

Standing on the Shoulders of Giants: Reprogramming Visual-Language Model for General Deepfake Detection

Kaiqing Lin^{1*}, Yuzhen Lin^{1*}, Weixiang Li¹, Taiping Yao², Bin Li^{1†}

¹Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, and SZU-AFS Joint Innovation Center for AI Technology, Shenzhen University, Shenzhen 518060, China

²Tencent YouTu Lab

linkaiqing2021@email.szu.edu.cn, libin@szu.edu.cn

Abstract

The proliferation of deepfake faces poses huge potential negative impacts on our daily lives. Despite substantial advancements in deepfake detection over these years, the generalizability of existing methods against forgeries from unseen datasets or created by emerging generative models remains constrained. In this paper, inspired by the zero-shot advantages of Vision-Language Models (VLMs), we propose a novel approach that repurposes a well-trained VLM for general deepfake detection. Motivated by the model reprogramming paradigm that manipulates the model prediction via input perturbations, our method can reprogram a pre-trained VLM model (e.g., CLIP) solely based on manipulating its input without tuning the inner parameters. First, learnable visual perturbations are used to refine feature extraction for deepfake detection. Then, we exploit information of face embedding to create sample-level adaptive text prompts, improving the performance. Extensive experiments on several popular benchmark datasets demonstrate that (1) the cross-dataset and cross-manipulation performances of deepfake detection can be significantly and consistently improved (e.g., over 88% AUC in cross-dataset setting from FF++ to WildDeepfake); (2) the superior performances are achieved with fewer trainable parameters, making it a promising approach for real-world applications.

Introduction

Deepfake refers to a series of deep learning-based facial forgery techniques (Li et al. 2020a; Xu et al. 2022; Chen et al. 2024a) that can swap or reenact the face of one person in a video to another. In recent years, deepfake videos (a.k.a, deepfakes) have gained substantial attention due to their potential by creating and spreading false information. Thus, detecting deepfakes has emerged as a crucial research topic to reduce such security risks.

Existing methods treat deepfake detection as a binary classification problem and predominantly utilize CNNs (e.g., Xception or EfficientNet) as the backbones of classifier. In addition, some works (Qian et al. 2020; Li et al. 2020b; Zhao et al. 2021; Yan et al. 2023b, 2024b) propose to introduce auxiliary clues, including modalities (e.g.,

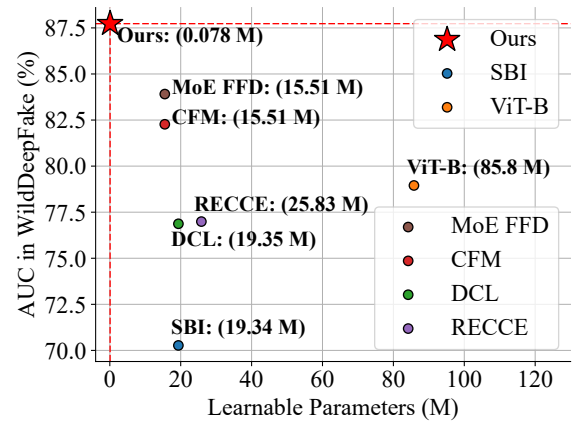


Figure 1: Comparison between our method and open-source deepfake detection models on the WildDeepfake dataset (trained on FF++). Our method with the fewest learnable parameters while achieves the best performance.

frequency) or supervision (e.g., forgery masks) information for learning subtle forgery artifacts. Despite these advancements, evaluating forgeries from unseen datasets and synthesized by unseen methods beyond the training data still poses a significant challenge for practical deepfake detection.

Vision Language Models (VLMs), such as CLIP (Contrastive Language-Image Pre-training) (Radford et al. 2021), demonstrate robust zero-shot and few-shot generalization capabilities across diverse downstream tasks. To improve deepfake detector generalization, VLFFD (Sun et al. 2023) fully fine-tunes a CLIP-based model, suggesting the effectiveness of CLIP. However, excessive adjustment of parameters in pre-trained models can disrupt the great pre-trained knowledge, potentially leading to suboptimal performance. Thus, developing a cost-efficient approach is demanding.

Due to the inherent vulnerability of deep networks to adversarial attacks, a novel concept called model reprogramming (Elsayed, Goodfellow, and Sohl-Dickstein 2019; Chen 2024) has been proposed to re-purpose a well-trained model trained in a source domain to perform a target-domain task, just through learning universal input perturbations without modifying the source-domain model parameters. In other

*These authors contributed equally.

†Corresponding author.

words, an additive offset applied to a deep network’s input would be sufficient to adapt the network to a new task without the need of re-training or fine-tuning its inner parameters. The reprogramming paradigm requires significantly fewer learnable parameters compared to fine-tuning numerous parameters within the model. Inspired by the above advancements, we advocate that reprogramming a well-trained foundation model to identify deepfakes is a promising way to improve generalization and training efficiency.

In this paper, we propose RepDFD, a novel method to reprogram a pre-trained CLIP model for effective and general deepfake detection. RepDFD solely learns task-specific visual perturbations (a.k.a, visual prompts) in pixel spaces for the deepfake detection task while keeping the entire CLIP model frozen. Specifically, we introduce *Input Transformation* to merge the image and the visual prompt and then feed it into the image encoder of CLIP. Furthermore, we propose *Face2Text Prompts* to generate text prompts merging information of face embedding, and then feed them into the text encoder of CLIP to guide the optimization of the visual prompts. This straightforward method enables the CLIP model to effectively detect deepfakes. Moreover, since the internal parameters are not trained, the foundation CLIP model can be reused for the other vision tasks. Extensive experiments on several popular deepfake benchmarks demonstrate that (1) the cross-dataset and cross-manipulation performances can be significantly and consistently enhanced by equipping RepDFD for a pre-trained CLIP model; and (2) the superior performances are achieved with fewer trainable parameters (see in Fig. 1), making it a promising approach for real-world applications.

