Multi-modal Deepfake Detection via Multi-task Audio-Visual Prompt Learning

Hui Miao, Yuanfang Guo*, Zeming Liu, Yunhong Wang

School of Computer Science and Engineering, Beihang University, China
{huimiao, andyguo, zmliu, yhwang}@buaa.edu.cn

Abstract

With the malicious use and dissemination of multi-modal deepfake videos, researchers start to investigate multi-modal deepfake detection. Unfortunately, most of the existing methods tune all the parameters of the deep network with limited speech video datasets and are trained under coarse-grained consistency supervision, which hinders their generalization ability in practical scenarios. To solve these problems, in this paper, we propose the first multi-task audio-visual prompt learning method for multi-modal deepfake video detection, by exploiting multiple foundation models. Specifically, we construct a two-stream multi-task learning architecture and propose sequential visual prompts and short-time audio prompts to extract multi-modal features, which are aligned at the frame level and utilized in subsequent fine-grained feature matching and fusion. Due to the natural alignment of visual content and audio signal in real data, we propose a frame-level cross-modal feature matching loss function to learn the fine-grained audio-visual consistency. Comprehensive experiments demonstrate the effectiveness and superior generalization ability of our method against the state-of-the-art methods.

Introduction

Benefiting from the advances in deep learning technology, malicious users can easily utilize DNN based generation techniques, such as faceswap[1] and SV2TTS (Jia et al. 2018), to generate visual or audio content for producing deep forged videos, which is also known as deepfakes, without requiring too much professional knowledge. These deepfake videos, which may be maliciously applied in fabricating political rumors[2] and spreading misinformation[3], can induce severe social problems and trust issues.

In real scenarios, since the majority of deepfake videos possess both the visual and audio signals to jointly convey information, these multi-modal deepfake data can be briefly produced via different types of manipulations, including visual-only manipulations, audio-only manipulations, misalignment manipulations and visual-audio manipulations. Under such circumstances, uni-modal deepfake detection techniques (Rossler et al. 2019; Shiohara and Yamasaki 2022; Haliassos et al. 2022; Xu et al. 2023) only focus on predicting the authenticity of the visual content, even though some methods (Haliassos et al. 2021, 2022) employ audio signals to learn high-level semantic irregularities in lip movements. They are less suitable for combating various types of multi-modal deepfake videos, such as real visual content with fake audio. Therefore, multi-modal deepfake detection has received increasing attention in recent years.

Based on the uni-modal deepfake detection methods, some approaches (Khalid et al. 2021a; Ilyas, Javed, and

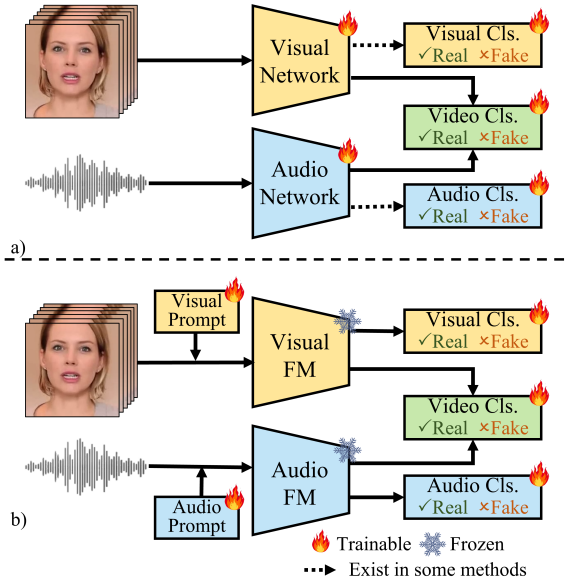

Figure 1: Overviews of multi-modal joint learning deepfake video detection methods. a) Previous methods usually tune all the visual and audio network parameters and some methods only consider audio-visual consistency to make the prediction. b) Our method proposes visual and audio prompts to adapt different foundation models to multi-modal deepfake video detection in a multi-task learning manner.

*Corresponding author.

[1]https://faceswap.dev/

[2]https://www.youtube.com/watch?v=30NvDC1zcL8

[3]https://www.youtube.com/watch?v=3wVpVH0Wa6E

Malik 2023) employ ensemble strategies, which directly combine the audio and visual prediction results, to detect deepfake videos. However, these methods tend to consider each modality independently. Thus, they have ignored the relationships between audio and visual content. Many researchers have realized that there exists the audio-visual consistency, such as the natural alignment between lip movement and spoken phoneme, which can be utilized to improve the detection performance. Therefore, a series of multi-modal joint learning approaches combined with audio-visual consistency constraints are proposed, including employing phoneme-viseme mismatches for specific pronunciation (Agarwal et al. 2020), introducing contrastive learning (Chugh et al. 2020; Zou et al. 2024), designing a novel multi-modal network architecture (Yang et al. 2023), and utilizing audio-visual feature representation learning (Oorloff et al. 2024).

Unfortunately, all the parameters of the above methods are tuned on limited speech video datasets, which hinders their generalization ability in practical scenarios. Besides, a certain number of methods (Mittal et al. 2020; Oorloff et al. 2024) focus on exploiting audio-visual inconsistencies to classify the multi-modal video yet ignoring the forgery artifacts within the forged modality. This mechanism tends to give less satisfactory performance when the audio-visual signals have high consistencies. Although other methods have considered the intra-modal forgery artifacts, they utilize learnable feature extractors to encode the raw input data directly under the supervision of standard classification loss function, which is less effective in capturing the forgery cues within data processing units according to the characteristic of the input modality (*e.g.*, audio deepfake feature within time windows or visual deepfake feature within frames). In addition, existing methods only construct the contrastive learning loss at the segment level, which prevents them from learning fine-grained frame-level consistency features. If the attackers generate lip-sync deepfake data, the coarse features may not be able to precisely model the slight inconsistencies in the forged videos.

To resolve the above problems, in this paper, we propose a multi-modal deepfake detection method, named multi-modal deepfake detection via multi-task audio-visual prompt learning, by exploiting foundation models and considering both the intra-modality artifacts and inter-modality consistencies, to extract detection features which are suitable for detecting different types of manipulations in multi-modal deepfakes with better generalization ability. The overviews of our method and the previous methods are shown in Fig. 1.

