	-
LipForensics (Haliassos et al. 2021)	FF++	0.824	-	0.735
FTCN (Zheng et al. 2021)	FF++	0.869	0.740	0.710*
Luo.et al. (Luo et al. 2021)	FF++	-	0.797	-
ResNet-34+ SBIs (Shiohara and Yamasaki 2022)	PD	0.870	0.822	0.664
EFNB4+ SBIs (Shiohara and Yamasaki 2022)	PD	<u>0.932</u>	<u>0.862</u>	0.724
RATF (Gu et al. 2022b)	FF++	0.765	0.691	-
Li.et al. (Li et al. 2022)	FF++	0.848	0.785	-
AltFreezing (Wang et al. 2023)	FF++	0.895	-	-
AUNet (Bai et al. 2023)	PD	0.928	<u>0.862</u>	<u>0.738</u>
Ours (ResNet-34)	FF++	**0.936**	**0.902**	**0.754**

Table 3: Cross-dataset comparison results (video-level AUC) on Celeb-DF-V2 (CD2), DFDC-Preview (DFDC-P), and DFDC. We train our method on FF++ (C23) and test it on other benchmark datasets. The 'PD' means private data. * is collected from (Shiohara and Yamasaki 2022; Bai et al. 2023), and other results are directly cited from the corresponding original paper. The bold and underline mark the best and second performances, respectively.

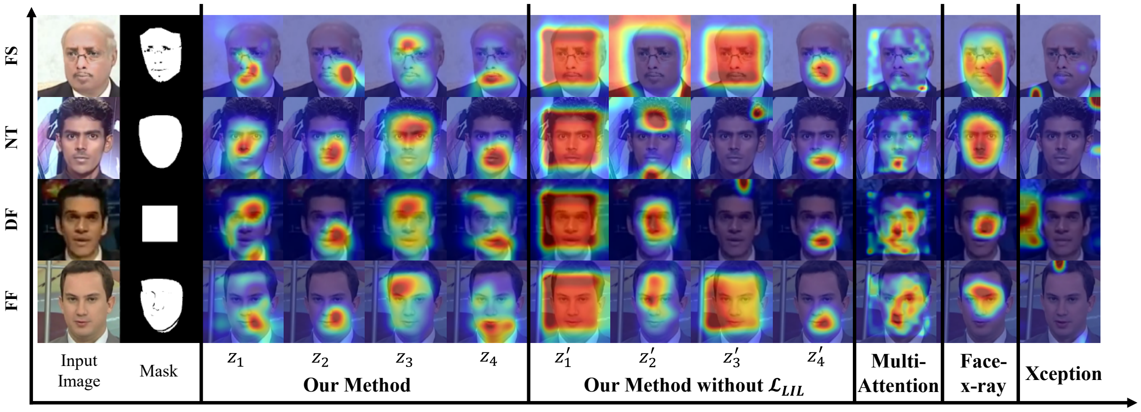


Figure 4: Visual examples of our method on various types of forgery methods within FF++ (C23), *i.e.,* Deepfakes (DF), Face2Face (FF), FaceSwap (FS) and NeuralTextures (NT). Comparison between our method with and without $\mathcal{L}_{LIL}$, Multi-Attentional, Face-x-ray, and Xception.

ID	Loss		FF++ (C23)		CD2	
	$\mathcal{L}_{\text{LIL}}$	$\mathcal{L}_{\text{GIL}}$	ACC↑	AUC↑	ACC↑	AUC↑
1	-	✓	94.26	0.979	78.79	0.840
2	✓	-	94.28	0.977	76.96	0.827
3	-	-	93.53	0.966	77.29	0.816
4	✓	✓	**94.98**	**0.983**	**80.70**	**0.864**

Table 4: Ablation study of the proposed $\mathcal{L}_{\text{LIL}}$ and $\mathcal{L}_{\text{GIL}}$ for our method. We show frame-level ACC (%) and AUC training on FF++ (C23) and testing on Celeb-DF-V2 (CD2). The bold mark best performance.