Briefly, the main contributions of this work can be summarized as follows:

- This is the first work to explore model reprogramming paradigm for deepfake detection tasks.
- We have proposed RepDFD to reprogram a pre-trained CLIP model by solely processing its image and text inputs without tuning the inner parameters. Thus, RepDFD can be seamlessly adapted to other foundation models.
- We have conducted extensive experiments on several benchmark datasets, and have demonstrated that RepDFD is general and efficiency for deepfake detection.

Related Work

Deepfake Detection

The past five years have witnessed a wide variety of methods proposed for defending against the malicious usage of deepfakes. Currently, the majority of deepfake detection methods are based on deep learning, leveraging generic CNNs (e.g., Xception, EfficientNet) as the backbones of the classifier. Furthermore, several works (Qian et al. 2020; Zhao et al. 2021) utilize frequency information or localize the forged regions to improve the performance of detectors. Nevertheless, the generalization challenge still hinders the application of deepfake detectors in real-world scenarios. To address such issue, several works (Li et al. 2020b; Shiohara and Yamasaki 2022; Nguyen et al. 2024; Lin et al. 2024b) introduce the augmented deepfake data, where two different faces

are blended, during the training. However, all of the above methods typically involve retraining backbone networks. In general, employing more powerful backbone networks (e.g., replacing CNNs with ViTs or VLMs) produces better performances but at the cost of increased computational costs in the training stage.

CLIP Model

CLIP is a vision-language model (Radford et al. 2021), that is able to perform flexible zero-shot transfer to unseen classes using text prompts. Since CLIP is pre-trained to predict whether an image matches a textual description, it naturally fits zero-shot recognition. This is achieved by comparing image features with the classification weights synthesized by the text encoder, which takes as input textual descriptions specifying classes of interest. The CLIP model contains an image-encoder $E_I(\cdot)$ and a text-encoder $E_T(\cdot)$ such that the cosine similarity between the features $E_I(x_k)$ and $E_T(t_k)$ are maximized with respect to each pair k . Compared with the traditional classifier learning approach where closed-set visual concepts are learned from random vectors, vision-language pre-training allows open-set visual concepts to be explored through a high-capacity text encoder, leading to a broader semantic space and in turn making the learned representations more transferable to various vision tasks. Several works (Sha et al. 2023; Ojha, Li, and Lee 2023; Cozzolino et al. 2024) has explore that using CLIP model for detecting generated images by GANs and Diffusion Models. VLFFD (Visual-Linguistic Face Forgery Detection) (Sun et al. 2023) proposes to fully fine-tune CLIP with fine-grained text prompts for deepfake detection. In this work, we propose a cost-efficient method to adapt a pre-trained CLIP model for general deepfake detection.

Model Reprogramming

Parameter Efficient Fine-Tuning (PEFT) methods (Tsao et al. 2024), which significantly reduce computational and storage overheads, have garnered increasing attention. By fine-tuning only a limited number of additional parameters, PEFT can adapt a large pre-trained model to achieve outstanding performance on targeted tasks. One of the PEFT methods, model reprogramming (Chen 2024), which incorporates prompts into inputs, offers an effective framework for repurposing models for various task-specific applications. The framework draws significant inspiration from adversarial reprogramming, which was first introduced by Elsayed et al (Elsayed, Goodfellow, and Sohl-Dickstein 2019). Model reprogramming aims to re-use and re-align the data representation, from an existing model, for a separate task without fundamental changes to the model’s inner parameters. This paradigm re-purposes existing knowledge by strategically transforming inputs and outputs, bypassing extensive inner model parameter fine-tuning. Reprogramming techniques have been widely applied in various tasks in the past few years (Wang et al. 2022; Cai et al. 2024). Visual prompting (VP) (Bahng et al. 2022) introduces a universal perturbation directly into the input data to facilitate task-specific fine-tuning while keeping the pre-trained model intact. Although these PEFT methods have developed in other

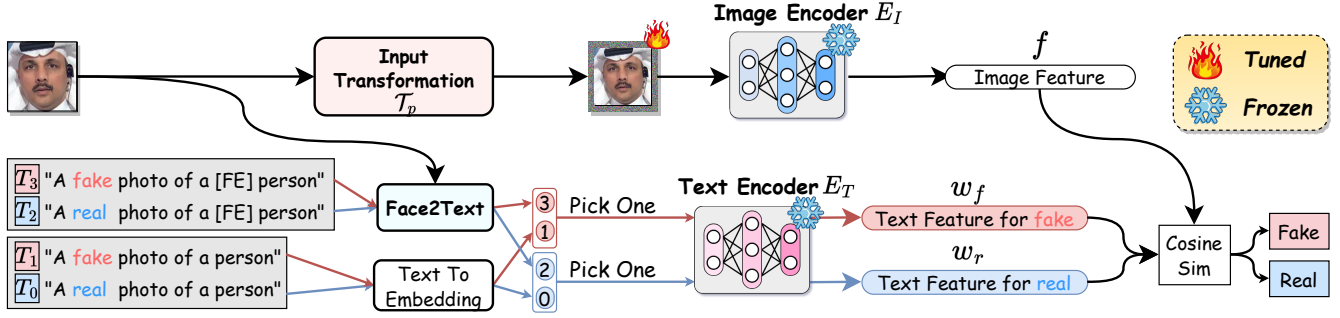


Figure 2: Overall framework of our proposed method. The core idea involves optimizing an universal visual prompt on a frozen CLIP model and generating sample-level text prompts (where the placeholder [FE] is replaced by a face embedding), aiming to adapt the model for the deepfake detection task.

research fields, the application of PEFT methods to face forgery detection remains largely unexplored. In this study, we introduce a novel approach to reprogramming a pre-trained CLIP model for deepfake detection using only a few number of learnable parameters, designed to leverage the great generalization abilities of the pre-trained Vision-Language Model (VLM).