Specifically, we construct a two-stream multi-task learning architecture and exploit multiple foundation models (FMs), which possess superior generalization abilities (Radford et al. 2021; Zhou et al. 2022; Cheng, Liang, and Tan 2024) than common DNNs, to improve the generalization ability hindered by insufficient training data and extract multi-modal detection features. To effectively fine-tune the foundation models with limited multi-modal deepfake data, we propose sequential visual prompts and short-time audio prompts. These proposed multi-modal prompts not only consider the intra-modality artifacts, but also contribute to the learning of natural audio-visual synchronizations. For the visual branch, since its inputs are sequential frames (images), we exploit the visual foundation model (*e.g.*, CLIP (Radford et al. 2021)) as our backbone network and propose sequential visual prompts, which preserve the learned information from each encoder layer, to extract frame-level visual features and can better describe the detection features for deepfake videos. For the audio branch, to extract the fine-grained phoneme feature, we exploit a speech recognition foundation model (*e.g.*, Whisper (Radford et al. 2023)) as the backbone network, and introduce learnable short-time audio prompts at feature dimension. This operation captures the audio deepfake cues in a short time and the learned embeddings can be aligned with the visual feature at the frame level, which is also beneficial for the subsequent fine-grained feature matching and fusion. To synchronize and fuse the visual and audio features, by improving the existing segment-level contrastive learning loss, we propose a frame-level cross-modal feature matching loss function, which can extract the fine-grained audio-visual consistency features.

Our major contributions are summarized as follows:

- To the best of our knowledge, we propose the first multi-task multi-modal prompt learning method for multi-modal deepfake detection, by jointly exploiting frozen visual and audio foundation models, to further learn task-related knowledge and improve the effectiveness and generalization ability with the limited deepfake training data.
- We construct a two-stream multi-task learning architecture to effectively combat the potential different manipulations in multi-modal deepfake videos.
- We propose sequential visual prompts and short-time audio prompts combined with an effective frame-level cross-modal feature matching loss function to efficiently utilize the intra-modality artifacts and learn the fine-grained audio-visual consistency features.
- Comprehensive experiments demonstrate the effectiveness and generalization ability of our method against the state-of-the-art method, and the qualitative analysis illustrates the efficacy of the proposed fine-grained cross-modal feature matching loss function.

Related Work

Foundation Models and Prompt Learning

In recent years, foundation models, which are trained on diverse datasets with a large number of parameters and possess strong zero-shot capability, have attracted many researchers to apply them to various downstream tasks. Visual Language Models (VLM) (*e.g.*, CLIP (Radford et al. 2021)), which play an important role in foundation models, have been extensively utilized in many vision tasks, such as semantic segmentation (Zhang et al. 2024), image denoising (Cheng, Liang, and Tan 2024), object detection (Vidit, Engilberge, and Salzmann 2023) and facial-related tasks (Zhao and Patras 2023; Lin et al. 2024; Zhou, Zhong, and Öztireli 2023).

In speech processing, benefiting from the large-scale training data and the supervision of multiple languages and tasks, the accuracy and robustness of the speech recognition foundation models (*e.g.*, Whisper (Radford et al. 2023)) are rapidly approaching that of human. Due to the strong visual and phoneme feature extraction ability of foundation models, we explore to adopt these visual and audio foundation models to deepfake video detection.

Inspired by the learnable prompts in the language model (Li and Liang 2021), a series of methods (Jia et al. 2022; Liu et al. 2024) apply this parameter-efficient fine-tuning technique to adapt foundation models to downstream vision and audio tasks. Although these prompts achieve good performance with a small amount of trainable parameters, they only utilize the inherent ability of the foundation models. In this paper, we further propose sequential visual prompts and short-time audio prompts to extract fine-grained deepfake embeddings, which are more suitable for multi-modal deepfake detection.

Deepfake Detection

Visual-related Deepfake Detection Researchers have conducted extensive studies on visual-related deepfake video detection (Khan and Dai 2021; Zhang et al. 2022; Tan et al. 2023), which consider the spatial artifacts as well as the temporal features. Early approaches explicitly employed specific sequence-level physiological signals, such as eye blinking (Li, Chang, and Lyu 2018) and head poses (Yang, Li, and Lyu 2019). However, these methods only focused on fixed biometric patterns, and usually ignored the potential spatial and temporal forgery features. Thus, they quickly became less effective due to the rapid evolvement of deepfake techniques. To address this limitation, a series of methods (Sabir et al. 2019; Gu et al. 2021, 2022) have been proposed to adaptively learn the spatial and temporal features, by constructing various network architecturess (Zheng et al. 2021; Xu et al. 2023, 2024), utilizing different learning strategies (Choi et al. 2024), or employing augmented input data (Wang et al. 2023).

Audio-visual Deepfake Detection Recently, (Khalid et al. 2021b) noticed the potential harms of the generated multi-modal content and introduced a novel audio-visual multi-modal deepfake dataset (FakeAVCeleb), which contains all possible combinations of audio and visual deepfake forgeries. To detect these different types of multi-modal manipulations, some approaches (Khalid et al. 2021a) utilize the ensemble-based strategy, which processes each modality separately yet ignores the relationships between audio and visual content. To address this limitation, many researchers (Mittal et al. 2020; Chugh et al. 2020; Zou et al. 2024) employ multi-modal joint learning and introduce audio-visual consistency constraints. AVFF (Oorloff et al. 2024) proposes a two-stage cross-modal learning method for audio-visual representation learning and deepfake classification. AVoid-DF (Yang et al. 2023) constructs a temporal-spatial encoder and a multi-modal joint decoder to extract and fuse the multi-modal features. (Salvi et al. 2023) utilizes frozen DNNs to extract visual and audio features and discusses various audio-visual feature fusion strategies. Multimodaltrace (Raza and Malik 2023) reformulates the multi-modal deepfake detection as a multi-label classification problem.

In general, the visual-related methods only utilize uni-modal cues, i.e., they can hardly utilize multi-modal information. On the other hand, most of the existing multi-modal methods tune all the parameters on limited datasets and cannot extract fine-grained audio-visual consistency features, which hinder the improvement of performance and generalization. To solve these problems, we propose two prompts to adapt foundation models, which posses superior generalization ability, to deepfake detection, and propose a frame-level cross-modal feature matching loss function to learn the fine-grained audio-visual consistency features.

Backgrounds

Since we adopt CLIP (Radford et al. 2021) and Whisper (Radford et al. 2023) as the foundation models in our method, a brief introduction to these two models is provided in this section.

