Detecting forgery videos is highly desirable due to the abuse of deepfake. Existing detection approaches contribute to exploring the specific artifacts in deepfake videos and fit well on certain data. However, the growing technique on these artifacts keeps challenging the robustness of traditional deepfake detectors. As a result, the development of these approaches has reached a blockage. In this article, we propose to perform deepfake detection from an unexplored voice-face matching view. Our approach is founded on two supporting points: first, there is a high degree of homogeneity between the voice and face of an individual (i.e., they are highly correlated), and second, deepfake videos often involve mismatched identities between the voice and face due to face-swapping techniques. To this end, we develop a voice-face matching method that measures the matching degree between these two modalities to identify deepfake videos. Nevertheless, training on specific deepfake datasets makes the model overfit certain traits of deepfake algorithms. We instead advocate a method that quickly adapts to untapped forgery, with a pre-training then fine-tuning paradigm. Specifically, we first pre-train the model on a generic audio-visual dataset, followed by the fine-tuning on downstream deepfake data. We conduct extensive experiments over three widely exploited deepfake datasets: DFDC, FakeAVCeleb, and DeepfakeTIMIT. Our method obtains significant performance gains as compared to other state-of-the-art competitors. For instance, our method outperforms the baselines by nearly 2%, achieving an AUC of 86.11% on FakeAVCeleb. It is also worth noting that our method already achieves competitive results when fine-tuned on limited deepfake data.

CCS Concepts: • Security and privacy → Human and societal aspects of security and privacy; • Applied computing → Computer forensics; • Computing methodologies → Artificial intelligence;

Additional Key Words and Phrases: Deepfake detection, cross-modal matching, voice, face

ACM Reference format:
Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. 2023. Voice-Face Homogeneity Tells Deepfake. ACM Trans. Multimedia Comput. Commun. Appl. 20, 3, Article 76 (November 2023), 22 pages.
https://doi.org/10.1145/3625231

Authors’ addresses: H. Cheng and Q. Li, School of Computer Science and Technology, Shandong University, 266237, Qingdao, China; e-mails: [xaCheng1996, iliqi.nice]@gmail.com; Y. Guo, School of Computing, National University of Singapore, 119391, Singapore; e-mail: guoyang.eric@gmail.com; T. Wang, Department of Computer Science, The University of Hong Kong, 999077, Hong Kong, China; e-mail: terry.ai.wang@gmail.com; X. Chang, Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia; e-mail: cxj273@gmail.com; L. Nie, School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), 518055, Shenzhen, China; e-mail: nieliqiang@gmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
1551-6857/2023/11-ART76 $15.00
https://doi.org/10.1145/3625231
1 INTRODUCTION

The role of deepfake [24, 30, 34, 47] is to synthesize the media in which a person is replaced with someone else’s portrait or vocal. Given its successful application in animation [67] and online education [51], deepfake has attracted increasing interest from academic and industrial practitioners. However, the abuse of such techniques, such as maliciously editing porn and violent videos, seriously challenges social functioning and ethics, spawning widespread concerns. It is hence imperative to detect deepfake abuse with effective measures.

Considerable research efforts have been dedicated to detecting deepfakes thus far [17, 19, 40, 58, 72]. Most of them explore the face manipulation artifacts of fake videos, including the visual artifacts from face attentive regions [73], apparent changes in the frequency domain resulting from up-sampling [37], or amplified artifacts via isolating manipulated faces [16]. Nevertheless, these methods are all limited by one critical downside, namely inferior generalization across datasets. For instance, a model trained on the FF++ dataset [53] suffers significant performance degradation when migrated to other deepfake datasets (e.g., DFDC [9] or Celeb-DF [36]). The key reason is that different datasets are built with distinctive manipulation algorithms. As a result, previous detection approaches tend to fit well on the specific training data, and the generalization is thereby hampered. For instance, Figure 1 illustrates that the models trained on FF++ pay more attention to the eyes yet fail on DFDC since the salient regions are instead the forehead and nose.

Existing solutions for this generally contribute from two directions. The first direction is to enhance the detection of manipulated traces with complementary modalities [31, 43, 57]. For instance, Zhou and Lim [74] leverage the speech content to unearth unharmonious mouth-related dynamics. However, these approaches focus on partial facial characteristics, such as lip movements [74] or emotional biases [43], which are easily attacked by specialized countermeasures [51]. The other direction is to extend current datasets with auxiliary visual features, such as blending regions [33] or facial landmarks [20]. Since these features are not fully covered by existing forgery algorithms, methods from this direction make it easy to spot the salient differences between real and fake frames. Nevertheless, building such datasets is time consuming, and the models often require computationally intensive training on both forgery and auxiliary data.

To overcome the shortcomings encountered by the existing studies, this article is devoted to addressing this task from an unexplored angle. Our solution is inspired by the recent progress from cross-modal biometric matching [44, 49], which certifies the fact that an individual’s voice and face are strongly correlated. Specifically, a person’s voice features, such as pitch and timbre, are closely associated with their facial features. Building on this observation, we conduct an in-depth analysis of the existing deepfake algorithms and recognize a serious voice-face mismatching problem. For example, specific deepfake videos may replace the speaker’s face with that of someone else, thereby creating a dissonance between the visual representation and the accompanying voice. This motivates us to speculate—can we perform deepfake detection from a discrimination view of voices and faces?

To answer this question, we tentatively propose a practical method called **Voice-Face Matching Detection (VFD)**. In particular, we perform matching between voices and faces rather than directly attacking the artifacts from deepfake. An overview of VFD is shown in Figure 2. We first train VFD on a generic audio-visual dataset (e.g., VoxCeleb2 [8]) without being manipulated by deepfake algorithms. Since this dataset is collected from diverse identity resources, we can strengthen VFD’s ability in modeling homogeneity. Specifically, given an anchor voice clip, the popular InfoNCE loss [61] is employed to pull matched voice clip and face instances closer while pushing

---

1. https://www.bbc.com/news/technology-42912529

ACM Trans. Multimedia Comput. Commun. Appl., Vol. 20, No. 3, Article 76. Publication date: November 2023.
Fig. 1. Heat maps and AUC scores (%) from Xception under different settings. When trained on FF++, the attention is mostly placed on eye regions and is shifted to the forehead and nose for DFDC. Since the editing algorithm focuses on different regions with respect to distinctive data, a well-trained model thus cannot adapt smoothly across datasets.

Fig. 2. Schematic illustration of our proposed VFD model. In the pre-training stage (a), we adopt the InfoNCE loss to pull matched voices (blue triangles) and faces (blue circles) closer while pushing unmatched faces (orange circles) far away. For fine-tuning (b), the upgraded RFC loss is employed to fine-tune the model on the deepfake dataset. The cosine similarity serves as evidence for judging the videos’ authenticity during inference.

unmatched ones farther away. As the pre-training dataset contains rich data for building an informative feature extractor, our VFD can easily adapt to various downstream deepfake datasets with few fine-tuning steps. On the basis of InfoNCE, we design an upgraded Real-Fake Contrastive (RFC) loss function to align the objective between pre-training and fine-tuning, given that the downstream deepfake dataset contains real (positive) and fake (negative) pairs. Arguably, our method offers a best-of-both-worlds solution. First, VFD focuses on the general matching objective of voices and faces and can be quickly migrated to various deepfake datasets, as opposed to paying attention to designated face regions (see Figure 1). Second, the pre-training then fine-tuning paradigm alleviates the requirement for auxiliary data. Our model builds upon easily collected generic videos and depends less on the expensive auxiliary data from deepfake.
We conduct extensive experiments over three widely exploited deepfake datasets: DFDC, FakeAVCeleb [26], and DeepfakeTIMIT [28]. The results demonstrate that our VFD achieves state-of-the-art performance—for instance, AUCs of 85.13% and 86.11% on DFDC and FakeAVCeleb, respectively. In addition, our method achieves remarkably competitive results with few deepfake data for fine-tuning as compared to some strong baselines.

The main contributions of this work are threefold:

- We address the deepfake detection from a voice-face matching view. Specifically, we perform deepfake detection via the intrinsic correlation of facial and audio excluding any additional auxiliary data, either more modalities or more visual features.
- We devise an effective multi-modal matching framework to justify real and fake videos. On the basis of the matching view, we enhance the traditional contrastive loss to align the objective between generic and deepfake datasets.
- Comprehensive quantitative and ablative experiments demonstrate that our method produces significant performance gains over a variety of state-of-the-art competitors. Further experiments demonstrate that our model can still achieve a practical detection capability even with limited fine-tuning data.