Methodology

Overview

This paper proposes RepDFD, which reprograms the pre-trained vision language model CLIPs for deepfake detection without altering internal parameters. Given a well-trained CLIP model (including the image encoder E_I and the text encoder E_T), it introduces visual and textual prompts on the inputs for adapting the frozen E_I and E_T to identify deepfakes. To achieve this, we propose two modules, *Input Transformation* and *Face2Text Prompts*, which process the image and text before inputting to E_I and E_T , respectively. In this way, we harness the power of the model reprogramming to guide the CLIP model to focus on the deepfake detection task while preserving its inner parameters and pre-trained knowledge. The overall framework is depicted in Fig. 2. In what follows, we elaborate on the details of *Input Transformation*, *Face2Text Prompts* and the optimization pipeline.

Input Transformation

RepDFD refines the visual features from the pre-trained CLIP model for deepfake detection tasks by introducing learnable visual prompts containing perturbations that improve the visual encoder’s features. A similar scheme like VP (Bahng et al. 2022) and AutoVP (Tsao et al. 2024) is adopted to initialize the visual prompt, placing it around images for input transformations. Input images are first resized smaller and subsequently incorporated with the visual prompts, aiming to match the input size of the pre-trained image encoder E_I . Fig. 3 demonstrates the details of Input Transformation. Given an original image \mathbf{X} , the input transformation aims to merge the universal visual prompt δ to \mathbf{X} . Formally, the process of input transformations can be de-

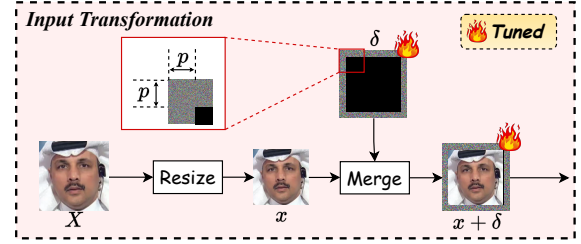


Figure 3: Illustration of Input Transformation

scribed as

$$\mathcal{T}_p(\mathbf{X}, \delta) = \text{Resize}_p(\mathbf{X}) + \delta, \quad (1)$$

where p is the width of the visual prompt. Finally, the original image \mathbf{X} of size $H \times W$ is resized to $H' \times W' = (H - 2p) \times (W - 2p)$, ensuring the transformed image does not overlap with δ . The number of trainable parameters in this method is model-agnostic, which can be computed by

$$\#Para = 3(HW - H'W') = 6p(H + W) - 12p^2. \quad (2)$$

This expression indicates that δ is applied to the RGB channels of the image. Therefore, for training or inference, $\mathcal{T}_p(\mathbf{X}, \delta)$ is fed into image encoder E_I to get the image feature f , i.e.,

$$f = E_I(\mathcal{T}_p(\mathbf{X}, \delta)). \quad (3)$$

Face2Text Prompts

To map model outputs to the target label, the CLIP model utilizes text prompts to align image features in a shared latent space, enabling classification by matching the closest text description to the image features. Therefore, a particular text prompt must be created for every class by injecting the information of the category. For instance, one of the official unified templates of text prompts is “A photo of a [cls]”, where [cls] is a placeholder of class label. However, in the context of deepfake detection, the binary class labels “real face” and “fake face” (or synonymous terms) lack specificity compared to labels such as “cat” or “dog”, making them difficult to accurately identify using existing text templates. To

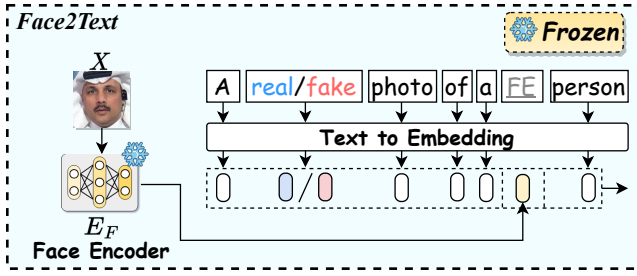


Figure 4: Illustration of Face2Text Prompts.

address this issue, VLFFD (Sun et al. 2023) introduces fine-grained text descriptions manually to fine-tune both $E_I(\cdot)$ and $E_T(\cdot)$ in the CLIP model. However, designing a hand-crafted text prompt that comprehensively describes forgery clues is difficult and labor-intensive.

In this work, we aim to design a simple but effective text prompts generation method to fully harness the generalizable capabilities of the CLIP model for deepfake detection. The success of personalized image generation methods (Chen et al. 2024b; Papantoniou et al. 2024) has shown that introducing pre-trained image embedding as textual embeddings is beneficial for maintaining the consistency of generated subjects. Considering that deepfakes involve the manipulation of faces, the inconsistency of facial representations between real and fake faces is a reliable clue. Inspired by this, we suggest exploiting the CLIP text encoder to integrate face embeddings into text embeddings. In this way, the sample-level face information, which is difficult to express by language, can be incorporated into textual prompts, thereby enhancing the CLIP model’s capability for efficient deepfake detection.