As a typical Vision-Language Model, CLIP utilizes an image encoder and a text encoder to transform images and text into a joint embedding space. In this paper, the image encoder of CLIP is utilized to extract the visual features. Therefore, the details of this encoder are presented here.

Given an input image $x_v \in \mathbb{R}^{C\times H\times W}$, where C, H, W are the number of channels, height and width, the image encoder firstly divides it into N patches with the size of $p \times p$, $\{x_v^i \in \mathbb{R}^{C\times p\times p} | i = 1, 2, ...N\}$. Then it performs a linear projection to embed each patch to D_v dimensions, as

$$\mathrm{E}_v^0 = [\mathrm{E}x_v^1, \mathrm{E}x_v^2, \ldots, \mathrm{E}x_v^N], \mathrm{E} \in \mathbb{R}^{D_v\times(C\cdot p\cdot p)}, \quad (1)$$

where $\mathrm{E}_v^0 \in \mathbb{R}^{N\times D_v}$ represents the image embedding, E stands for the linear projection parameters, and $[\cdot,\cdot]$ refers to concatenation along the token length dimension. Then, a learnable classification token $\mathrm{z}_{v_{cls}}^0 \in \mathbb{R}^{1\times D_v}$ is prepended and a positional embedding p_v is added to the token sequence, as

$$\mathrm{z}_v^0 = [\mathrm{z}_{v_{cls}}^0, \mathrm{E}_v^0] + \mathrm{p}_v, \quad (2)$$

where $\mathrm{z}_v^0 \in \mathbb{R}^{(N+1)\times D_v}$ refers to the feature fed into the first transformer layer L_v^1. The output of the previous layer z_v^{l-1} is then input to the next transformer layer L_v^l, which can be formulated as

$$\mathrm{z}_v^l = L_v^l(\mathrm{z}_v^{l-1}), l = 1...\mathbb{L}_v, \quad (3)$$

where l denotes the index of the transformer layer and $\mathbb{L}_v$ is the number of layers in the image encoder. The output of the classification token in the last layer $z_{v_{cls}}^{\mathbb{L}_v}$ represents the feature embedding of the input image.

Whisper, which is utilized as our audio foundation model, is an excellent transformer-based sequence-to-sequence speech recognition foundation model. Since we adopt the encoder of Whisper, it is briefly introduced here.

The input audio is firstly transformed into a log-mel spectrogram $x_a \in \mathbb{R}^{N_m\times N_t}$, where N_m and N_t are the number of channels and tokens in the spectrogram. Then, a 1D convolution module is applied to project the spectrogram from

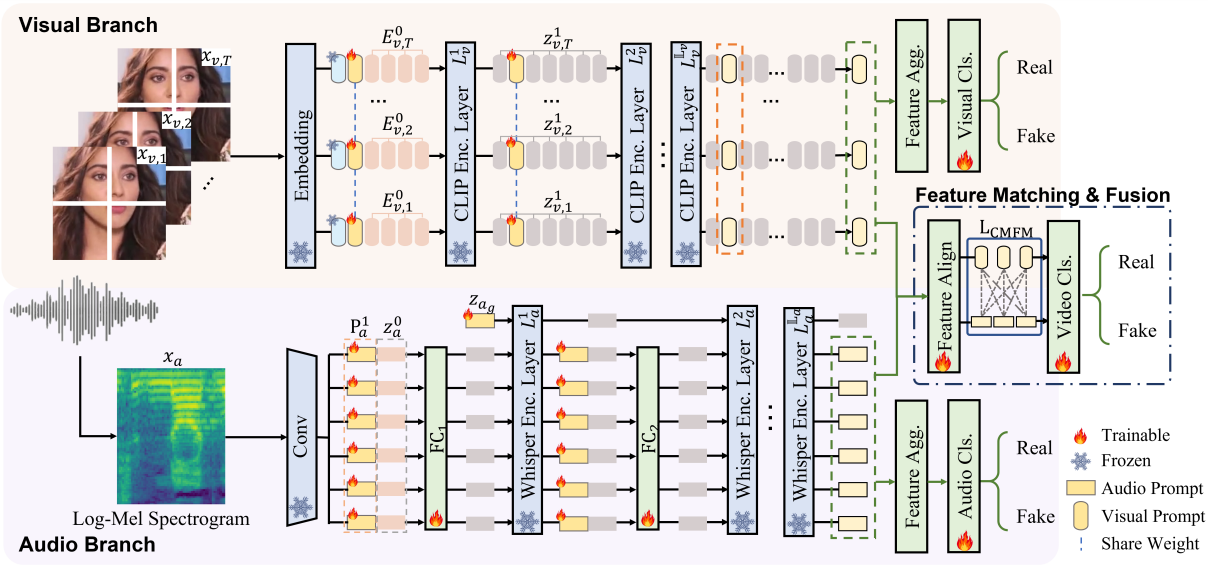

Figure 2: Overview of our proposed multi-modal deepfake detection via multi-task audio-visual prompt learning. For each visual frame, we inject learnable visual prompts before being input to the next layer while maintaining the existing injected prompts in the previous layers. In the audio branch, we inject short-time audio prompts, which are concatenated with other embeddings along the feature dimension. After feature alignment, the audio and visual features are fused to make a prediction with the help of the frame-level cross-modal feature matching (CMFM) loss function.

N_m channels to D_a channels. Similar to CLIP, a positional embedding p_a is introduced, as

$$\mathrm{z}_a^0 = ConvModule(x_a) + \mathrm{p}_a, \tag{4}$$

where $\mathrm{z}_a^0 \in \mathbb{R}^{\frac{N_t}{2} \times D_a}$ is the audio embedding. Then, the transformer layers L_a^l are adopted to process the audio embedding z_a^{l-1}, via

$$\mathrm{z}_a^l = L_a^l(\mathrm{z}_a^{l-1}), l = 1...\mathbb{L}_a, \tag{5}$$

where $\mathbb{L}_a$ is the number of layers in the Whisper encoder. The encoded embeddings will then be fed to the decoder, which is not utilized in this paper.

Methodology

In this paper, we propose a multi-modal deepfake detection method, named multi-modal deepfake detection via multi-task audio-visual prompt learning, by exploiting foundation models and considering both the intra-modality artifacts and inter-modality consistencies. Based on the characteristics of the visual and audio data, we propose sequential visual prompts and short-time audio prompts to specifically model the intra-modality artifacts and inter-modality consistencies. As shown in Fig. 2, the visual content and the corresponding audio signal are fed into the visual and audio branches, which consists of frozen foundation models and learnable prompts, respectively, to extract the multi-modal features. After feature alignment, we fuse the multi-modal features to further obtain the fine-grained audio-visual consistency features, via the help of our frame-level cross-modal feature matching loss function. Then, with respect to the final features, which contain the uni-modal and consistency information, we can predict the authenticity of the input data.