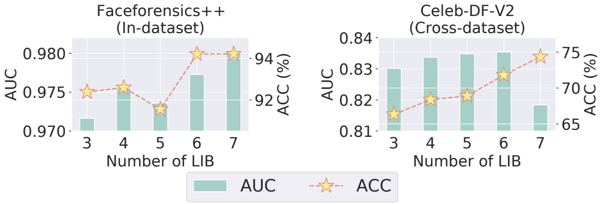


Figure 5: In-dataset and cross-dataset performance effects within different numbers of LIBs. We train models on FF++ (C23) with 10 epochs and test them on Celeb-DF-V2.

tribution to our method, with AUC on FF++ improving from 0.966 (without both) to 0.983 (with both) and AUC on CD2 from 0.816 to 0.864. The absence of either loss will bring about a significant drop in model performance.

Ablation Study on local information block We first investigate the effect of varying the number of local information blocks (LIB), *i.e.* local feature quantity. We train models on FF++ (C23) and test them on CD2, and report the frame-level AUC and ACC. The number of LIB is varied from three to seven, while other hyper-parameters are fixed. Fig. 5 shows the results for different numbers of LIB. We observe that both in-dataset and cross-dataset performance improve with an increasing number of LIB increases. However, when the number of LIBs becomes excessively high ($n = 7$), the model's generalization performance experiences a significant decline. This aligns with our intuition, as gradually augmenting the number of LIBs enlarges the number of trainable network parameters, directly affecting the in-dataset performance. Simultaneously, this expansion results in a rise in local feature quantity, contributing to the enhancement of the model's generalization performance. Nevertheless, as the number of LIBs continues to rise, an overabundance of parameters induces model overfitting, ultimately diminishing the model's capacity for generalization.

4.4 Visualization

To further assess the model interpretability and the efficacy of the Local Information Loss $\mathcal{L}_{LIL}$, we visualize four samples subjected to various forgery methods on FF++. We apply Grad-CAM (Selvaraju et al. 2017) for representation visualization. As shown in Fig. 4, our approach offers several noteworthy insights. Firstly, it becomes evident that our method excels in extracting more forgery clues. Other detection techniques fixate on specific regions, disregarding subtle cues present elsewhere. This leads to confined regions of focus for detection. The second insight reveals our method focuses on different forgery regions with little overlap. It provides evidence that the orthogonality within extracted local representations. Specifically, our method identifies manipulated cues in the nose, cheek, forehead, and mouth, corresponding to z_1 through z_4 respectively. In contrast, results without $\mathcal{L}_{LIL}$ depict local representations possess an imbalanced capacity to signify forgery features. While some local representations contain ample information (z'_1), others offer duplicated (z'_3) or scanty (z'_2 and z'_4) forgery-related clues.

4.5 Limitations

Our method is a purely data driven approach relying on information theoretic constraints to search for forgery clues. For some classes or forgeries, employing prior knowledge as guidance could be more optimal. For future work, we seek to incorporate heuristic guidance into our model, which could further boost performance and interpretability.

5 Conclusion

In this paper, we propose an information bottleneck based framework for deepfake detection, which aims to extract broader forgery clues. In this context, we derive local information losses to obtain task-related independent local features. We further theoretically analyze the global information objective to aggregate local features into a sufficient and purified global representation for classification. Extensive experiments demonstrate that our method achieves state-of-the-art *in-dataset* and *cross-dataset* performance on five benchmark datasets, indicating its potential as a reliable solution for deepfake detection in various real-world scenarios.

Acknowledgements

This work is partially supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020010, the National Natural Science Foundation of China (62172359, 62372402, 61972348, 62102354), the Key R&D Program of Zhejiang Province (2023C01217), the Fundamental Research Funds for the Central Universities (No. 2021FZZX001-27), and the Hangzhou Leading Innovation and Entrepreneurship Team (TD2020003).

References

Afchar, D.; Nozick, V.; Yamagishi, J.; and Echizen, I. 2018. Mesonet: a compact facial video forgery detection network. 1–7.