To achieve this, a Face2Text module is proposed and the details are shown in Fig. 4. The prompt template \mathbf{T} is designed as “A [cls] photo of a [FE] person”. Given a face encoder E_F , we obtain a face embedding S^* for an input image \mathbf{X} :

$$S^* = f_{\text{map}}(E_F(\mathbf{X})). \quad (4)$$

where f_{map} represents a frozen linear layer with a random initialized, projecting the face embeddings to align the input dimension of the text encoder E_T . By the operation TTE (Text To Embedding), the text prompt \mathbf{T} is initially tokenized and converted into word embeddings \mathbf{t} , which can be expressed as

$$\mathbf{t} = TTE(\mathbf{T}). \quad (5)$$

Then, the placeholder token [FE] is replaced with the face embedding S^* , as shown below

$$\mathbf{t} = \text{Face2Text}(\mathbf{t}, S^*). \quad (6)$$

We also consider the plain prompt template “A [cls] photo of a person,” which is not processed by Eq. (6). Thus, we get four different prompt templates for real and fake labels (see Tab. 1). An interesting finding according to the results is that an asymmetric setting for real and fake categories ($\{T_0, T_3\}$) achieves the best performance. The in-depth analysis is discussed later.

Index	Textual Template
T_0	“A real photo of a person”
T_1	“A fake photo of a person”
T_2	“A real photo of a [FE] person”
T_3	“A fake photo of a [FE] person”

Table 1: Candidate templates of text prompts.

Therefore, for training or inference, a text embedding t is fed into text encoder E_T of the pre-trained CLIP to get the text feature w , i.e.,

$$w = E_T(\mathbf{t}), \quad (7)$$

Optimization Pipeline

Following the paradigm of CLIP, the prediction probability are calculated as,

$$P(y_i | x) = \frac{\exp(\cos(w_i, f)/\tau)}{\sum_{j=0}^1 \exp(\cos(w_j, f)/\tau)}, i = \{0, 1\} \quad (8)$$

where $\cos(\cdot)$ denotes the cosine similarity, τ is the temperature parameter of CLIP, f is the image feature, and w_i is the text feature from the text prompt of label y_i .

During training, the learnable visual prompt δ is optimized by maximizing the likelihood of the correct label. During inference, δ is pad around the shrunk test samples and for predictions. On the training dataset \mathcal{D} , the optimization target is to minimize \mathcal{L} by tuning δ , which can be formulated as

$$\mathcal{L} = \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} \left[- \sum_n^N P(y_n | x_n) \log(P(y_n | x_n)) \right], \quad (9)$$

where N is the size of training dataset \mathcal{D} , x_n is the n -th image sample, and y_n is the ground truth label for x_n . We initialize the δ to zero, and follow a simple gradient-based approach to directly optimize the visual prompt via back-propagation which updates δ at step k by

$$\delta^{k+1} = \delta^k - \gamma \nabla_{\delta} \mathcal{L}, \quad (10)$$

where γ is the learning rate.

Notably, Eq. (8) has a similar formulation to a traditional fully connected layer with input dimension N and output dimension 2 if we treat the text features $[w_0, w_1]$ as the ‘weight’ of a fully connected layer for the real and fake categories. Given that the adaptative text feature for each input image in our method, it implies that the model will classify samples based on sample-level dynamic classifiers. Therefore, the Face2Text module will also significantly enrich the supervised information during optimization.

Experiments

Experimental Settings

Datasets Following most previous works, we mainly conducted training on the FaceForensics++ (FF++) (Rossler et al. 2019). It contains 1000 Pristine (PT) videos (i.e., the

Method	Frame-level				Method	Video-level			
	CDF	Wild	DFDCP	DFDC		CDF	Wild	DFDCP	DFDC
UIA-ViT (ECCV 2022)	82.41	-	75.80	-	DCL (AAAI 2022)	88.24	76.87	77.57	75.03
CFM (TIFS 2024)	82.78	78.39	-	75.82	AUNet (CVPR 2023)	92.77	-	86.16	73.82
SLADD (CVPR 2022)	79.70	-	-	77.20	SBI (CVPR 2022)	88.61	70.27	84.80	71.70
FoCus (TIFS 2024)	76.13	73.31	76.62	68.42	TALL (ICCV 2023)	90.79	-	-	76.78
UCF (ICCV 2023)	75.27	-	75.94	71.91	TALL++ (IJCV 2024)	91.96	-	-	78.51
Ba et al. (AAAI 2024)	86.40	-	85.10	72.10	SeeABLE (ICCV 2023)	87.30	-	86.30	75.90
LSDA (CVPR 2024)	83.00	-	81.50	73.60	LAA-Net (CVPR 2024)	95.40	80.03	86.94	-
VLFFD (arXiv 2023)	84.80	83.55	84.74	-	IID (CVPR 2023)	83.80	-	-	81.23
SA3WT (IJCV 2024)	83.80	-	-	76.02	Bi-LIG (TIFS 2024)	97.93	83.00	91.24	82.57
Ours (DF)	78.61	86.60	86.15	72.43	Ours (DF)	88.41	87.73	90.68	77.19
Ours (FF++)	80.00	85.42	90.57	77.34	Ours (FF++)	89.94	88.05	95.03	80.99

Table 2: AUC (%) of cross-datasets evaluations. The results of other SOTA methods are directly cited from their corresponding original paper. The best results are highlighted.

real sample) and 5000 fake videos forged by five manipulation methods, i.e., Deepfakes (DF), Face2Face (F2F), FaceSwap (FS), NeuralTextures (NT) and FaceShifter (FSh). Besides, FF++ provides three quality levels in compression for these videos: raw, high-quality (HQ) and low-quality (LQ). The **HQ version of FF++** is adopted by default in this paper. The samples were split into disjoint training, validation, and testing sets at the video level follows the official protocol. To demonstrate the performances in cross-dataset settings, four additional datasets are adopted, i.e., Celeb-DF-v2 (CDF) (Li et al. 2020c), DeepFake Detection Challenge preview (DFDCP) (Dolhansky et al. 2019), DeepFake Detection Challenge public (DFDC) (Dolhansky et al. 2020) and WildDeepfake (Wild) (Zi et al. 2020).