Visual Prompt Learning and Classification

Considering that the input to the visual branch is a sequence of images, we construct visual prompts for each frame to learn the fine-grained visual features. Compared to the original VPT (Jia et al. 2022), our sequential visual prompts not only utilize the inherent high-level feature extraction ability of foundation models, but also inject learnable tokens into each transformer layer to assign flexible learnability to foundation models to extract the detection feature.

Specifically, we freeze all the parameters of the CLIP image encoder. Then, we inject visual prompts into each layer while maintaining the existing injected prompts in the previous layer. This operation preserves the useful features learned from each layer, to promote the learning of deepfake-related cues. As shown in Fig. 2, the video frames $\{x_{v,t} \in \mathbb{R}^{C \times H \times W} | t = 1, 2, ...T\}$ extracted from the segment $v \in \mathbb{R}^{T \times C \times H \times W}$ are processed via Eq. 1 to obtain the frame embeddings $\{\mathrm{E}_{v,t}^0 \in \mathbb{R}^{N \times D_v} | t = 1, 2, ...T\}$. Then, the frozen classification token $z_{v_{cls}}^0$ is prepended to the embedding, and a set of learnable tokens $\{\mathbf{P}_v^l \in \mathbb{R}^{N_{vpt} \times D_v} | l = 1, 2, ...\mathbb{L}_v\}$, i.e., sequential visual prompts, are inserted before input to every transformer layer. For each layer L_v^l, the output visual feature at the t-th frame is

$$\mathrm{z}_{v,t}^l = L_v^l([\mathrm{z}_{v,t}^{l-1,0}, \mathbf{P}_v^l, \mathrm{z}_{v,t}^{l-1,1} ..., \mathrm{z}_{v,t}^{l-1,N+(l-1)\times N_{vpt}}]), \tag{6}$$

where $\mathrm{z}_{v,t}^{l,i}$ represents the i-th token vector in the l-th layer output at the t-th frame.

Then, the output $\mathrm{z}_{v,t}^{\mathbb{L}_v,1}$, which corresponds to the first newly injected learnable prompt in the last transformer layer,

is the final image representation with deepfake-related cues. It is worth noting that our proposed visual prompt is different from the original deep visual prompt (Jia et al. 2022), which replaces the prompt tokens rather than only inserts the learnable tokens. After extracting the frame-level features, the segment-level feature is aggregated by calculating the average of $\mathrm{z}_{v,t}^{\mathbb{L}_v,1}$ along the temporal dimension, i.e.,

$$\mathbf{f_v} = \frac{1}{T}\sum_{t=1}^{T}\mathrm{z}_{v,t}^{\mathbb{L}_v,1}. \tag{7}$$

Based on the segment-level visual feature $\mathbf{f_v}$, a visual classification head is introduced to predict the authenticity of the visual content. The visual classification head is formed by two linear layers with a ReLU activation. In this branch, only the visual classification head and sequential visual prompts $\mathbf{P}_v^l$ are trainable and the cross-entropy loss $\mathrm{L_v}$ is adopted as the objective function.

Audio Prompt Learning and Classification

In the audio branch, we propose short-time audio prompts to adapt the audio foundation model, *e.g.*, Whisper, to extract appropriate audio features. Different from the prompt learning in NLP, which directly injects prompts to the input sequence, we concatenate the learnable audio prompts to the embeddings along the feature dimension, before input to each transformer layer.

Specifically, according to the frame rate and audio sampling rate, the audio segment corresponding to the input frames in the visual branch is obtained. As shown in Fig. 2, similar to the data preprocessing of Whisper, the audio data is firstly transformed to the log-mel spectrogram x_a. Then, the spectrogram is processed via Eq. 4. Based on the original architecture of Whisper, we introduce a set of trainable audio prompts $\mathbf{P}_a^l \in \mathbb{R}^{\frac{N_t}{2}\times D_p}$, which are concatenated to z_a^{l-1} along the feature dimension, before input to the l-th layer L_a^l. Since each token in the embeddings contains certain information of a few neighboring time windows, our proposed audio prompts are termed as short-time audio prompts, which can capture the audio deepfake cues in a short time period. To ensure that the feature dimensions are consistent before and after audio prompts concatenation, a fully connected layer is introduced. Meanwhile, we also inject a learnable token z_{a_g} to capture global information along the temporal dimension. Then, the above process can be formulated as

$$\mathrm{z}_a^l = \begin{cases} L_a^l([\mathrm{z}_{a_g}, \mathbf{FC}_l(cat(\mathbf{P}_a^l, \mathrm{z}_a^0))]) & l = 1 \\ L_a^l([\mathrm{z}_a^{l-1,0}, \mathbf{FC}_l(cat(\mathbf{P}_a^l, \mathrm{z}_a^{l-1,1\ldots\frac{N_t}{2}}))]) & l \neq 1, \end{cases} \tag{8}$$

where $cat(\cdot,\cdot)$ refers to the concatenation operation along the feature dimension.

Since audio is a sequential signal, we directly calculate the average of the output embeddings from the last layer along the temporal dimension, without considering the first global token, to obtain the segment-level audio feature $\mathbf{f_a}$, as

$$\mathbf{f_a} = \frac{2}{N_t}\sum_{i=1}^{\frac{N_t}{2}}\mathrm{z}_a^{\mathbb{L}_a,i}. \tag{9}$$

Similar to the visual branch, the segment-level audio feature is also fed into an audio classification head, which also contains two linear layers with a ReLU activation. The trainable parts of the audio branch are the fully-connected layers $\mathbf{FC}_{1\ldots\mathbb{L}_a}$, audio prompts $\mathbf{P}_a^{1\ldots\mathbb{L}_a}$, a global token z_{a_g}, and a classification head. The loss function is the cross-entropy loss $\mathrm{L_a}$.

Frame-level Feature Matching and Fusion

Due to the natural alignment of visual content (*e.g.,* lip movement) and audio signals (*e.g.,* spoken phoneme), both the uni-modal and multi-modal manipulations can induce inconsistencies between the two modalities, which is a significant clue for multi-modal deepfake video detection. To better utilize this disharmony, we align the visual and audio features along the temporal dimension and propose a frame-level cross-modal feature matching loss function for fine-grained audio-visual consistency learning.