The rest of this article is structured as follows. Section 2 briefly reviews the related literature. Method intuition and architecture are presented in Section 3. Section 4 elaborates the experimental settings and results, followed by our conclusion and discussion in Section 5.

2 RELATED WORK

2.1 Deepfake

Benefiting from the continuous development of portrait synthesis, deepfake has recently emerged as a prevailing research problem. Existing algorithms either leverage the image only or the 3D information [50] to edit videos. The image-only methods synthesize fake faces for the target identities, which are then blended into the given video [2]. For example, Li et al. [32] utilized cascaded AAD blocks to integrate identities and face attributes within multiple feature levels, and realistic human faces can be generated. To achieve realistic face synthesis, Generative Adversarial Networks (GANs) [12] have grown in popularity due to their promising performance. StyleGAN [23] modifies high-level facial attributes with a progressive growing approach and adaptive instance normalization. IPGAN [69] disentangles the identity and attributes of the source and target faces, respectively. These two are thereafter blended for identity-preserving face synthesis. Different from these methods, Kim et al. [27] applied 3DMM [3] to produce the portraits with controllable poses. HifiFace [64] generates photo-realistic videos via the 3D shape-aware identity extractor. However, the existing deepfake approaches pay much attention to the face regions while the voice-face consistency is hard to maintain.

2.2 Deepfake Detection

Deepfake detection is often cast as a binary (real or fake) classification task. Preliminary efforts often endeavor to detect the specific traces of manipulation [11, 13, 18, 56]. Masi et al. [41] proposed a two-branch network to separately extract artifacts of color and frequency domains [37]. SSTNet [66] detects edited faces through spatial, steganalysis, and temporal features. Contrary to these approaches utilizing the vision modality only, studies today exploit the multi-modal information [14] for deepfake detection [38, 70, 73]. For instance, lip-syncing and dubbing models [29] are employed to identify the audio-visual inconsistency from a speaker. Zhou and Lim [74] predicted the probability of voices and faces being edited to judge video credibility—to this end, they sufficiently trained the model by modifying the distribution of the existing datasets. Mittal et al. [43]
extracted the emotional biases that video and audio jointly mention, based on which the detection objective can be achieved. Previous approaches have gained certain improvements on some datasets. Nonetheless, when transferring to unknown data, inferior performance is often confronted.

To address this lack of generalization issue, several cross-dataset detection approaches are proposed. Li et al. [33] constructed auxiliary data from extracted blending regions in large-scale videos to enhance the robustness. Haliassos et al. [20] utilized a pre-trained lip-reading model to explore the irregularities in mouth movements, followed by delicate fine-tuning on the forgery data. Nonetheless, these methods always require auxiliary data and yield increased training overload. We propose tackling deepfake detection from a novel matching view in this work. Specifically, the matching between voices and faces is taken as the proxy for discriminating real and fake videos, since the voices and faces show a severe mismatch from deepfake algorithms. Our method exhibits promising generalization over various deepfake datasets, which requires only paired voices and faces for training.

2.3 Cross-Modal Biometric Matching

Cross-modal biometric matching aims to retrieve the corresponding video for a given audio from multiple candidates, or vice versa [44]. Among the initial efforts, researchers extracted video and audio features via pre-trained models and then employed cross entropy [6, 44] or cosine loss [15, 21] to measure the matching degree. Later studies take into consideration the interactions among different modalities. For instance, Wen et al. [65] devised a two-level loss, which leverages both local and global features on modality alignment. ADSM [7] adopts an adversarial matching network to extract the high-level semantical features. A specially designed discriminator is then employed to bridge the voice and face gap while maintaining semantic consistency. Moreover, Speech2Face [49] applies a pre-trained face decoder network to reconstruct the face from speech clips. The methods in this category indeed provide certain support that the voices and faces of the same person are strongly correlated.

3 METHODOLOGY

In this section, we first present the evidence from two aspects for the intuition of our method, followed by a detailed introduction of our multi-modal matching model pre-trained on generic audio-visual data. We end this section with the fine-tuning and detection on downstream deepfake datasets.

3.1 Method Intuition

Research on human perception and neurology has shown that humans can outline pictures of a person’s appearance based on the voices, or vice versa [22]. In other words, voices and faces exhibit a high degree of homogeneity in the human brain and are associated closely with identities. In light of this, we contribute to exploring whether the matching between these two can serve as a proxy in detecting deepfake. To this end, we perform some probing tests, which mainly answer the following two questions:

- Q1: Are the voices and faces matched in deepfake videos?
- Q2: Can the voices and faces be leveraged to discriminate different identities?

3.1.1 Voice-Face Mismatching in Deepfake Data (Q1). We evaluate whether the voices and faces are matched in deepfake videos via measuring their similarity. In particular, we randomly sample 1,800 real and 4,000 corresponding fake videos from DFDC. Two plain transformer-based [44, 63] models are employed to extract the voice and face features. Specifically, we leveraged two feature...
extractors for this experiment, VIT-B/32 and VIT-B/16 \[10\], resulting in four groups of features in total. These extractors share a similar 12-layer self-attention structure, followed by a linear layer that maps the output to the same dimension. To pre-train the extractors, we used the VoxCeleb2 dataset, which is a general-purpose audio-visual dataset, and subsequently fine-tuned them on the DFDC dataset. Thereafter, we calculate the corresponding cosine similarity between features as follows:

\[
\text{similarity} = \frac{v \cdot f}{\max(||v||_2 \cdot ||f||_2, \epsilon)},
\]

where \(v\) and \(f\) are voice and face features, respectively. \(\epsilon\) is a small value to avoid division by zero.

We present the results in Figure 3, which shows the comparison of feature similarity among different groups of features. From this figure, we make the following observations. First, the voice-face similarity in real videos is much larger than those in fake ones, which is evident across different extractors. In other words, the voice and face from real videos match better in the feature space, whereas fake videos demonstrate serious mismatch evidence. Second, a small fraction of real videos share similar values with fake ones, which is partially due to the presence of ambient noise during recording.

3.1.2 Voice Discrimination over Identities (Q2). Faces encode essential cues for distinguishing different identities, which has been extensively proven by considerable studies \[55, 68\]. To testify whether the voices can achieve similar effects, we extract the voice features and cluster them in the following way. We employ the VGG-based \[59\] model as the backbone and the InfoNCE loss to learn the voice features, where the voice clips from the same identity are deemed as positive and from other identities as negative. In the second step, we randomly sample 5,120 voice clips from 163 people in the VoxCeleb2 dataset and show the voice feature manifold embedding in Figure 4.

Figure 4 reveals that voices from the same identity tend to cluster together with apparent boundaries, demonstrating the discrimination capability of voices. For details, we zoom in the bottom right region and highlight 10 identities whose associated videos are over 40. This detailed view illustrates that the learned features correlate tightly with group characteristics. For example, the female identities (id00656, id08130, id00902, id07039, and id03059) distribute on the top left. Moreover, given the same gender, the voice features can also discriminate ages. For example, id01262, id01746, and id00266 share similar ages (in their 60s), driving the voice features closer (top right region).

3.2 VFD for Pre-Training

Based on the preceding findings that a single person’s voices and faces have certain homogeneity, in this work we propose to detect deepfake videos by judging the matching degree of these two. To achieve this goal, we design a VFD method. As shown in Figure 5(a), our VFD is embodied with a dual-stream network, wherein the voices and faces are separately processed. Two modulators (i.e.,
Fig. 4. t-SNE [62] visualization of voice features from 163 identities. We zoom in the bottom right region and highlight 10 identities with more than 40 video instances. Note that the color of each cluster is unique even though some clusters may look visually the same due to the excessive number of identities.

Fig. 5. Overall architecture of the voice-face matching network (a) and the voice/face feature extractor (b).

face and voice modulators) are employed to guide the extractors to focus on the identity-related features, followed by a matching function (e.g., InfoNCE loss [61]) to determine the matching degree.