Implementation details We employed CLIP-ViT-L (Radford et al. 2021) as the foundation model, which is pretrained on LAION-400M. A pre-trained Transface (Dan et al. 2023) was employed as E_F . The size of input for foundation model is 224×224 . We set $p = 34$ for Input Transformations, so that trainable parameters of our method is 0.078M according to Eq. (2). We employed the AdamW optimizer with the learning rate 1.0, and the weight decay was fixed at 0. Besides, the data preprocessing transform was as same as the original CLIP (Radford et al. 2021) Notably, the visual prompt δ was initialized by zero.

Evaluation metrics In this work, we mainly report the area under the ROC curve (AUC) to compare with prior works. The video-level results are obtained by averaging predictions over each frame on an evaluated video. To facilitate a comprehensive comparison of our method with others, we also present the results of the equal error rate (EER) in our appendix (Lin et al. 2024a).

Comparisons with State-of-the-Arts

To comprehensively evaluate the generalizability of our method, we compare the performances of cross-datasets and cross-manipulation evaluations with several SOTA methods published in the past three years.

Training Data	Method	DF	FS	FSh
DF	DCL	99.98	61.01	68.45
	IID	99.51	63.83	73.49
	Ours	99.36	94.94	81.51
FS	DCL	74.80	99.90	64.86
	IID	75.39	99.73	66.18
	Ours	98.31	99.59	85.21
FSh	DCL	63.98	58.43	99.49
	IID	65.42	59.50	99.50
	Ours	89.99	81.22	99.82

Table 3: AUC (%) on cross-manipulation evaluations. The best cross-manipulation results are highlighted.

Cross-Dataset Evaluations The cross-datasets evaluation is still a challenging task because the unknown domain gap between the training and testing datasets can be caused by different source data, forgery methods, and/or post-processing. In this part, we evaluate the generalization performances in a cross-dataset setting, in which detection models were trained on the FF++ (only containing DF, F2F, FS, and NT subsets for fair comparisons) and tested on other datasets. Our method is compared with several state-of-the-art (SOTA) methods proposed in the past three years, including: UiA-ViT (Zhuang et al. 2022), DCL (Sun et al. 2022), CFM (Luo et al. 2023), AUNet (Bai et al. 2023), SLADD (Chen et al. 2022), SBI (Shiohara and Yamasaki 2022), FoCus (Tian et al. 2024), UCF (Yan et al. 2023a), TALL++ (Xu et al. 2024), TALL (Xu et al. 2023), (Ba et al. 2024), SeeABLE (Larue et al. 2023), LSDA (Yan et al. 2024a), LAA-Net (Nguyen et al. 2024), VLFFD (Sun et al. 2023), IID (Huang et al. 2023), SA3WT (Li et al. 2024), and Bi-LIG (Jiang et al. 2024). The experimental results in terms of frame-level and video-level AUC are shown in Tab. 2. Aside from its moderate performance on the CDF dataset, our method outperforms most competitors on the Wild, DFDC, and DFDCP datasets. We also report the results trained on FF++/DF, and found our method still performs better than all the competitors on the Wild, DFDC, and DFDCP datasets.

It is important to highlight that all competitors re-train backbone networks (e.g., ResNet18 and ViT), which consist of at least 10M trainable parameters. In contrast, our method utilizes a mere 0.078M parameters, illustrating its superior general performance with fewer learnable parameters.

Cross-Manipulation Evaluations Existing face forgery detectors often struggle to handle emerging manipulation techniques. In this part, we conduct cross-manipulation experiments involving three forgery techniques, i.e., Deepfakes (DF), FaceSwap (FS), and FaceShifter (FSh). We examine models trained on one manipulation type and tested across the other three. As shown in Tab. 3, it can be observed that our method can improve cross-manipulation performances. These results highlight the effectiveness of our method in combating emerging unseen forgery methods.

Ablation Studies

In this part, we perform several evaluations to explore the effectiveness of ReDFD. The main results of these experiments are cross-dataset performances trained on FF++/DF.

Impact of reprogramming paradigm In this part, we evaluate the effectiveness of our reprogramming paradigm compared with other tuning paradigms. Specifically, for the fine-tuning of the image feature encoder E_I , two established methods were assessed: Full Fine-Tuning (FFT), entailing the adjustment of all parameters within E_I , and Linear Probing (LP), which incorporated a learnable linear layer while maintaining E_I as frozen. In addition, we reference a very recent work MoE-FFD (Kong et al. 2024), which presents a deepfake detection method jointly utilizing the LoRA and Adapter paradigms to tune a frozen E_I .

Beyond these three image feature extractor-related tuning paradigms, our study also incorporated a text feature extractor-related tuning paradigm, CoOp (Zhou et al. 2022), into the comparison. It maintains the CLIP frozen while introducing learnable text prompts to adapt CLIP for target tasks. As shown in Tab. 4, our approach outperforms other methods in most scenarios. LP and CoOp inadequately facilitate the transfer of the base CLIP model to the deepfake detection task, resulting in suboptimal generalization performance. We consider that these two paradigms share a common issue: they affect the utilization of image features rather than their extraction. Compared with them, FFT, MoE-FFD, and ours can directly impact image feature extraction, consequently improving general performance across multiple datasets. Furthermore, our method achieves superior generalization using significantly fewer parameters. Although the number of trainable parameters does not increase, the performance gains when using a larger CLIP model with our method. Conversely, the performance of MoE-FFD significantly declines on a larger model due to more trainable parameters, suggesting possible overfitting. We speculate that utilizing a limited number of parameters to fit deepfake-related knowledge potentially preserves the generalization capabilities of the base model. It reveals the potential of our approach to be effectively scaled to larger models.