After the feature extraction in the visual and audio branch, the visual feature f_v and the audio feature f_a are obtained:

$$f_v = [\mathrm{z}_{v,1}^{\mathbb{L}_v,1}, \mathrm{z}_{v,2}^{\mathbb{L}_v,1}, \ldots \mathrm{z}_{v,T}^{\mathbb{L}_v,1}], f_v \in \mathbb{R}^{T\times D_v}, \tag{10}$$

$$f_a = [\mathrm{z}_a^{\mathbb{L}_a,1}, \mathrm{z}_a^{\mathbb{L}_a,2}, \ldots \mathrm{z}_a^{\mathbb{L}_a,\frac{N_t}{2}}], f_a \in \mathbb{R}^{\frac{N_t}{2}\times D_a}. \tag{11}$$

To align the audio and visual features along the temporal dimension and unify the feature dimensions to D_f, two 1D convolutional layers are introduced, as

$$f_v^{'} = \mathbf{conv}_v(f_v, k_v, s_v), f_v^{'} \in \mathbb{R}^{T\times D_f}, \tag{12}$$

$$f_a^{'} = \mathbf{conv}_a(f_a, k_a, s_a), f_a^{'} \in \mathbb{R}^{T\times D_f}, \tag{13}$$

where k_v and s_v respectively denote the kernel size and stride of the convolutional layer $\mathbf{conv}_v$ for processing the visual features. k_a and s_a respectively represents the kernel size and stride of the convolutional layer $\mathbf{conv}_a$ for processing the audio features. Note that we set $s_v = k_v = 1$ and $s_a = k_a = \frac{N_t}{2T}$ to align the sequence length of these two features to T.

Then we concatenate the aligned features along the feature dimension to obtain the fused feature $f_f \in \mathbb{R}^{T\times(2\cdot D_f)}$. At last, the fused feature is aggregated along the temporal dimension and input to a video classification head, which contains two linear layers with a ReLU activation.

Besides of the standard cross-entropy loss $\mathrm{L_f}$, we propose a novel frame-level cross-modal feature matching (CMFM) loss function, which promotes the learning of the fine-grained audio-visual consistency features in the final features. Specifically, we consider all the frame-level audio-visual pairs $\{(f_{v,m}^{'p}, f_{a,n}^{'q}), y_m\}$ in a mini-batch. Here, $f_{v,m}^{'p}$ refers to the visual feature at the p-th frame of the m-th segment in the mini-batch. Similarly, $f_{a,n}^{'q}$ stands for the audio feature at the q-th frame of the n-th segment in a mini-batch. y_m denotes the label of the m-th segment. For each pair, we not only consider the labeled data but also utilize the natural inconsistency of audio and visual content from different

Method	Acc.	AUC
MDS (Chugh et al. 2020)*	96.97	51.93
AV-DFD (Zhou and Lim 2021)*	96.97	52.04
BA-TFD (Cai et al. 2022)*	96.96	54.31
MMtrace (Raza and Malik 2023)	92.90	-
MRDF-CE (Zou et al. 2024)	94.05	92.43
AVFF (Oorloff et al. 2024)	<u>98.60</u>	99.10
AVGraph (Yin et al. 2024)	**99.84**	<u>99.94</u>
Ours	**99.84**	**99.98**

Table 1: Intra-dataset results on FakeAVCeleb. The best result is highlighted in bold font, and the second-best value is underlined. The results with * are obtained from the paper of AVGraph.

segments, to form our loss function as

$$\mathrm{L_{CMFM}} = max(\frac{1}{N_l^+}\sum_{m=1}^{B}\sum_{p=1}^{T}\sum_{n=1}^{B}\sum_{q=1}^{T} max(l_{mp,nq}, 0) + \frac{1}{N_l^-}\sum_{m=1}^{B}\sum_{p=1}^{T}\sum_{n=1}^{B}\sum_{q=1}^{T} min(l_{mp,nq}, 0), 0), \tag{14}$$

$$l_{mp,nq} = \begin{cases} \alpha \times \mathbf{D}(f_{v,m}^{'p}, f_{a,n}^{'q}) & \text{if } m = n, y_m = 0, p = q \\ -\beta \times \mathbf{D}(f_{v,m}^{'p}, f_{a,n}^{'q}) & \text{if } m = n, y_m = 1 \\ -\gamma \times \mathbf{D}(f_{v,m}^{'p}, f_{a,n}^{'q}) & \text{if } m! = n \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Note that $\mathbf{D}(\cdot,\cdot)$ denotes the cosine distance, B is the size of a mini-batch, and $N_l^+(N_l^-)$ represents the number of the values greater (less) than zero in the summation process. Intuitively, Eq. 14 minimizes the distances between matched audio-visual features and maximizes the distances between inconsistent features. Here, we define the inconsistent features as the audio-visual embeddings of the fake segments and the representations of naturally mismatched audio and visual data from different segments.

Then, the overall loss L of our method is

$$\mathrm{L} = \mathrm{L_v} + \mathrm{L_a} + \mathrm{L_f} + \mathrm{L_{CMFM}}. \tag{16}$$

Experiments

Experiment Settings

Datasets To validate the performance of our method, we evaluate our model on FakeAVCeleb (Khalid et al. 2021b), which contains 500 real videos and 19500 fake videos from different ethnic groups. To avoid identity leakage, we divide the training, validation, and testing set according to these ethnic groups. Specifically, the training set consists of South Asian, East Asian, and American Caucasian, the validation set contains African, and the testing set contains European Caucasian. In addition, by following (Feng, Chen, and Owens 2023; Oorloff et al. 2024), we also evaluate the methods on a subset of KoDF (Kwon et al. 2021) to assess the cross-dataset generalization. It is worth noting that our method in different experiments are exclusively trained on the training set of FakeAVCeleb, unless otherwise specified. In this paper, accuracy (Acc.), average precision (AP), and area under the ROC curve (AUC) are employed as evaluation metrics.