We first train the model on a generic audio-visual dataset. Our motivation for implementing pre-training stems from two aspects. First, learning cross-modality homogeneity has long been recognized non-trivial, which demands large quantity of multi-modal paired data [5, 65, 71]. Second, different deepfake datasets focus on distinctive forgery angles, and the volume of one dataset is often less sufficient to train a generalizable matching model [9, 26, 45]. Therefore, directly training VFD leads to sub-optimal performance, as revealed in Section 4.7. We instead train our VFD on a generic dataset. In what follows, we first present the data preprocessing protocol. After that, we elaborate on the overall architecture of our VFD and its corresponding pre-training strategy sequentially.

3.2.1 Data Preprocessing. For each input audio \( \mathcal{A} \), we extract a 3-second voice clip with a 16-kHz sample rate, which will be represented as a spectrogram \( C \in \mathbb{R}^{L_1 \times H_1 \times W_1} \), where \( L_1 \) denotes the channel number, and \( H_1 \) and \( W_1 \) are the height and width of the spectrograms, respectively. However, for the video input \( \mathcal{V} \), we randomly leverage one random face frame \( I \in \mathbb{R}^{L_2 \times H_2 \times W_2} \) to represent it as each video involves one identity only [8, 9, 26].
The key to performing contrastive learning is to construct efficacious positive and negative samples. In the pre-training, we utilize the voice as the anchor and build <positive, negative > pairs from identity faces. Specifically, for a given voice clip \( C \), it is straightforward to sample the faces from the same identity as its positive \( I^+ \). As to the negative \( I^- \) sampling, we simply adopt the faces from other identities as the counterpart while leaving cumbersome hard negative mining as future work. After this, the voice and face features (both positive and negative) are inputted to two independent feature extractors, which is detailed as follows.

### 3.2.2 Feature Extraction

Extracting both voice and face features associated with identity is of vital importance for a multi-modal matching approach. To this end, an elaborated multi-modal feature extractor is presented, and the overall architecture is shown in Figure 5(b). In the next, we take the processing for the face input \( I \) as an example, and the voice feature extractor is designed in a similar fashion.

We first utilize a deep forward convolutional projection to cut the faces into patches and extract the initial face features:

\[
I_v(i, j, m) = \sum_{u, v=0}^{N} \sum_{t=0}^{F} h(u, v, t) \otimes I(i-u, j-v, m-t),
\]

where \( N \) is the image size (the height and width are identical in face images), \( F \) denotes the output channel number, and \( h(\cdot) \) represents the convolutional kernel. In this way, we obtain the \( F \)-channel feature maps \( I_v \in \mathbb{R}^{F \times D} \), representing the modality features.

It is non-trivial to learn the overall identity features of voices or faces with the preceding module, as only local information is aggregated by this fashion while the global context is overlooked. Therefore, we develop another novel transformer-like [10] module to learn the identity features based on the initial multi-modal representations, wherein the self-attention mechanism is employed to extract non-local joint features. To this end, we first build a new feature matrix \( M_l \) via concatenating a learnable vector \( c \in \mathbb{R}^{D} \) with \( I_v \) to gather identity-related features from the modal representations. Subsequently, the newly constructed map \( M_l \) will be element-wisely added by a positional encoding \( E \) for retaining the positional information of the image feature maps:

\[
M_p = M_l \oplus E,
\]

where \( M_p, M_l, \) and \( E \in \mathbb{R}^{F' \times D}, F' = F + 1 \). Then, we map \( M_p \) into three matrices \( M_q, M_k, \) and \( M_v \), which share a same size of \( F' \times D \). We then utilize the self-attention mechanism to perform non-local learning:

\[
g(M_q, M_k, M_v) = \text{softmax} \left( \frac{M_q M_k^T}{\sqrt{D}} \right) M_v.
\]

We perform this block several times and finally map the outputs of \( g(\cdot, \cdot, \cdot) \) into the \( K_f \in \mathbb{R}^{F' \times V} \) with a linear transformation. In this calculation, the vector \( c \) has aggregated the global identity related information via the self-attention mechanism [10]. We thus detach \( c \) from \( K_f \) and reformulate this vector as \( f_v \in \mathbb{R}^{V} \) to represent the features of the whole face frame. In a similar fashion, the voice branch will transform voice input to \( K_v \) with the same shape as \( K_f \), and a feature vector \( f_v \) would be selected as the representative. Note that due to the modality gap between voices and faces [44], we do not share parameters between these two extractors but instead expect the InfoNCE loss (see the following) to enhance the homogeneity modeling.

\[^2\]It is also feasible to choose faces as anchors, which will serve as a possible extension of this work.
3.2.3 Pre-Training Protocol

We employ three loss functions to update the parameters of VFD. The first two are naive cross-entropy loss, and the last one is the InfoNCE loss [61] for judging matchness.

The cross-entropy loss is applied to both the voice and face modulators for binding identity information to the feature vectors. Specifically, we utilize the identity in the given video as the label for feature extraction:

\[
L_{cls} = -\log \frac{\exp \left(W_y f\right)}{\sum_{i=1}^{C} \exp \left(W_i f\right)},
\]

where \(C\) is the number of identities, \(W\) is a weighted matrix to calculate the feature vector \(f\) or \(f^v\) to the propensity of a particular identity, and \(y\) is the label. In this manner, the voice and face extractors will be empowered to extract identity-associated features from the input modalities. Thereafter, the InfoNCE loss [61] is applied to achieve the homogeneity as follows:

\[
L_I = \mathbb{E}_{f^+ \sim p^+, f^- \sim p^-} \left[ -\log \frac{e^{d(f^v, f^+_f)/\tau}}{e^{d(f^v, f^-_f) + \sum_{i=1}^{U} e^{d(f^v, f^-_i)}}} \right],
\]

where \(f^+_f\) corresponds to a positive instance feature extracted from the matched face image \(I^+_f\); \(f^-_f\) is the negative instance feature as described in Section 3.2.1; \(U\) is the number of negative samples; \(\tau\) represents a hyperparameter temperature, and \(d(\cdot, \cdot)\) is defined as the cosine similarity in the joint space between voices and faces as shown in Equation (1).

By means of calculating the mutual information between \(d(f^v, f^+_f)\) and \(d(f^v, f^-_f)\), VFD pulls the matched voice clip \(f^v\) and the positive faces \(f^+_f\) closer while pushing the unmatched negative faces \(f^-_f\) farther away. The \(\tau\) contributes to make the positive and negative instances more separable.

3.3 VFD for Fine-Tuning

Benefiting from the large-scale pre-training, we then leverage the pre-trained model for downstream specific deepfake detection. Inspired by prior work [33], we advocate one that quickly adapts in a fine-tuning fashion to improve the model generalization. In particular, we transfer the pre-trained VFD model to deepfake videos where voices and faces might be from some mismatched identities. Notably, the pre-training has made the model quite sensitive to mismatched voices and faces. Therefore, applying part of the deepfake dataset in the fine-tuning stage is feasible to quickly adapt the model to deepfake data, wherein significant computational overhead is thereby circumvented.

During fine-tuning, the optimization of voice and face extractors is performed simultaneously. One might intuitively apply Equation (6) to perform homogeneity modeling on deepfake datasets. In this scenario, the matching voice \(f^v\) and face \(f^+_f\) are extracted from a real video, denoted as \(V^{real}\), whereas the mismatched face \(f^-_f\) is derived from a fake video \(V^{fake}\). Nevertheless, this leads to an issue since deepfake videos are created by interchanging the faces of genuine ones. Consequently, \(V^{fake}\) might be a product of a face swap with \(V^{real}\). This implies that \(f^+_f\) and \(f^-_f\) could represent the same face, leading to an indistinguishable difference between \(d(f^v, f^+_f)\) and \(d(f^v, f^-_f)\), thereby compromising the effectiveness of Equation (6). To better align the objectives of pre-training and fine-tuning phases, we propose an enhancement to the vanilla contrastive loss, introducing a straightforward RFC loss function. Specifically, we first calculate the similarity between voices and faces sourced from real and fake videos:

\[
\begin{align*}
    v_{pos} &= d(f^+^v, f^+_f), \\
    v_{neg} &= d(f^-^v, f^-_f),
\end{align*}
\]

[61] Gidaris et al., “Unsupervised Learning of Invariant Visual Representations using Moment Matching Networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
where $f^+_V$ and $f^-_V$ denote voice features extracted from $V_{\text{real}}$ and $V_{\text{fake}}$, respectively. As per the analysis in Section 3.1.1, the similarity between a real voice and face, denoted as $v_{\text{pos}}$, surpasses that of deepfake ones, denoted as $v_{\text{neg}}$. We thereby make the pre-trained model adapt to the manipulation patterns in the deepfake dataset according to the difference of similarity. This is achieved via a simple cross-entropy method:

$$L_{RFC} = -\sum_t \log \left( \text{softmax} \left( \left[ v_{\text{pos}}, v_{\text{neg}} \right]_{i=1}^Q \right) \right),$$

where $Q$ represents the count of selected fake videos. The label vector $t \in \mathbb{R}^Q$, $Q = Q + 1$, is a series of artificially generated tags wherein $t_i = 1$ and $t_i = 0$ are respectively assigned to the matched similarity $v_{\text{pos}}$ and the mismatched ones. RFC can effectively draw $f^+_V$ and $f^+_F$ closer and push $f^-_V$ and $f^-_F$ farther away—that is, the voices and faces in the real video will match better, whereas the fake ones will exhibit a high degree of inconsistency. In essence, RFC serves as an advanced version of traditional methods like InfoNCE, specifically tailored for the deepfake dataset. It estimates the mismatch between the fake voices and faces, thereby eliminating the need for manually selecting negative examples.