Model	Method	# Para	CDF	Wild	DFDCP	Avg
ViT-L	FFT	303M	83.12	70.20	90.58	81.30
	LP	0.002M	75.78	74.33	76.73	75.62
	MoE	41.34M	86.21	80.00	77.51	81.24
	CoOp	0.057M	74.72	74.07	75.82	74.87
	Ours	0.078M	89.94	88.05	95.03	91.01
ViT-B	FFT	86M	79.64	66.84	89.86	78.78
	LP	0.001M	61.96	68.81	76.91	69.23
	MoE	15.51M	91.28	83.91	84.97	86.72
	CoOp	0.038M	67.43	64.47	76.05	69.32
	Ours	0.078M	86.81	81.53	91.93	86.76

Table 4: Comparison of AUC (%) across different tuning paradigms in a cross-dataset setting. Results for MoE correspond to MoE-FFD and are sourced from the original publication. The 'Avg' column denotes the mean AUC computed over various datasets

Impact of different text prompts In this part, we investigate the effects of various text prompt configurations on RepDFD, including fixed text prompts, randomly initialized text prompts, and our adaptative face-related text prompts (termed *dynamic text prompts*). It can be concluded from the experiment that the dynamic text prompts are more effective than the fixed those. We consider all groups of the real/fake text templates in Tab. 1, i.e., $\{T_0, T_1\}$, $\{T_2, T_1\}$, $\{T_2, T_3\}$. Besides, we consider a special text prompt setting, named 'Rand Text', which contains two completely random initialized and frozen text embedding as long as the 'Fixed Text' ($\{T_0, T_1\}$). As shown in Tab. 5, the best performance occurred when the face embeddings were only utilized in the text prompt for the fake class, corresponding to the dynamic text prompts setting $\{T_0, T_3\}$. We observed that the 'Fixed Text' ($\{T_0, T_1\}$) demonstrated limited effectiveness, performing similarly to 'Rand Text', which lacks substantive semantic content. This observation implies that semantic content in language may not significantly influence deepfake detection tasks, as the primary focus is on trace analysis. In contrast, integrating face embeddings into text prompts to create dynamic text prompts, corresponding to $\{T_2, T_3\}$, $\{T_2, T_1\}$, $\{T_0, T_3\}$, boosted model performance. It may introduce fine-grained and face-related visual information into text prompts, thereby supplementing details that are challenging to describe linguistically. Thus, the dynamic text prompts can not only enrich the supervision information during training, but also provide sample-level adaptative information to support classification.

Impact of different face embeddings To further verify the universality of Face2Text, we investigate the impact of different face embeddings on Face2Text prompts. We compare the results obtained by ArcFace (Deng et al. 2019), BlendFace (Shiohara, Yang, and Taketomi 2023) and Transface (Dan et al. 2023) and fixed text (i.e., using the text prompt group $\{T_0, T_1\}$). As shown in Fig. 5, our method demonstrates good performance across various face encoders E_F .

Method	CDF	Wild	DFDCP
Rand Text	82.46	82.30	89.00
$\{T_0, T_1\}$	82.88	84.91	88.62
$\{T_2, T_3\}$	85.98	80.85	87.28
$\{T_2, T_1\}$	85.26	84.80	87.38
$\{T_0, T_3\}$ (Ours)	88.41	87.73	90.68

Table 5: Comparisons of AUC (%) across different text prompt configurations in a cross-dataset setting. These models were trained on FF++ (DF)

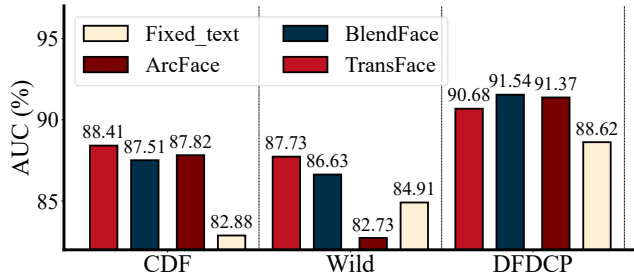


Figure 5: Comparisons of AUC (%) of our method incorporating with various face embeddings. These models were trained on FF++ (DF)

Discussion

In this subsection, we present several in-depth analyses below, aiming to explore the effectiveness of our method.

Why are asymmetric text prompts effective? In Tab. 5, it is noteworthy that optimal performance is achieved when the $\{T_0, T_3\}$ configuration is selected, which solely incorporates face embedding into the text for the fake label. To investigate the reason, as shown in Tab. 6, we calculated the cosine similarity between the image features extracted by the initialized visual prompt δ and the text prompts, both with and without face embedding. Our findings indicate that cosine similarities tend to be higher for the real label (T_0) when using the text prompt without face embeddings, and for the fake label (T_3) when using the text prompt with face embeddings. We speculate that text prompts with higher cosine similarities offer a better initialization for the CLIP model, enabling more effective fine-tuning of target models. This observation may benefit future methodological design.

The visualization of visual feature distribution In this experiment, we provide the t-SNE visualization of visual feature distributions as shown in Fig. 6. The influence of

Data	T_0	T_1	T_2	T_3
FF++	0.1889	0.1989	0.1757	0.2053
CDF	0.1996	0.2002	0.1851	0.2021

Table 6: The cosine similarity calculated between image features and various configurations of text prompts.