Implementation Details To better extract the audio-visual consistency feature, all the visual and audio data are firstly re-sampled to 25fps and 16,000Hz, respectively. According to the frame rate and the audio sampling rate, 16 continuous frames with corresponding 10240 audio samples are sampled as input. For convenience, ViT-B/32 CLIP and the base version of Whisper are employed as our foundation models. The number of the visual prompt tokens N_{vpt} is set to 1. The weights of the CMFM loss are as set to $\alpha = 2, \beta = 2, \gamma = 1$. For the training process, we randomly sample segments from each video and utilize 15 training epochs with the Adam (Kingma and Ba 2014) optimizer. The initial learning rate is set to 0.0001, with a reduction by a factor of 10 occurring at the 12th epoch. In the testing process, we sample continuous non-overlapping segments and compute the averaged segment predictions as the final result. Note that our method is trained on a single RTX 3080ti GPU with 25G CPU memory on Ubuntu 20.04.

Comparisons with the Existing Methods

Similar to (Feng, Chen, and Owens 2023; Oorloff et al. 2024; Yin et al. 2024), we evaluate the effectiveness and generalization ability of our method via intra-dataset evaluation, cross-manipulation evaluation, and cross-dataset evaluation. A variety of state-of-the-art multi-modal deepfake detection methods are selected, including MDS (Chugh et al. 2020), AV-DFD (Zhou and Lim 2021), BA-TFD (Cai et al. 2022), MMtrace (Raza and Malik 2023), AVAD (Feng, Chen, and Owens 2023), MRDF-CE (Zou et al. 2024), AVFF (Oorloff et al. 2024), and AVGraph (Yin et al. 2024). By following (Feng, Chen, and Owens 2023; Oorloff et al. 2024), we also select visual-related methods for cross-dataset evaluation, including Xception (Rossler et al. 2019), LipForensics (Haliassos et al. 2021), FTCN (Zheng et al. 2021) and RealForensics (Haliassos et al. 2022). In the tables, the modality employed by each method has been indicated, i.e., 'V' denotes the visual-related method and 'AV' represents the audio-visual multi-modal approach.

Intra-dataset Evaluation As shown in Tab. 1, our method achieves state-of-the-art performance in terms of both accuracy and AUC. Besides, benefiting from the prompt learning technique, our method only possesses 4.4M learnable parameters and the training only requires 9 hours on a single GPU. Meanwhile, the representation learning network of AVFF consists of two VideoMAE (Tong et al. 2022) architectures, which possess more than 174M learnable parameters. Similarly, MRDF-CE utilizes modified ResNet-18 as audio/visual encoders and 12 transformer blocks for feature fusion, and the overall number of learnable parameter numbers is more than 100M. Therefore, compared to the existing

Method	RVFA		FVRA-WL		FVFA-FS		FVFA-GAN		FVFA-WL		AVG-FV	
	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
AV-DFD	74.9	73.3	97.0	97.4	99.6	99.7	58.4	55.4	**100.**	**100.**	88.8	88.1
AVAD (LRS2)	62.4	71.6	93.6	93.7	95.3	95.8	94.1	94.3	93.8	94.1	94.2	94.5
AVAD (LRS3)	70.7	80.5	91.1	93.0	91.0	92.3	91.6	92.7	91.4	93.1	91.3	92.8
AVFF	93.3	92.4	94.8	98.2	**100.**	**100.**	99.9	**100.**	99.4	99.8	98.5	99.5
Ours	**97.1**	**95.5**	**99.9**	**99.9**	**100.**	**100.**	**100.**	**100.**	**100.**	**100.**	**99.9**	**99.9**

Table 2: Cross-manipulation results on FakeAVCeleb when testing on each forgery type after training on the remaining fours. The best results are highlighted in bold font, and the second-best results are underlined. The results of other methods are obtained from the paper of AVFF.

Method	Modality	AP	AUC
Xception	V	76.9	77.7
LipForensics	V	89.5	86.6
FTCN	V	66.8	68.1
RealForensics	V	95.7	93.6
AV-DFD	AV	79.6	82.1
AVAD	AV	87.6	86.9
AVFF	AV	93.1	**95.5**
Ours	AV	**96.6**	94.5

Table 3: Cross-dataset results on a subset of KoDF, after training on FakeAVCeleb. The best results are highlighted in bold, and the second-best results are underlined. The results of other methods are obtained from the paper of AVFF.

approaches, our proposed single-stage method can achieve state-of-the-art performance with significantly lower training cost.

Cross-manipulation Evaluation Here, we consider five manipulation categories in FakeAVCeleb by following (Feng, Chen, and Owens 2023; Oorloff et al. 2024): a) RVFA: real visual content with fake audio manipulated by SV2TTS; b) FVRA-WL: real audio with fake visual content manipulated by Wav2Lip; c) FVFA-FS: fake visual content manipulated by Faceswap and Wav2Lip, and fake audio manipulated by SV2TTS; d) FVFA-GAN: fake visual content manipulated by FSGAN and Wav2Lip, and fake audio manipulated by SV2TTS; e) FVFA-WL: fake visual content manipulated by Wav2Lip, and fake audio manipulated by SV2TTS. Benefiting from our proposed multi-modal prompts and fine-grained audio-visual consistency learning, Tab. 2 shows that our method outperforms other multi-modal methods when tested on the unseen category after training on the remaining four categories. Note that the performance of our method on RVFA outperforms the state-of-the-art method AVFF by 3.8% in AP and 3.1% in AUC, when training on the fake visual content.

Cross-dataset Evaluation We also investigate the generalization ability of the proposed method by conducting cross-dataset evaluation by following (Feng, Chen, and Owens 2023; Oorloff et al. 2024). The experimental results on a subset of KoDF (Kwon et al. 2021) are shown in Tab. 3. It can be observed that our method outperforms RealForensics by 0.9% in both AP and AUC. Note that our performance is on par with AVFF, which is a two-stage method with much more learnable parameters than our single-stage FM-based prompt learning method.

Ablation Study

Effect of Multi-modal Learning To demonstrate the importance of multi-modal learning, we train the audio branch and the visual branch separately. Besides, we report the performance without using $\mathrm{L_{CMFM}}$ to validate the efficacy of our frame-level cross-modal feature matching loss function. To reveal the uni-modal performances, we additionally introduce Acc.$_{\{v,a\}}$ and AUC$_{\{v,a\}}$. As shown in Tab. 4, though the uni-modal methods can achieve good performances in the corresponding modality, practical deepfake videos tend to possess multi-modal manipulations, which limits the applicability of uni-modal methods. In addition, it can be observed that the introduction of $\mathrm{L_{CMFM}}$ greatly improves the cross-dataset generalization ability, i.e., it brings 17.0% gains in terms of accuracy and 4.4% gains in term of AUC, which reveals that the natural alignments between audio and visual content are effective in mitigating the overfitting problem.