### 3.4 VFD for Deepfake Detection

On the basis of the aforementioned protocols, our VFD is pre-trained and then fine-tuned on a generic audio-visual dataset and a deepfake dataset, respectively. For deepfake detection, we extract the face images $I_d$ and voice clips $C_d$, and apply VFD to determine the matchness:

$$\text{Matching} = \begin{cases} \text{True}, & \text{VFD}(C_d, I_d) \geq \lambda, \\ \text{False}, & \text{VFD}(C_d, I_d) < \lambda, \end{cases}$$

where the output of $\text{VFD}(\cdot, \cdot)$ is the similarity computed via Equation (1) in the joint latent space, and $\lambda$ is selected from the validation set. According to Equation (9), the matching result of the $I_d$ and $C_d$ being $\text{True}$ denotes a matched voice-face pair, namely the input video is real. On the contrary, the $\text{False}$ result corresponds to fake videos edited by deepfake algorithms.

### 4 EXPERIMENT

#### 4.1 Dataset

We utilized the VoxCeleb2 dataset for the pre-training of our VFD model. Specifically, VoxCeleb2 is a generic audio-visual dataset collected from YouTube videos, containing more than 1 million utterances from 6,112 celebrities with an average duration of 7.44 seconds. We split them into training, validation, and testing sets with a ratio of 8:1:1. Thereafter, three deepfake datasets—DeepfakeTIMIT, DFDC, and FakeAVCeleb—are employed to fine-tune and evaluate the effectiveness of VFD:

- **DeepfakeTIMIT** [28] contains two subsets of fake videos, namely **Lower Quality (LQ)** via a 64 $\times$ 64 input/output size model and **Higher Quality (HQ)** with a 128 $\times$ 128 one. Each subset involves 16 similar-looking pairs of subjects, wherein each subject has 10 face-swapping videos. We selected the original videos from VidTIMIT [54] and all fake ones from DeepfakeTIMIT for our fine-tuning and testing. On average, the duration of these videos is 4.25 seconds. For each subset, we divided the videos into fine-tuning and testing sets according to the ratio of 4:1.

- **DFDC** [9] involves 23,654 real videos recorded from 960 identities and 104,500 fake videos. The dataset contains eight different forgery algorithms, including GAN-based methods like FSGAN [48] and StyleGAN [23], as well as traditional face-swapping techniques. Following
previous work [20, 43], we sampled 6,089 real and 32,245 fake videos as the fine-tuning set, whereas 1,700 real videos with corresponding 5,810 fake videos were considered the testing set due to computational overhead. The average duration of these videos is 10.02 seconds, which is consistent with the original dataset.

- As for FakeAVCeleb [26], it comprises 500 real videos and 19,500 fake videos with an average duration of 5.35 seconds. These forged videos are generated with various forgery methods, including FSGAN and Wav2Lip [51]. We used 391 real videos and the corresponding 16,869 fake ones as the fine-tuning set and the remaining part as the testing set. It is worth noting that the pre-training, fine-tuning, and testing sets in the datasets mentioned previously do not contain any duplicated identities.

### 4.2 Implement Details

We implemented our model with the PyTorch toolkit. The AdamW optimizer [39] is adopted with a base learning rate of $1 \times 10^{-4}$ and $5 \times 10^{-6}$ for parameter updating of pre-training and fine-tuning, respectively. The learning rate decays gradually following the training process with the weight decay 0.2. All model parameters are initialized using a random normal distribution with a mean of 0 and a standard deviation of 0.02. The model is trained with a mini-batch size of 128 on four Tesla V100 GPUs. The threshold $\epsilon$ used in Equation (1) is $1 \times 10^{-8}$, the $\tau$ employed in Equation (6) is set at 0.1, and the $\lambda$ in Equation (9) is $-0.1$ as selected in the validation set of VoxCeleb2. We use 12 transformer blocks with 12-head self-attention in identity feature extractors (i.e., the $N = 12$ in Figure 5(b)). The input face images are resized to 224 x 224, whereas the voice clips are represented as 512 x 300 spectrograms since each second of voice clips is divided into 100 small windows in the sliding window manner.

### 4.3 Compared Baselines and Evaluation Metrics

We compared our model with multiple state-of-the-art baselines regarding the metrics of ACC and AUC scores. They can be roughly divided into four groups:

- **Single modality** methods detect visual artifacts based on the vision modality only: (1) MesoInception-4 and (2) Meso-4 [1], (3) Capsule [46], (4) F³-Net [52], (5) Xception [53], (6) ViT [10], (7) EfficientNet [60], and (8) VGG16 [59].

- **Visual-auditory detection** models include (9) BA-TFD [4], and (10) Emotional Forensics [43]. In addition, the uni-modal baselines (11) MesoInception-4_MM, (12) EfficientNet_MM, and (13) VGG16_MM were reformulated as multi-modal versions via ensembling different classifiers [25].

- **Models using auxiliary data**, such as the facial landmarks, guided manipulation traces detection: (14) VA-MLP and (15) VA-LogReg [42], (16) FWA and DSP-FWA [35], and (17) Headpose [70].

- **Fine-tuning strategy-based** approaches include (18) Face X-ray [33], and (19) LipForensics [20]. In addition, we reproduced the traditional three uni-modal approaches, namely (20) Xception_F, (21) F³-Net_F, and (22) ViT_F, with a fine-tuning method, wherein the first two are pre-trained on the FF++ dataset and the ViT_F possesses the identical training strategy with VFD—that is, pre-training on VoxCeleb2 dataset and fine-tuning on DFDC and FakeAVCeleb.

The first group of the aforementioned baselines is the classic method of recognizing fake videos through visual artifacts. The subsequent ones aim to improve robustness and generalizability via introducing various auxiliary modalities or data.
Table 1. Performance (%) of VFD and Baselines on DFDC, FakeAVCeleb, and DeepfakeTIMIT