Ba, Z.; Wen, Q.; Cheng, P.; Wang, Y.; Lin, F.; Lu, L.; and Liu, Z. 2023. Transferring Audio Deepfake Detection Capability across Languages. In *Proceedings of the ACM Web Conference 2023*, 2033–2044.

Bai, W.; Liu, Y.; Zhang, Z.; Li, B.; and Hu, W. 2023. AUNet: Learning Relations Between Action Units for Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24709–24719.

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

Chen, L.; Zhang, Y.; Song, Y.; Liu, L.; and Wang, J. 2022a. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. 18710–18719.

Chen, L.; Zhang, Y.; Song, Y.; Wang, J.; and Liu, L. 2022b. Ost: Improving generalization of deepfake detection via one-shot test-time training. *Advances in Neural Information Processing Systems*, 35: 24597–24610.

Chiou, M.-J.; Liu, Z.; Yin, Y.; Liu, A.-A.; and Zimmermann, R. 2020. Zero-shot multi-view indoor localization via graph location networks. In *Proceedings of the 28th ACM International Conference on Multimedia*, 3431–3440.

Chugh, K.; Gupta, P.; Dhall, A.; and Subramanian, R. 2020. Not made for each other-audio-visual dissonance-based deepfake detection and localization. 439–447.

Deng, J.; Guo, J.; Ververas, E.; Kotsia, I.; and Zafeiriou, S. 2020. Retinaface: Single-shot multi-level face localisation in the wild. 5203–5212.

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

Dong, S.; Wang, J.; Liang, J.; Fan, H.; and Ji, R. 2022a. Explaining deepfake detection by analysing image matching. In *European Conference on Computer Vision*, 18–35. Springer.

Dong, X.; Bao, J.; Chen, D.; Zhang, T.; Zhang, W.; Yu, N.; Chen, D.; Wen, F.; and Guo, B. 2022b. Protecting celebrities from deepfake with identity consistency transformer. 9468–9478.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging frequency analysis for deep fake image recognition. 3247–3258.

Fung, S.; Lu, X.; Zhang, C.; and Li, C.-T. 2021. Deepfakeucl: Deepfake detection via unsupervised contrastive learning. 1–8.

Gu, Z.; Yao, T.; Chen, Y.; Ding, S.; and Ma, L. 2022a. Hierarchical Contrastive Inconsistency Learning for Deepfake Video Detection. 596–613.

Gu, Z.; Yao, T.; Yang, C.; Yi, R.; Ding, S.; and Ma, L. 2022b. Region-aware temporal inconsistency learning for deepfake video detection. 1.

Haliassos, A.; Vougioukas, K.; Petridis, S.; and Pantic, M. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. 5039–5049.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. 770–778.

Hu, J.; Liao, X.; Liang, J.; Zhou, W.; and Qin, Z. 2022. Finfer: Frame inference-based deepfake detection for high-visual-quality videos. 36(1): 951–959.

Hua, Y.; Shi, R.; Wang, P.; and Ge, S. 2023. Learning Patch-Channel Correspondence for Interpretable Face Forgery Detection. *IEEE Transactions on Image Processing*, 32: 1668–1680.

Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; and Ye, D. 2023. Implicit Identity Driven Deepfake Face Swapping Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4490–4499.

Li, J.; Xie, H.; Yu, L.; and Zhang, Y. 2022. Wavelet-enhanced Weakly Supervised Local Feature Learning for Face Forgery Detection. 1299–1308.

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

Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020a. Face x-ray for more general face forgery detection. 5001–5010.

Li, Y.; and Lyu, S. 2018. Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020b. Celeb-df: A large-scale challenging dataset for deepfake forensics. 3207–3216.

Liebel, L.; and Körner, M. 2018. Auxiliary tasks in multi-task learning. *arXiv preprint arXiv:1805.06334*.

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. 772–781.