Effect of Multi-task Learning Our proposed method not only outputs the result of the input video, but also predicts the authenticity of each modality. Here, we explore the effect of the multi-task learning strategy in our method. As shown in Tab. 5, after discarding $\mathrm{L_a}$ or $\mathrm{L_v}$, the impact is not significant when the distributions of the training and testing sets are consistent, i.e., both the training and testing sets are from FakeAVCeleb. On the contrary, the performance drops sharply when the method is tested on the unseen dataset, i.e., KoDF. These experimental results demonstrate that the multi-task learning strategy in our method can obviously improve the generalization ability.

Effect of Sequential Visual Prompts To reveal the effectiveness of our sequential visual prompts, we replace our injection operation with the original prompt replacing operation in VPT (Jia et al. 2022). The 4th row in Tab. 5 shows that our proposed sequential visual prompts possess better generalization ability than the original approach in VPT, which

Audio Branch	Visual Branch	L_{CMFM}	FakeAVCeleb						KoDF	
			Acc.	AUC	Acc.$_v$	AUC$_v$	Acc.$_a$	AUC$_a$	Acc.	AUC
✓			55.2	79.6	-	-	99.4	99.9	-	-
	✓		97.6	98.5	99.8	99.9	-	-	61.5	93.6
✓	✓		99.8	99.9	99.8	99.9	99.7	100.	75.0	90.1
✓	✓	✓	**99.8**	**99.9**	**99.8**	**99.9**	**100.**	**100.**	**92.0**	**94.5**

Table 4: Performances of each branch and the effect of L_{CMFM} when conducting intra-dataset evaluation on FakeAVCeleb and cross-dataset evaluation on KoDF. The best results are highlighted in bold font.

Method	FakeAVCeleb		KoDF	
	Acc.	AUC	Acc.	AUC
Ours	99.8	**99.9**	**92.0**	**94.5**
w/o L_a	99.5	99.9	73.5	86.9
w/o L_v	99.8	99.9	77.5	90.3
VPT-replacing	99.8	99.9	86.5	90.2
segment-level loss	**99.9**	99.9	90.0	90.6

Table 5: Results on different losses and prompt adding operation. The best results are highlighted in bold font.

proves the superiority of our visual prompts.

Effect of Frame-level Cross-Modal Feature Matching Experimental results (the 5th row in Tab. 5) indicate that the introduction of the proposed frame-level loss function can bring better generalization ability compared to the segment-level loss function. Meanwhile, we also visualize the cosine distance of the frame-level audio-visual features of the FakeAVCeleb test samples, under the supervision of segment-level loss (Fig. 3 (a)) and frame-level loss (Fig. 3 (b)). It is obvious that the distance distribution depicted in Fig. 3 (b) exhibits greater orderliness, with real samples displaying smaller audio-visual feature distances compared to fake samples. In addition, in Fig. 3 (b), the distance between the real data distribution and the distributions of different types of deepfakes meets our expectations. Specifically, due to the decent performance of Wav2Lip, the lip movement and audio of the FVFA data are more synchronized than others, which results in smaller audio-visual feature distances than the other fake data. More than half of the FVRA data is manipulated by face swapping methods, i.e., FaceSwap and FSGAN, which swap identities while maintaining the facial expression. However, they have limitations in fine-grained expression generation, which tends to induce inconsistencies between lip movement and corresponding audio. Thus, the audio-visual feature distance of FVRA manipulated data is usually larger than FVFA. Since the RVFA data is not lip-synchronized in FakeAVCeleb, its corresponding audio-visual feature distances are usually the largest. Based on the above analysis, our frame-level cross-modal feature matching loss function is effective for fine-grained multi-modal feature matching.

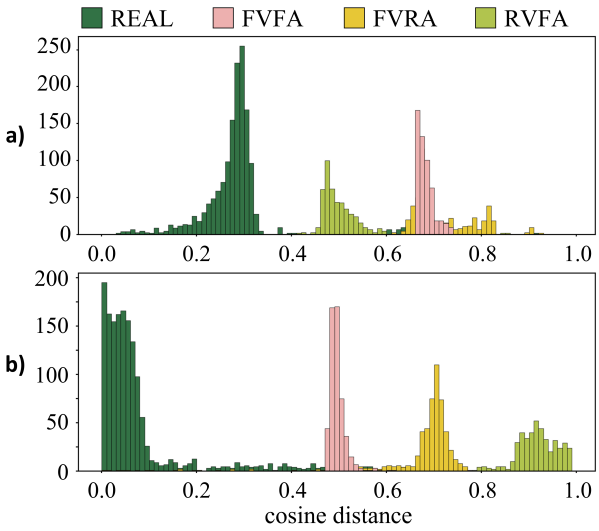

Figure 3: The visualizations of the cosine distances between the frame-level audio-visual features of the FakeAVCeleb test samples under the supervision of (a) segment-level and (b) frame-level cross-modal feature matching loss.

Conclusion

In this paper, we propose the first multi-task audio-visual prompt learning method for multi-modal deepfake video detection, by exploiting multiple foundation models. Based on the characteristics of input data, we propose novel sequential visual prompts and short-time audio prompts, which consider both the intra-modality artifacts and inter-modality consistencies. To utilize the natural alignment of the visual content and the audio signal and learn fine-grained audio-visual consistency embeddings, we propose a frame-level cross-modal feature matching loss function. Extensive experiments demonstrate the effectiveness and superior generalization ability of our method.

In this work, we utilized CLIP and Whisper as the visual and audio foundation models. In the future, we will further explore to apply the proposed approach to other foundation models, and solve more complicated problems such as deep forgery localizations.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62272020, U20B2069 and 62406015, in part by the State Key Laboratory of Complex & Critical Software Environment under Grant SKLSDE2023ZX-16, and in part by the Fundamental Research Funds for Central Universities.

References

Agarwal, S.; Farid, H.; Fried, O.; and Agrawala, M. 2020. Detecting deep-fake videos from phoneme-viseme mismatches. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 660–661.

Cai, Z.; Stefanov, K.; Dhall, A.; and Hayat, M. 2022. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *International Conference on Digital Image Computing: Techniques and Applications*, 1–10.

Cheng, J.; Liang, D.; and Tan, S. 2024. Transfer CLIP for Generalizable Image Denoising. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25974–25984.