| Model               | Modality     | Auxiliary | DFDC         | FakeAVCeleb | DeepfakeTIMIT |
|---------------------|--------------|-----------|--------------|--------------|---------------|
|                     | Visual       | Audio     | ACC          | AUC          | ACC           | LQ     | HQ     |
| Meso-4 [1]          | ✓            | ×         | 49.23        | 52.92        | 43.65         | 49.17  | 62.10  | 55.25  |
| MesoInception-4 [1] | ✓            | ×         | 56.37        | 60.56        | 72.22         | 75.82  | 78.45  | 60.70  |
| EfficientNet [60]   | ✓            | ×         | –            | –            | 81.03         | –      | –      | –      |
| VGG16 [59]          | ✓            | ×         | –            | –            | –             | –      | –      | –      |
| Capsule [46]        | ✓            | ×         | 57.65        | 61.20        | 73.27         | 76.19  | 84.58  | 81.69  |
| Xception [53]       | ✓            | ×         | 73.09        | 75.52        | 71.67         | 76.19  | 97.90  | 95.48  |
| F³-Net [52]         | ✓            | ×         | 74.16        | 75.40        | 81.08         | 84.54  | 98.39  | 94.60  |
| Vit [10]            | ✓            | ×         | 74.97        | 76.05        | 74.35         | 80.49  | 99.59  | 98.84  |
| BA-TFD [4]          | ✓            | ✓         | –            | –            | 84.60         | –      | –      | –      |
| Emotional Forensics [43] | ✓      | ✓         | –            | –            | 84.40         | –      | –      | 96.30  | 94.90  |
| MesoInception_MM [25] | ✓       | ✓         | Ensemble    | –            | –             | –      | –      | –      |
| EfficientNet_MM [25] | ✓       | ✓         | Ensemble    | –            | –             | 63.18  | –      | –      |
| VGG16_MM [25]       | ✓            | ✓         | Ensemble    | –            | –             | 78.04  | –      | –      |
| VA-MLP [42]         | ✓            | ×         | Landmark    | –            | 61.90         | –      | 67.00  | 61.40  | 62.10  |
| VA-LogReg [42]      | ✓            | ×         | Landmark    | –            | 66.20         | –      | 67.90  | 77.00  | 77.30  |
| FWA [35]            | ✓            | ×         | Landmark    | –            | 72.70         | –      | –      | 99.90  | 93.20  |
| DSP-FWA [35]        | ✓            | ×         | Landmark    | –            | 75.50         | –      | –      | 99.90  | 99.70  |
| Headpose [70]       | ✓            | ×         | Landmark    | –            | 55.90         | –      | 49.00  | 55.10  | 53.20  |
| Xception_F [35]     | ✓            | ×         | Fine-tune   | 67.84        | 71.13         | 70.53  | 71.20  | 95.17  | 93.69  |
| F³-Net_F [35]       | ✓            | ×         | Fine-tune   | 71.29        | 72.76         | 75.13  | 78.46  | 98.25  | 96.66  |
| Vit_F [35]          | ✓            | ×         | Fine-tune   | 65.04        | 71.16         | 70.05  | 74.30  | 99.00  | 98.17  |
| Face X-ray (Net18-100K) [33] [35] | ✓  | ×  | Fine-tune  | 43.42        | 59.36         | 72.88  | 73.52  | 96.95  | 94.47  |
| Face X-ray (Net18-500K) [33] [35] | ✓  | ×  | Fine-tune  | 44.80        | 58.98         | 75.65  | 77.94  | 98.61  | 97.54  |
| Face X-ray (Net32-100K) [33] [35] | ✓  | ×  | Fine-tune  | 46.49        | 61.57         | 76.75  | 79.72  | 99.72  | 98.53  |
| LipForensics [20]   | ✓            | ×         | Fine-tune   | –            | 73.50         | –      | –      | –      | –      |
| VFD                 | ✓            | ✓         | Fine-tune   | 80.96        | 85.13         | 81.52  | 86.11  | 99.95  | 99.82  |

For comparison purposes, we reported the AUC on two subsets of DeepfakeTIMIT, namely LQ and HQ, following previous work [35]. ‡, the model is reproduced by ourselves; -, the authors did not report this metric on this dataset. The multi-modal models and their uni-modal versions are highlighted in pink, whereas the fine-tuning ones and their fine-tuning-free counterparts are highlighted in blue.

4.4 Performance Comparison

The results of baselines and our method on the three datasets are demonstrated in Table 1. It can be observed that VFD achieves state-of-the-art performance on all datasets, expressing the effectiveness of our method as well as the validity of tackling deepfake detection using the voice-face matching view. Among these three datasets, DFDC and FakeAVCeleb are more challenging due to the inclusion of considerable videos and the plural forgery algorithms. Our VFD achieves around 85% AUCs, whereas the baselines even show accuracy below 50% (e.g., 49.23% and 43.65% of Meso-4). In contrast, the performance of baselines on DeepfakeTIMIT is more satisfying (i.e., most baselines are over 90% AUC). This relates to the fact that DeepfakeTIMIT contains standard faces without occlusion or rotation from a naive deepfake approach. Therefore, the models can easily capture minor visual and auditory artifacts. In what follows, we will analyze the experimental results in different groups.
### Table 2. Cross-Dataset Evaluation

| Performance | Xception | F³-Net | ViT | VFD |
|-------------|----------|--------|-----|-----|
| ACC         | 65.03    | 45.73  | 70.87 | **77.30** |
| AUC         | 73.20    | 58.49  | 74.64 | **82.53** |

The models are trained or fine-tuned on the DFDC dataset and tested on FakeAVCeleb.

#### 4.4.1 Comparison on Modality.
We can observe that some single-modality models perform unfavorably, with Capsule’s accuracy less than 60% and that of Meso-4 around 50%. One possible reason for this is that the single-modality approaches rely heavily on the artifacts extraction capability of the backbones. Hence, some strong models are prone to fail on the latest data with more realistic visual artifacts. In the multi-modal model zoo, the audio or emotion-based multi-modal methods, namely BA-TFD and Emotional Forensics, can significantly outperform the single-modality ones, demonstrating the promising potential of multi-modal cues. However, the ensembling models show both enhancements and decreases over their uni-modal counterparts (pink base color). We attributed this phenomenon to the fact that the ensembling model is designed to fuse the prediction of the two plain classifiers without considering the relations among modalities, which would lead to unstable performance [25]. Therefore, the multi-modal detection models and training strategies must be carefully designed to exploit the complementary properties of multiple modalities.

#### 4.4.2 Comparison on Auxiliary Data.
The models that employ auxiliary data, such as landmarks, perform comparably to traditional uni-modal methods. For example, DSP-FWA gains AUC over 75% and 99% on DFDC and DeepfakeTIMIT, which is on par with ViT, whereas Headpose only outperforms Meso-4. One possible reason is that the landmarks are representations of facial movements, whereas some powerful visual models may be sufficient to capture such features and thus achieve proximate performance.

#### 4.4.3 Comparison on Fine-Tuning Strategy.
Out of expectation, the fine-tuning on traditional approaches is harmful (blue base color). For instance, Xception decreases 4.99% on FakeAVCeleb, whereas F³-Net and ViT degrade close to 2.5% and 5.0% on DFDC, respectively. One dominant reason is that these three methods leverage only 1/4 of the data during fine-tuning, which trades certain performance degradation for faster training and efficient overheads. By contrast, the full fine-tuning methods are more competitive. For example, Face X-ray models significantly surpass Xception on FakeAVCeleb under two settings, and LipForensics outperforms other fine-tuning baselines on DFDC. Finally, our VFD performs the best over all the baselines (e.g., around 12% improvement over LipForensics). It is worth noting that VFD exploits limited data similar to Xception_F with only 1/4 of the dataset while achieving much satisfactory results, implying the robustness of VFD across different datasets from one side.

#### 4.4.4 Comparison on Cross-Dataset Validation.
We conducted a cross-dataset evaluation to further validate the effectiveness of VFD. Specifically, we selected three strong baselines—Xception, F³-Net, and ViT—which are trained on the DFDC dataset and directly evaluated on FakeAVCeleb. As for our VFD, we fine-tuned the model using the DFDC dataset. The results are presented in Table 2. From this table, VFD exhibits significantly higher cross-dataset capability compared to the other three baselines. A possible explanation is that VFD detects fake videos based on the audio-visual mismatch, a characteristic that is more prevalent across various deepfake videos compared to the forged artifacts relied upon by previous methods.

#### 4.4.5 Comparison on More Metrics.
In Table 3, we employ two new sets of metrics, namely Macro and Weighted, to evaluate both our VFD and the baselines. It can be observed that VFD
Table 3. Macro and Weighted Metrics on the DFDC Dataset

| Model   | Macro   |          |          | Weighted |          |          |
|---------|---------|----------|----------|----------|----------|----------|
|         | Precision | Recall   | F1       | Precision | Recall   | F1       |
| Xception| 70.24    | 56.53    | 55.57    | 72.88    | 75.15    | 69.24    |
| F3-Net  | 66.32    | 53.80    | 51.52    | 71.45    | 75.62    | 68.66    |
| VFD     | 76.97    | 68.93    | 71.40    | 81.22    | 82.55    | 81.04    |

Fig. 6. (a) Heatmaps produced by baselines and our method. The top two rows are from Xception and F3-Net, followed by the third row of heatmaps yielded by our VFD and the original face images. (b) Non-cherry-picked heatmaps of VFD on the FakeAVCeleb dataset.

demonstrates significant superiority, such as achieving nearly a 20% improvement in Macro-F1 compared to F3-Net. Additionally, the Weighted metrics consistently outperform the Macro ones. This is attributed to the data imbalance between real and fake videos in the DFDC dataset. The Macro metrics highlight the baselines' tendency to misclassify real videos as fake, whereas our method significantly mitigates this issue.