Liu, Z.; Qian, P.; Yang, J.; Liu, L.; Xu, X.; He, Q.; and Zhang, X. 2023. Rethinking smart contract fuzzing: Fuzzing with invocation ordering and important branch revisiting. *IEEE Transactions on Information Forensics and Security*, 18: 1237–1251.

Liu, Z.; Wu, S.; Xu, C.; Wang, X.; Zhu, L.; Wu, S.; and Feng, F. 2022. Copy Motion From One to Another: Fake Motion Video Generation. *arXiv preprint arXiv:2205.01373*.

Luo, Y.; Zhang, Y.; Yan, J.; and Liu, W. 2021. Generalizing face forgery detection with high-frequency features. 16317–16326.

Masi, I.; Killekar, A.; Mascarenhas, R. M.; Gurudatt, S. P.; and AbdAlmageed, W. 2020. Two-branch recurrent network for isolating deepfakes in videos. 667–684.

McGill, W. 1954. Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4): 93–111.

Nirkin, Y.; Keller, Y.; and Hassner, T. 2019. Fsgan: Subject agnostic face swapping and reenactment. 7184–7193.

Pan, K.; Yin, Y.; Wei, Y.; Lin, F.; Ba, Z.; Liu, Z.; Wang, Z.; Cavallaro, L.; and Ren, K. 2023. DFIL: Deepfake Incremental Learning by Exploiting Domain-invariant Forgery Clues. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8035–8046.

Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; and Tucker, G. 2019. On variational bounds of mutual information. 5171–5180.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. 86–103.

Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Nießner, M. 2019. Faceforensics++: Learning to detect manipulated facial images. 1–11.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. 618–626.

Shiohara, K.; and Yamasaki, T. 2022. Detecting deepfakes with self-blended images. 18720–18729.

Shuai, C.; Zhong, J.; Wu, S.; Lin, F.; Wang, Z.; Ba, Z.; Liu, Z.; Cavallaro, L.; and Ren, K. 2023. Locate and Verify: A Two-Stream Network for Improved Deepfake Detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 7131–7142.

Song, X.; Chen, J.; and Jiang, Y.-G. 2023. Relation Triplet Construction for Cross-modal Text-to-Video Retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4759–4767.

Sun, K.; Yao, T.; Chen, S.; Ding, S.; Li, J.; and Ji, R. 2022. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2316–2324.

Tian, X.; Zhang, Z.; Lin, S.; Qu, Y.; Xie, Y.; and Ma, L. 2021. Farewell to mutual information: Variational distillation for cross-modal person re-identification. 1522–1531.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Wang, J.; Sun, Y.; and Tang, J. 2022. LiSiam: Localization invariance Siamese network for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 17: 2425–2436.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; and Li, H. 2023. AltFreezing for More General Video Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4129–4138.

Wei, Y.; Sun, Y.; Zheng, R.; Vemprala, S.; Bonatti, R.; Chen, S.; Madaan, R.; Ba, Z.; Kapoor, A.; and Ma, S. 2023. Is Imitation All You Need? Generalized Decision-Making with Dual-Phase Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16221–16231.

Wu, Y.; Song, X.; Chen, J.; and Jiang, Y.-G. 2023. Generalizing Face Forgery Detection via Uncertainty Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1759–1767.

Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; and Ge, S. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, 5384–5392.

Zhang, D.; Li, C.; Lin, F.; Zeng, D.; and Ge, S. 2021. Detecting Deepfake Videos with Temporal Dropout 3DCNN. 1288–1294.

Zhang, X.; Hong, H.; Hong, Y.; Huang, P.; Wang, B.; Ba, Z.; and Ren, K. 2023. Text-CRS: A Generalized Certified Robustness Framework against Textual Adversarial Attacks. In *2024 IEEE Symposium on Security and Privacy (SP)*, 53–53. IEEE Computer Society.

Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021a. Multi-attentional deepfake detection. 2185–2194.

Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; and Xia, W. 2021b. Learning self-consistency for deepfake detection. 15023–15033.

Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; and Wen, F. 2021. Exploring temporal coherence for more general video face forgery detection. 15044–15054.