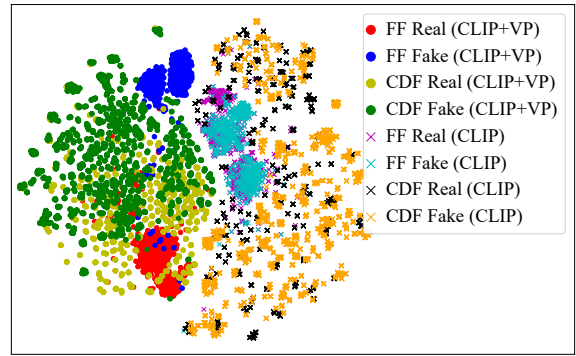


Figure 6: The t-SNE visualization of image features. ‘CLIP’ denotes features extracted by the original CLIP, while ‘CLIP+VP’ refers to features extracted by the visual prompt.

the visual prompt δ is limited exclusively to the input images and not the models, ensuring that the image features can occupy a consistent feature latent space. Initial observations suggest that there are significant differences in the original feature distributions between the CDF and FF datasets, which indicates the intrinsic domain discrimination capability of the unmodified CLIP. Moreover, without visual prompts, the original CLIP model failed to identify real and fake images. In contrast, within the FF dataset, real and fake images can be efficiently discriminated by equipping visual prompts, suggesting the effectiveness of our method. Furthermore, although the CDF dataset was not used during training, visual prompts trained on the FF dataset effectively endowed the model with the capability to detect deepfakes in the CDF dataset. We speculate that the visual prompt δ , consisting of a limited number of adjustable parameters, potentially tends to exploit the inherent capabilities of frozen models instead of introducing additional deepfake information directly, which may protect the generalization ability of pre-trained models and improve models’ performance. Except for the observation of visual features, we also provide the visualization of text features and face embeddings in our appendix (Lin et al. 2024a), which reveals a mid-domain between different datasets in common. These observations are necessary to be further explored in future works.

Conclusion

In this paper, we have proposed RepDFD, a general but parameter-efficient method for detecting face forgeries by reprogramming a well-trained CLIP model. Specifically, we employ the Input Transformation to merge the image with learnable perturbations before feeding it into the CLIP image encoder. Moreover, we introduce the Face2Text Prompts to asymmetrically incorporate facial embedding information into the text prompts for real and fake categories, which are then fed into the CLIP text encoder to guide the optimization of perturbations. RepDFD has effectively employed a CLIP to detect deepfakes by processing only the input images and texts, excluding the internal model parameters. Comprehensive experiments have demonstrated that our superior performance can be achieved with fewer trainable parameters.

Acknowledgements

This work was supported in part by NSFC (Grant U23B2022, 62202310, U22B2047); in part by Guangdong Provincial Key Laboratory (Grant 2023B1212060076); and in part by the Special Funds for Graduate Students at Shenzhen University (Grant 868-000002020210).

References

- Ba, Z.; Liu, Q.; Liu, Z.; Wu, S.; Lin, F.; Lu, L.; and Ren, K. 2024. Exposing the Deception: Uncovering More Forgery Clues for Deepfake Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2): 719–728.
- Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; and Isola, P. 2022. Exploring Visual Prompts for Adapting Large-Scale Models. arXiv:2203.17274.
- 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 (CVPR)*, 24709–24719.
- Cai, C.; Ye, Z.; Feng, L.; Qi, J.; and Liu, F. 2024. Sample-Specific Masks for Visual Reprogramming-based Prompting. In *Proceedings of the 41st International Conference on Machine Learning*, 5383–5408. PMLR.
- Chen, L.; Zhang, Y.; Song, Y.; Liu, L.; and Wang, J. 2022. Self-Supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18710–18719.
- Chen, P.-Y. 2024. Model Reprogramming: Resource-Efficient Cross-Domain Machine Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20): 22584–22591.
- Chen, X.; Ni, B.; Liu, Y.; Liu, N.; Zeng, Z.; and Wang, H. 2024a. SimSwap++: Towards Faster and High-Quality Identity Swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1): 576–592.
- Chen, Z.; Fang, S.; Liu, W.; He, Q.; Huang, M.; and Mao, Z. 2024b. DreamIdentity: Enhanced Editability for Efficient Face-Identity Preserved Image Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(2): 1281–1289.
- Cozzolino, D.; Poggi, G.; Corvi, R.; Nießner, M.; and Verdoliva, L. 2024. Raising the Bar of AI-generated Image Detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 4356–4366.
- Dan, J.; Liu, Y.; Xie, H.; Deng, J.; Xie, H.; Xie, X.; and Sun, B. 2023. TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20642–20653.
- Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.
- Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The DeepFake Detection Challenge (DFDC) Dataset. arXiv:2006.07397.
- Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; and Ferrer, C. C. 2019. The Deepfake Detection Challenge (DFDC) Preview Dataset. arXiv:1910.08854.
- Elsayed, G. F.; Goodfellow, I.; and Sohl-Dickstein, J. 2019. Adversarial Reprogramming of Neural Networks. In *International Conference on Learning Representations*.
- 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 (CVPR)*, 4490–4499.
- Jiang, P.; Xie, H.; Yu, L.; Jin, G.; and Zhang, Y. 2024. Exploring Bi-Level Inconsistency via Blended Images for Generalizable Face Forgery Detection. *IEEE Transactions on Information Forensics and Security*, 19: 6573–6588.
- Kong, C.; Luo, A.; Bao, P.; Yu, Y.; Li, H.; Zheng, Z.; Wang, S.; and Kot, A. C. 2024. MoE-FFD: Mixture of Experts for Generalized and Parameter-Efficient Face Forgery Detection. arXiv:2404.08452.
- Larue, N.; Vu, N.-S.; Struc, V.; Peer, P.; and Christophides, V. 2023. SeeABLE: Soft Discrepancies and Bounded Contrastive Learning for Exposing Deepfakes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21011–21021.
- Li, L.; Bao, J.; Yang, H.; Chen, D.; and Wen, F. 2020a. Advancing High Fidelity Identity Swapping for Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5074–5083.
- Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; and Guo, B. 2020b. Face X-Ray for More General Face Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5001–5010.
- Li, Y.; Yang, X.; Sun, P.; Qi, H.; and Lyu, S. 2020c. Celeb-DF: A Large-Scale Challenging Dataset for Deep-Fake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3207–3216.
- Li, Y.; Zhang, Y.; Yang, H.; Chen, B.; and Huang, D. 2024. SA 3 WT: Adaptive Wavelet-Based Transformer with Self-Paced Auto Augmentation for Face Forgery Detection. *International Journal of Computer Vision*, 132: 4417–4439.
- Lin, K.; Lin, Y.; Li, W.; Yao, T.; and Li, B. 2024a. Standing on the Shoulders of Giants: Reprogramming Visual-Language Model for General Deepfake Detection. arXiv:2409.02664.
- Lin, Y.; Song, W.; Li, B.; Li, Y.; Ni, J.; Chen, H.; and Li, Q. 2024b. Fake It till You Make It: Curricular Dynamic Forgery Augmentations Towards General Deepfake Detection. In *Computer Vision – ECCV 2024*, 104–122.
- Luo, A.; Kong, C.; Huang, J.; Hu, Y.; Kang, X.; and Kot, A. C. 2023. Beyond the prior forgery knowledge: Mining critical clues for general face forgery detection. *IEEE Transactions on Information Forensics and Security*, 19: 1168–1182.