4.5 Qualitative Results

4.5.1 Heatmap. To qualitatively compare our method with baselines, we demonstrated some generated heatmaps. One can observe that in Figure 6(a), compared with Xception and F3-Net, VFD focuses on the whole face of targets, indicating that VFD recognizes fake videos based on global identity information rather than specific regional artifacts. Moreover, Figure 6(b) illustrates more VFD heatmaps in FakeAVCeleb. It can be seen that VFD yields stable attention regions on different datasets, which further proves that VFD is more generalizable and will not fail due to the migration of datasets or the update of deepfake algorithms.

4.5.2 Similarity Comparison. We computed the voice-face similarity from VFD via Equation (1) and show the results in Figure 7. This demonstrates that the real and fake videos are evidently split. In addition, the split line of real and fake videos is around −0.1 for both datasets, implying that VFD does not calculate the distance via specific features of the dataset but from the general matching view. Otherwise, different cutoff values will be learned by our method.

Figure 8 illustrates the voice-face similarity on the LQ and HQ subsets of DeepfakeTIMIT. We first show the performance via fine-tuning and testing on LQ and HQ in Figure 8(a) and (b), respectively. A clear cutoff line around -0.1 can be observed, as in DFDC and FakeAVCeleb.
that the two subsets are synthesized from the same real videos, we swapped fine-tuning sets to verify the reliability of VFD. Specifically, we conducted a new experiment with fine-tuning on LQ and then testing on HQ directly, or vice versa. The results are displayed in Figure 8(c) and (d). It can be seen that the model still discriminates between real and fake videos, proving that the image quality does not significantly affect the effectiveness of the VFD, the dominant reason being that our model focuses on the identity features of the whole face rather than the limited facial manipulation trajectories that are easily covered by compressing or blurring. Another interesting point is that the real parts share a similar distribution when processed via the models fine-tuning on a specific dataset (as in Figure 8(a) and (c)). We attribute this phenomenon to the real videos in both LQ and HQ testing sets sampled from VidTIMIT. In other words, the real instances in the LQ and HQ sets are identical, thus bringing the same distribution.

4.5.3 Detection Failure Cases. We also illustrate the typical failure cases of VFD in Figure 9. There are three notable causes leading to the detection failure: illumination, facial perspectives, and attributes. From Figure 9(a), one can see that the faces are mixed with backgrounds due to inadequate illumination, which undermines the model’s capability. Figure 9(b) shows some extreme facial angles (e.g., 90 degrees of head-turning), making the facial features difficult to extract. Moreover, some face attribute editing videos [23] are displayed in Figure 9(c). The images with a blue border are the faces from real videos in each rectangular box, whereas red ones are extracted from the corresponding fake videos. One can see that the modification of specific facial attributes, such as style of the glasses or upper lip beard, will lead to invalid identification as identities are frozen in this scenario.
Fig. 9. Some detection failure cases. (a) Blurred faces due to the scarcity of illumination. (b) Extreme facial angles, making the facial features under-represented. (c) The modification of facial attributes (e.g., glasses or beard), wherein identities are unchanged.

Fig. 10. Hyperparameter study. (a) The AUC and ACC results with respect to different numbers of real videos in the deepfake dataset. (b) The influence of the negative samples.

4.6 Hyperparameter Study

We studied VFD with extensive parameter settings. We first illustrate the AUC and ACC results with different fine-tuning dataset sizes on DFDC in Figure 10(a). As we can see, when fine-tuning on a limited subset of fine-tuning dataset (e.g., 1,000 or 2,000 real samples), VFD already achieves an AUC of 77%, which outperforms most baselines. Furthermore, even if the dataset degrades to 100, VFD’s performance (65.87% on AUC) is still comparable to some baselines such as MesoInception-4 (60.56%) and VA-LogReg (66.20%), proving that VFD can achieve an acceptable identification ability in a few-shot way and quickly adapt to newly forgery algorithms. VFD reaches its best performance when fine-tuning instances reach 5,000. Compared to traditional approaches where the entire DFDC training set is used (i.e., more than 20,000 videos [52, 53, 66]), VFD requires fewer samples to achieve better performance, making our method distinguished from these baselines.

In addition, we studied the influence of different number of negatives (i.e., the Q in Equation (8)) and show the results in Figure 10(b). The number 0 denotes that no fake pairs are leveraged for fine-tuning. We can observe that the performance is significantly boosted with more negatives.
Table 4. Performance of VFD under Different Ablation Settings on DFDC and FakeAVCeleb

| Model        | Modality  | Modulator | InfoNCE | RFC | DFDC  | FakeAVCeleb |
|--------------|-----------|-----------|---------|-----|-------|--------------|
|              | Visual    | Audio     |         |     | AUC   | ACC          | AUC   | ACC          |
| Visual-only  | ✓         |           |         |     | 65.04 | 71.16        | 70.05 | 74.30        |
| Visual-auditory | ✓         | ✓         |         |     | 52.45 | 54.52        | 64.16 | 69.12        |
| Modulator    | ✓         | ✓         | ✓       |     | 56.93 | 59.54        | 65.01 | 69.22        |
| InfoNCE-based | ✓         | ✓         | ✓       | ✓   | 59.64 | 60.15        | 80.15 | 85.40        |
| VFD          | ✓         | ✓         | ✓       | ✓   | 80.96 | 85.13        | 81.52 | 86.11        |

Table 5. GFLOPs of the Models in Different Stages

| Stage | Fine-Tuning | Testing |
|-------|-------------|---------|
|       | Voice+  | Face+  | Voice−  | Face−  | VFD | Xception | F3-Net |
| GFLOPs | 2.83   | 4.36   | 2.83×Q   | 4.36×Q   | 7.19 | 8.39     | 17.25 |

In the fine-tuning stage, the number of negative samples Q is correlated to time consumption. During the testing, the absence of negative examples results in a significant reduction in time requirements.

4.7 Ablation Study

4.7.1 Module Evaluation. To study the effectiveness of different modules in VFD, we explored the performance of the following variants. First, Visual-only is the visual-modality only version of VFD, wherein a cross-entropy loss is employed to classify the real and fake videos, namely the ViT_F in Table 1. Second, Visual-auditory utilizes both voice and face extractors, while the two modulators in Figure 5 are removed, and a linear classifier is applied to bind multi-modal features instead of contrastive loss. Third, Modulator adds the voice and face modulators. Fourth, the InfoNCE-based model integrates InfoNCE into the pre-training phase while replacing the proposed RFC with InfoNCE in the fine-tuning stage.

As can be seen from Table 4, all these four variants will jeopardize the performance to some extent. Among them, the Visual-auditory and Modulator models perform inferior to the uni-modal counterpart Visual-only in most cases, proving that detecting deepfakes via multi-modal cues require elaborated approaches to model the homogeneity. Furthermore, Modulator surpasses Visual-auditory due to the introduction of modality modulators. Moreover, it can be observed that applying InfoNCE to pre-training will improve the model performance, validating the effectiveness of our pre-training strategy. Meanwhile, the InfoNCE-based model performs comparably with RFC on FakeAVCeleb but fails severely on DFDC during fine-tuning. We attributed this result to the fact that the FakeAVCeleb is built with relatively detailed annotation (i.e., both voice and face manipulations are labeled), which enables InfoNCE to obtain adequate features. Nevertheless, when a similar strategy is applied to coarsely annotated datasets (e.g., DFDC), some crucial properties are overlooked, as introduced in Section 3.3. In a nutshell, when combining all the devised modules, our method can achieve the best result.

4.7.2 Time Consumption of VFD. We conducted an analysis on the time costs associated with various modules in VFD. First, we report the GFLOPs (giga floating-point operations) during the fine-tuning and testing stages in Table 5. During the fine-tuning stage, the model simultaneously processes both positive and negative videos to derive \( f_v^+, f_v^-, f_f^+, \) and \( f_f^- \), as specified by Equation (7). The time required in this stage depends on the number of negative videos, represented by the
variable Q. In the testing stage, the selection of negative samples is no longer necessary, leading to a substantial reduction in time costs. The time requirement of VFD in processing a video is 7.19 GFLOPs, which is noticeably lower than that of Xception (8.39 GFLOPs) and F³-Net (17.25 GFLOPs). To further illustrate the influence of the number of negative samples in fine-tuning stage, we present the video throughput (videos/s) in Figure 11. The results demonstrate a significant decline in the video throughput of VFD as the number of negative samples increases. Consequently, to strike an optimal balance between efficiency and performance, meticulous consideration must be given to the selection of the value for Q.