Choi, J.; Kim, T.; Jeong, Y.; Baek, S.; and Choi, J. 2024. Exploiting Style Latent Flows for Generalizing Deepfake Video Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1133–1143.

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

Feng, C.; Chen, Z.; and Owens, A. 2023. Self-supervised video forensics by audio-visual anomaly detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10491–10503.

Gu, Z.; Chen, Y.; Yao, T.; Ding, S.; Li, J.; Huang, F.; and Ma, L. 2021. Spatiotemporal inconsistency learning for deepfake video detection. In *ACM International Conference on Multimedia*, 3473–3481.

Gu, Z.; Chen, Y.; Yao, T.; Ding, S.; Li, J.; and Ma, L. 2022. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *AAAI Conference on Artificial Intelligence*, volume 36, 744–752.

Haliassos, A.; Mira, R.; Petridis, S.; and Pantic, M. 2022. Leveraging real talking faces via self-supervision for robust forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14950–14962.

Haliassos, A.; Vougioukas, K.; Petridis, S.; and Pantic, M. 2021. Lips don't lie: A generalisable and robust approach to face forgery detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5039–5049.

Ilyas, H.; Javed, A.; and Malik, K. M. 2023. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing*, 136: 110124.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European Conference on Computer Vision*, 709–727.

Jia, Y.; Zhang, Y.; Weiss, R.; Wang, Q.; Shen, J.; Ren, F.; Nguyen, P.; Pang, R.; Lopez Moreno, I.; Wu, Y.; et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in Neural Information Processing Systems*, 31.

Khalid, H.; Kim, M.; Tariq, S.; and Woo, S. S. 2021a. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In *The 1st workshop on synthetic multimedia-audiovisual deepfake generation and detection*, 7–15.

Khalid, H.; Tariq, S.; Kim, M.; and Woo, S. S. 2021b. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*.

Khan, S. A.; and Dai, H. 2021. Video transformer for deepfake detection with incremental learning. In *ACM International Conference on Multimedia*, 1821–1828.

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

Kwon, P.; You, J.; Nam, G.; Park, S.; and Chae, G. 2021. Kodf: A large-scale korean deepfake detection dataset. In *IEEE/CVF International Conference on Computer Vision*, 10744–10753.

Li, X. L.; and Liang, P. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Li, Y.; Chang, M.-C.; and Lyu, S. 2018. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *IEEE International Workshop on Information Forensics and Security*, 1–7.

Lin, L.; Papabathini, S.; Wang, X.; and Hu, S. 2024. Robust Light-Weight Facial Affective Behavior Recognition with CLIP. *arXiv preprint arXiv:2403.09915*.

Liu, Y.; Liu, X.; Zhao, Y.; Wang, Y.; Xia, R.; Tain, P.; and Wang, Y. 2024. Audio Prompt Tuning for Universal Sound Separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 1446–1450.

Mittal, T.; Bhattacharya, U.; Chandra, R.; Bera, A.; and Manocha, D. 2020. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *ACM International Conference on Multimedia*, 2823–2832.

Oorloff, T.; Koppisetti, S.; Bonettini, N.; Solanki, D.; Colman, B.; Yacoob, Y.; Shahriyari, A.; and Bharaj, G. 2024. AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27102–27112.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 28492–28518.

Raza, M. A.; and Malik, K. M. 2023. Multimodaltrace: Deepfake detection using audiovisual representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop*, 993–1000.

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

Sabir, E.; Cheng, J.; Jaiswal, A.; AbdAlmageed, W.; Masi, I.; and Natarajan, P. 2019. Recurrent convolutional strategies for face manipulation detection in videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 80–87.

Salvi, D.; Liu, H.; Mandelli, S.; Bestagini, P.; Zhou, W.; Zhang, W.; and Tubaro, S. 2023. A robust approach to multimodal deepfake detection. *Journal of Imaging*, 9(6): 122.

Shiohara, K.; and Yamasaki, T. 2022. Detecting deepfakes with self-blended images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18720–18729.

Tan, L.; Wang, Y.; Wang, J.; Yang, L.; Chen, X.; and Guo, Y. 2023. Deepfake video detection via facial action dependencies estimation. In *AAAI Conference on Artificial Intelligence*, volume 37, 5276–5284.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35: 10078–10093.

Vidit, V.; Engilberge, M.; and Salzmann, M. 2023. Clip the gap: A single domain generalization approach for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3219–3229.

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

Xu, Y.; Liang, J.; Jia, G.; Yang, Z.; Zhang, Y.; and He, R. 2023. TALL: Thumbnail Layout for Deepfake Video Detection. In *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*, 1–18.

Yang, W.; Zhou, X.; Chen, Z.; Guo, B.; Ba, Z.; Xia, Z.; Cao, X.; and Ren, K. 2023. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18: 2015–2029.

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

Yin, Q.; Lu, W.; Cao, X.; Luo, X.; Zhou, Y.; and Huang, J. 2024. Fine-Grained Multimodal DeepFake Classification via Heterogeneous Graphs. *International Journal of Computer Vision*, 1–15.

Zhang, B.; Li, S.; Feng, G.; Qian, Z.; and Zhang, X. 2022. Patch Diffusion: A General Module for Face Manipulation Detection. In *AAAI Conference on Artificial Intelligence*, volume 36, 3243–3251.

Zhang, B.; Yu, S.; Wei, Y.; Zhao, Y.; and Xiao, J. 2024. Frozen CLIP: A Strong Backbone for Weakly Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3796–3806.

Zhao, Z.; and Patras, I. 2023. Prompting Visual-Language Models for Dynamic Facial Expression Recognition. In *British Machine Vision Conference*, 1–14.

Zheng, Y.; Bao, J.; Chen, D.; Zeng, M.; and Wen, F. 2021. Exploring temporal coherence for more general video face forgery detection. In *IEEE/CVF International Conference on Computer Vision*, 15044–15054.

Zhou, C.; Zhong, F.; and Öztireli, C. 2023. CLIP-PAE: projection-augmentation embedding to extract relevant features for a disentangled, interpretable and controllable text-guided face manipulation. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–9.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16816–16825.

Zhou, Y.; and Lim, S.-N. 2021. Joint audio-visual deepfake detection. In *IEEE/CVF International Conference on Computer Vision*, 14800–14809.

Zou, H.; Shen, M.; Hu, Y.; Chen, C.; Chng, E. S.; and Rajan, D. 2024. Cross-Modality and Within-Modality Regularization for Audio-Visual Deepfake Detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4900–4904.