- Nguyen, D.; Mejri, N.; Singh, I. P.; Kuleshova, P.; Astrid, M.; Kacem, A.; Ghorbel, E.; and Aouada, D. 2024. LAA-Net: Localized Artifact Attention Network for Quality-Agnostic and Generalizable Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17395–17405.
- Ojha, U.; Li, Y.; and Lee, Y. J. 2023. Towards Universal Fake Image Detectors That Generalize Across Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24480–24489.
- Papantoniou, F. P.; Lattas, A.; Moschoglou, S.; Deng, J.; Kainz, B.; and Zafeiriou, S. 2024. Arc2Face: A Foundation Model for ID-Consistent Human Faces. In *Computer Vision – ECCV 2024*, 241–261.
- Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-Aware Clues. In *ECCV*, 86–103.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. PMLR.
- Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; and Niessner, M. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1–11.
- Sha, Z.; Li, Z.; Yu, N.; and Zhang, Y. 2023. DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 3418–3432.
- Shiohara, K.; and Yamasaki, T. 2022. Detecting Deepfakes With Self-Blended Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18720–18729.
- Shiohara, K.; Yang, X.; and Taketomi, T. 2023. Blend-Face: Re-designing Identity Encoders for Face-Swapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7634–7644.
- Sun, K.; Chen, S.; Yao, T.; Sun, X.; Ding, S.; and Ji, R. 2023. Towards General Visual-Linguistic Face Forgery Detection. arXiv:2307.16545.
- 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, J.; Chen, P.; Yu, C.; Fu, X.; Wang, X.; Dai, J.; and Han, J. 2024. Learning to Discover Forgery Cues for Face Forgery Detection. *IEEE Transactions on Information Forensics and Security*, 19: 3814–3828.
- Tsao, H.-A.; Hsiung, L.; Chen, P.-Y.; Liu, S.; and Ho, T.-Y. 2024. AutoVP: An Automated Visual Prompting Framework and Benchmark. In *The Twelfth International Conference on Learning Representations*.
- Wang, Q.; Liu, F.; Zhang, Y.; Zhang, J.; Gong, C.; Liu, T.; and Han, B. 2022. Watermarking for Out-of-distribution Detection. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 15545–15557.
- Xu, C.; Zhang, J.; Han, Y.; Tian, G.; Zeng, X.; Tai, Y.; Wang, Y.; Wang, C.; and Liu, Y. 2022. Designing One Unified Framework for High-Fidelity Face Reenactment and Swapping. In *Computer Vision – ECCV 2022*, 54–71.
- Xu, Y.; Liang, J.; Jia, G.; Yang, Z.; Zhang, Y.; and He, R. 2023. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 22658–22668.
- Xu, Y.; Liang, J.; Sheng, L.; and Zhang, X.-Y. 2024. Learning Spatiotemporal Inconsistency via Thumbnail Layout for Face Deepfake Detection. *International Journal of Computer Vision*, 132: 5663–5680.
- Yan, Z.; Luo, Y.; Lyu, S.; Liu, Q.; and Wu, B. 2024a. Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8984–8994.
- Yan, Z.; Yao, T.; Chen, S.; Zhao, Y.; Fu, X.; Zhu, J.; Luo, D.; Wang, C.; Ding, S.; Wu, Y.; and Yuan, L. 2024b. DF40: Toward Next-Generation Deepfake Detection. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Yan, Z.; Zhang, Y.; Fan, Y.; and Wu, B. 2023a. UCF: Uncovering Common Features for Generalizable Deepfake Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 22412–22423.
- Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; and Wu, B. 2023b. DeepfakeBench: A Comprehensive Benchmark of Deepfake Detection. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zhao, H.; Zhou, W.; Chen, D.; Wei, T.; Zhang, W.; and Yu, N. 2021. Multi-Attentional Deepfake Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2185–2194.
- Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to Prompt for Vision-Language Models. *International Journal of Computer Vision*, 130: 2337–2348.
- Zhuang, W.; Chu, Q.; Tan, Z.; Liu, Q.; Yuan, H.; Miao, C.; Luo, Z.; and Yu, N. 2022. UIA-ViT: Unsupervised Inconsistency-Aware Method Based on Vision Transformer for Face Forgery Detection. In *Computer Vision – ECCV 2022*, 391–407.
- Zi, B.; Chang, M.; Chen, J.; Ma, X.; and Jiang, Y.-G. 2020. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2382–2390.