4.7.3 Strategy Evaluation. To further demonstrate the necessity of the pre-training then fine-tuning paradigm, we studied the following variants: (1) VFD w/o pre-training replaces the pre-training strategy with training from scratch, and (2) VFD w/o fine-tuning applies the pre-trained model to detect forgery directly. From Table 6, one can observe that the pre-training strategy contributes more than 10% and 5% AUC improvements on DFDC and FakeAVCeleb, respectively, wherein the homogeneity between voices and faces are accordingly modeled. Meanwhile, the removal of fine-tuning is more detrimental to the performance (e.g., the AUC on DFDC drops by 30%), indicating that fine-tuning is essential to narrow the gap between generic and deepfake datasets.

4.7.4 Feature Extractor Evaluation. We investigated the impact of feature extractors structures. Specifically, we modified the transformer-based models in Section 3.2.2 into VGG-based [59] ones. We then evaluated their performance in two dimensions: pre-training and fine-tuning. The pre-training reports the capability of the each feature extractor for identifying the faces and voices on the generic dataset, and the accuracy on matching voice-face pairs from the same identities. In addition, fine-tuning focuses on the performance on the deepfake dataset, and we thus compare the AUC and ACC results on DFDC dataset.

From Table 7, one can observe that the transformer-based model outperforms the VGG-based model in all metrics in this task. We attribute this to the transformer’s superiority to model the...
global features of faces and voices, which will be more suitable for identity-based matching. In addition, both types of extractors show superior performance in face recognition (i.e., over 98% and 94% accuracy in classifying the faces, respectively). The identification of voice is the bottleneck that limits the model performance, implying that accurate classification for voice is both challenging and cutting edge. Moreover, VFD exhibits a matching accuracy close to 97% on the generic dataset. A possible explanation is that the multi-modal inputs are mutually supportive via contrastive learning such that the errors of the voice feature extractor are compensated to some extent. Meanwhile, the decrease after migration to the deepfake dataset suggests that our matching view has considerable potential.

5 CONCLUSION AND DISCUSSION

Detecting forgery videos in deepfake is challenging due to the continuous progress from deepfake techniques. In this work, we empirically recognized the severe mismatch of the voices and faces in deepfake videos, based on which we proposed to tackle this task from an unexplored homogeneity modeling perspective. In particular, our proposed method followed a pre-training then fine-tuning pipeline, wherein the voices and faces from a single identity were learned to closely match with a novel RFC loss. The experimental and visualization results demonstrated the effectiveness of the proposed method against existing elaborately designed competitors.

Despite its effectiveness on existing deepfake datasets, VFD shows a certain limitation in the unusual case where both voices and faces are maliciously edited. Given the fact that our method addresses deepfake detection from either face swapping or voice editing, simultaneously performing these two may lead to the matching collapse. However, to the best of our knowledge, no such high-quality datasets, or associated approaches, have been studied in the literature so far. Yet, this challenging setting remains promising and demands extensive exploration in the future.

REFERENCES

[1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. 2018. MesoNet: A compact facial video forgery detection network. In Proceedings of the International Workshop on Information Forensics and Security. IEEE, Los Alamitos, CA, 1–7.
[2] Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter N. Belhumeur, and Shree K. Nayar. 2008. Face swapping: Automatically replacing faces in photographs. ACM Transactions on Graphics 27, 3 (2008), 1–8.
[3] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. 2004. Exchanging faces in images. Computer Graphics Forum 23, 3 (2004), 669–676.
[4] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. 2022. Do you really mean that? Content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. arXiv:2204.06228 (2022).
[5] Harry Cheng, Yangyang Guo, Liqiang Nie, Zhiyong Cheng, and Mohan S. Kankanahalli. 2023. Sample less, learn more: Efficient action recognition via frame feature restoration. arXiv:2307.14866 (2023).
[6] Harry Cheng, Yangyang Guo, Jianhua Yin, Haonan Chen, Jiafang Wang, and Liqiang Nie. 2021. Audio-driven talking video frame restoration. IEEE Transactions on Multimedia. Early access, October 7, 2021.
[7] Kai Cheng, Xin Liu, Yu-ming Cheung, Rui Wang, Xing Xu, and Bineng Zhong. 2020. Hearing like seeing: Improving voice-face interactions and associations via adversarial deep semantic matching network. In Proceedings of the ACM International Conference on Multimedia. ACM, New York, NY, 448–455.
[8] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep speaker recognition. In *Proceedings of the Conference of the International Speech Communication Association*. 1086–1090.

[9] Brian Dolhansky, Joanna Bitton, Ben Pliaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton-Ferrero. 2020. The DeepFake Detection Challenge (DFDC) dataset. arXiv:2006.07397 (2020).

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissennbom, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*. 1–12.

[11] Shiming Ge, Fanzhao Lin, Chenyu Li, Daichi Zhang, Weiping Wang, and Dan Zeng. 2022. Deepfake video detection via predictive representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 2s (2022), 1–21.

[12] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM* 63, 11 (2020), 139–144.

[13] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022), e211003119.

[14] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Alberto Del Bimbo. 2021. AdVAQA: Overcoming language priors with adapted margin cosine loss. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 708–714.

[15] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. 2022. Loss re-scaling VQA: Revisiting the language prior problem from a class-imbalance view. *IEEE Transactions on Image Processing* 31 (2022), 227–238.

[16] Zhiqing Guo, Gaobo Yang, Jiyou Chen, and Xingming Sun. 2021. Fake face detection via adaptive manipulation traces extraction network. *Computer Vision and Image Understanding* 204 (2021), 103170.

[17] Zhiqing Guo, Gaobo Yang, Jiyou Chen, and Xingming Sun. 2023. Exposing deepfake face forgeries with guided residuals. *IEEE Transactions on Multimedia*. Early access, January 18, 2023.

[18] Zhiqing Guo, Gaobo Yang, Dewang Wang, and Dengyong Zhang. 2023. A data augmentation framework by mining structured features for fake face image detection. *Computer Vision and Image Understanding* 226 (2023), 103587.

[19] Zhiqing Guo, Gaobo Yang, Dengyong Zhang, and Ming Xia. 2023. Rethinking gradient operator for exposing AI-enabled face forgeries. *Expert Systems with Applications* 215 (2023), 119361.

[20] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 2021. Lips don’t lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 5039–5049.

[21] Shota Horiguchi, Naoyuki Kanda, and Kenji Nagamatsu. 2018. Face-voice matching using cross-modal embeddings. In *Proceedings of the ACM International Conference on Multimedia*. ACM, New York, NY, 1011–1019.

[22] Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson. 2003. ‘Putting the face to the voice’: Matching identity across modality. *Current Biology* 13, 19 (2003), 1709–1714.

[23] Tero Karras, Samuli Laine, and Timo Aila. 2018. *A style-based generator architecture for generative adversarial networks*. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 12 (2021), 4217–4228.

[24] Ira Kemelmacher-Shlizerman. 2016. Transfiguring portraits. *ACM Transactions on Graphics* 35, 4 (2016), Article 94, 8 pages.

[25] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon S. Woo. 2021. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In *Proceedings of the ACM Workshop on Synthetic Multimedia*. ACM, New York, NY, 7–15.

[26] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. 2021. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS ’21): Track on Datasets and Benchmarks*. 1–15.

[27] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. 2018. Deep video portraits. *ACM Transactions on Graphics* 37, 4 (2018), Article 163, 14 pages.

[28] Pavel Korshunov and Sébastien Marcel. 2018. DeepFakes: A new threat to face recognition? Assessment and detection. *CoRR abs/1812.08685* (2018).

[29] Pavel Korshunov and Sébastien Marcel. 2018. Speaker inconsistency detection in tampered video. In *Proceedings of the European Signal Processing Conference*. IEEE, Los Alamitos, CA, 2375–2379.

[30] Mohammad Rami Koujan, Michail Christos Doukas, Anastasios Roussos, and Stefanos Zafeiriou. 2020. Head2Head: Video-based neural head synthesis. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*. IEEE, Los Alamitos, CA, 16–23.

[31] John K. Lewis, Imad Eddine Toubal, Helen Chen, Vishal Sandesera, Michael Lomnitz, Zigfried Hampel-Arias, Prasad Calyam, and Kannappan Palaniappan. 2020. Deepfake video detection based on spatial, spectral, and temporal
inconsistencies using multimodal deep learning. In Proceedings of the Applied Imagery Pattern Recognition Workshop. IEEE, Los Alamitos, CA, 1–9.

[32] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2020. Advancing high fidelity identity swapping for forgery detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition. IEEE, Los Alamitos, CA, 5073–5082.

[33] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face x-ray for more general face forgery detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition. IEEE, Los Alamitos, CA, 5000–5009.

[34] Yidong Li, Wenhua Liu, Yi Jin, and Yuanzhouhan Cao. 2021. SPGAN: Face forgery using spoofing generative adversarial networks. ACM Transactions on Multimedia Computing, Communications, and Applications 17, 1s (2021), 1–20.

[35] Yuezun Li and Siwei Lyu. 2019. Exposing DeepFake videos by detecting face warping artifacts. In Proceedings of the Conference on Computer Vision and Pattern Recognition Workshops. IEEE, Los Alamitos, CA, 46–52.

[36] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 2020. Celeb-DF: A large-scale challenging dataset for DeepFake Forensics. In Proceedings of the Conference on Computer Vision and Pattern Recognition. IEEE, Los Alamitos, CA, 3204–3213.

[37] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. 2021. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In Proceedings of the Conference on Computer Vision and Pattern Recognition. IEEE, Los Alamitos, CA, 772–781.

[38] Xiaolong Liu, Yang Yu, Xiaolong Li, Yao Zhao, and Guodong Guo. 2023. TCSD: Triple complementary streams detector for comprehensive deepfake detection. ACM Transactions on Multimedia Computing, Communications, and Applications 19, 6 (2023), Article 213, 22 pages.

[39] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations. 1–8.

[40] Bahar Uddin Mahmud, Guan Yue Hong, and Bernard Fong. 2023. A study of human-AI symbiosis for creative work: Recent developments and future directions in deep learning. ACM Transactions on Multimedia Computing, Communications, and Applications 20, 2 (2023), Article 47, 21 pages.

[41] Iacopo Masi, Aditya Killekar, Royston Marijan Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. 2020. Two-branch recurrent network for isolating deepfakes in videos. In Proceedings of the European Conference on Computer Vision. 667–684.

[42] Falko Matern, Christian Riess, and Marc Stamminger. 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In Proceedings of the Winter Applications of Computer Vision Workshops. IEEE, Los Alamitos, CA, 83–92.

[43] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. 2020. Emotions don’t lie: An audio-visual deepfake detection method using affective cues. In Proceedings of the ACM International Conference on Multimedia. ACM, New York, NY, 2823–2832.

[44] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Seeing voices and hearing faces: Cross-modal biometric matching. In Proceedings of the Conference on Computer Vision and Pattern Recognition. IEEE, Los Alamitos, CA, 8427–8436.

[45] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: A large-scale speaker identification dataset. In Proceedings of the Conference of the International Speech Communication Association. 2616–2620.

[46] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. 2019. Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. IEEE, Los Alamitos, CA, 2307–2311.

[47] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. FSGAN: Subject agnostic face swapping and reenactment. In Proceedings of the International Conference on Computer Vision. IEEE, Los Alamitos, CA, 7183–7192.

[48] Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. FSGAN: Subject agnostic face swapping and reenactment. In Proceedings of the International Conference on Computer Vision. IEEE, Los Alamitos, CA, 7183–7192.

[49] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T. Freeman, Michael Rubinstein, and Wojciech Matusik. 2019. Speech2Face: Learning the face behind a voice. In Proceedings of the Conference on Computer Vision and Pattern Recognition. IEEE, Los Alamitos, CA, 7539–7548.

[50] Pietro Pala and Stefano Berrett. 2019. Reconstructing 3D face models by incremental aggregation and refinement of depth frames. ACM Transactions on Multimedia Computing, Communications, and Applications 15 (2019), 1–24.

[51] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In Proceedings of the ACM International Conference on Multimedia. ACM, New York, NY, 484–492.

[52] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Proceedings of the European Conference on Computer Vision. 86–103.
[53] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Face-Forensics++: Learning to detect manipulated facial images. In *Proceedings of the International Conference on Computer Vision*. IEEE, Los Alamitos, CA, 1–11.

[54] Conrad Sanderson and Brian C. Lovell. 2009. Multi-region probabilistic histograms for robust and scalable identity inference. In *Proceedings of the International Conference on Biometrics*. 199–208.

[55] Huiru Shao, Jing Li, Jia Zhang, Hui Yu, and Jiande Sun. 2020. Eye-based recognition for user identification on mobile devices. *ACM Transactions on Multimedia Computing, Communications, and Applications* 16, 4 (2020), 1–19.

[56] Nitin Arvind Shelke and Singara Singh Kasana. 2021. A comprehensive survey on passive techniques for digital video forgery detection. *Multimedia Tools and Applications* 80, 4 (2021), 6247–6310.

[57] Nitin Arvind Shelke and Singara Singh Kasana. 2022. Multiple forgeries identification in digital video based on correlation consistency between entropy coded frames. *Multimedia Systems* 28, 1 (2022), 267–280.

[58] Nitin Arvind Shelke and Singara Singh Kasana. 2022. Multiple forgery detection and localization technique for digital video using PCT and NBAP. *Multimedia Tools and Applications* 81, 16 (2022), 22731–22759.

[59] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*. 1–10.

[60] Mingxing Tan and Quoc V. Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning*. 6105–6114.

[61] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR abs/1807.03748* (2018).

[62] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, 86 (2008), 2579–2605.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems*. 5998–6008.

[64] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021. HiFiFace: 3D shape and semantic prior guided high fidelity face swapping. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 1136–1142.

[65] Peisong Wen, Qianqian Xu, Yangbangyan Jiang, Zhiyong Yang, Yuan He, and Qingming Huang. 2021. Seeking the shape of sound: An adaptive framework for learning voice-face association. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 16347–16356.

[66] Xi Wu, Zhen Xie, YuTao Gao, and Yu Xiao. 2020. SSTNet: Detecting manipulated faces through spatial, steganalysis and temporal features. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE, Los Alamitos, CA, 2952–2956.

[67] Han Xue, Jun Ling, Anni Tang, Li Song, Rong Xie, and Wenjun Zhang. 2023. High-fidelity face reenactment via identity-matched correspondence learning. *ACM Transactions on Multimedia Computing, Communications, and Applications* 19, 3 (2023), Article 122, 23 pages.

[68] Chenggang Yan, Lixuan Meng, Liang Li, Jiehua Zhang, Zhan Wang, Jian Yin, Jiayong Zhang, Yaoqi Sun, and Bolun Zheng. 2022. Age-invariant face recognition by multi-feature fusion and decomposition with self-attention. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 1s (2022), 1–18.

[69] Lan Yan, Wenbo Zheng, Chao Gou, and Fei-Yue Wang. 2022. IFGAN: Identity-preservation generative adversarial network for unsupervised photo-to-caricature translation. *Knowledge-Based Systems* 241 (2022), 108223.

[70] Xin Yang, Yuezu Li, and Siwei Lyu. 2019. Exposing deep fakes using inconsistent head poses. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE, Los Alamitos, CA, 8261–8265.

[71] Yongxiang Yao, Yongjun Zhang, Yi Wan, Xinyi Liu, Xiaohu Yan, and Jiayuan Li. 2022. Multi-modal remote sensing image matching considering co-occurrence filter. *IEEE Transactions on Image Processing* 31 (2022), 2584–2597.

[72] Yang Yu, Rongrong Ni, Wenjie Li, and Yao Zhao. 2022. Detection of AI-manipulated fake faces via mining generalized features. *ACM Transactions on Multimedia Computing, Communications, and Applications* 18, 4 (2022), Article 94, 23 pages.

[73] Hangxing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. 2021. Multi-attentional deepfake detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*. IEEE, Los Alamitos, CA, 2185–2194.

[74] Yipin Zhou and Ser-Nam Lim. 2021. Joint audio-visual deepfake detection. In *Proceedings of the International Conference on Computer Vision*. IEEE, Los Alamitos, CA, 14800–14809. 

Received 27 January 2023; revised 18 May 2023; accepted 18 September 2023