
CAD: A General Multimodal Framework for Video Deepfake Detection via Cross-Modal Alignment and Distillation

Yuxuan Du^{1*}, Zhendong Wang^{2*}, Yuhao Luo^{3*}, Caiyong Piao^{3*},
Zhiyuan Yan¹, Hao Li¹, Li Yuan^{1†}

* Equal Contributors, † Corresponding Authors

¹ Peking University, Shenzhen Graduate School,

² University of Science and Technology of China,

³ Chinese University of Hong Kong, Shenzhen

dyx2290246176@gmail.com, yuanli-ece@pku.edu.cn

Abstract

The rapid emergence of multimodal deepfakes (visual and auditory content are manipulated in concert) undermines the reliability of existing detectors that rely solely on modality-specific artifacts or cross-modal inconsistencies. In this work, we first demonstrate that modality-specific forensic traces (e.g., face-swap artifacts or spectral distortions) and modality-shared semantic misalignments (e.g., lip-speech asynchrony) offer complementary evidence, and that neglecting either aspect limits detection performance. Existing approaches either naively fuse modality-specific features without reconciling their conflicting characteristics or focus predominantly on semantic misalignment at the expense of modality-specific fine-grained artifact cues. To address these shortcomings, we propose a general multimodal framework for video deepfake detection via **Cross-Modal Alignment and Distillation** (CAD). CAD comprises two core components: 1) Cross-modal alignment that identifies inconsistencies in high-level semantic synchronization (e.g., lip-speech mismatches); 2) Cross-modal distillation that mitigates feature conflicts during fusion while preserving modality-specific forensic traces (e.g., spectral distortions in synthetic audio). Extensive experiments on both multimodal and unimodal (e.g., image-only/video-only) deepfake benchmarks demonstrate that CAD significantly outperforms previous methods, validating the necessity of harmonious integration of multimodal complementary information.

1 Introduction

The rapid advancement of deepfake generation has led to highly sophisticated multimodal forgeries that seamlessly integrate manipulated visual, audio, and textual elements to produce realistic and deceptive content [41, 59, 47]. Modern multimodal generation techniques increasingly exploit cross-modal coherence, such as syncing synthetic voices with lip movements [42, 72], generating contextually plausible scripts [40, 60], or blending facial reenactments with cloned speech [29, 16], thereby making the **detection of such multimodal forgeries increasingly challenging** for existing approaches. Existing unimodal detection methods (e.g., facial blending anomaly detectors [48, 11, 62] or spectral artifact analysis in audio [65, 31]) mainly rely on isolated artifacts in a single modality, *making them inherently blind to cross-modal discrepancies and failing to identify cross-modal manipulations* like synchronized lip movements with AI-generated voices.

Generally, the multimodal forgeries essentially expose detection traces from two key perspectives. 1) **Modality-Specific Inconsistencies**: These include artifacts confined to a single modality, such as unnatural facial texture blending in manipulated videos ([30, 48]) or spectral distortions in synthesized audio ([2, 36]). 2) **Modality-Shared Semantic Misalignments**: These involve inconsistencies between modalities, such as mismatches between lip movements and speech ([20, 21]), or incongruent emotional tones across audio and visual streams ([39]). However, we find that **most existing multimodal detectors fail to consider how to fully utilize both modality-specific and shared forgeries for comprehensive detection**. To illustrate, most multimodal frameworks (such as [33, 39, 63]) primarily focus on semantic coherence (*e.g.*, lip-speech alignment) but often overlook modality-specific forensic traces, such as unnatural texture blending or flickering in synthesized faces and subtle acoustic artifacts in synthetic speech. Additionally, approaches like [67] directly combine modality-specific cues without effectively leveraging modality-shared information, *missing the opportunity to exploit cross-modal correlations that could strengthen forgery detection*.

By revisiting the existing detection works, we argue that *a truly effective deepfake detection system must integrate both modality-specific and shared perspectives, ensuring that neither local artifacts nor cross-modal mismatches escape detection*. By leveraging the full multimodal space rather than treating modalities independently, we can build a stronger defense against increasingly sophisticated fakes. As illustrated in Figure 1, we divide the traces of multimodality deepfake detection into modality-specific and modality-shared inconsistencies. *The visual domain* (x_1) contains artifacts like blending boundaries or unrealistic motion, while the *audio domain* (x_2) may exhibit synthetic spectral anomalies. More critically, their *joint representation* (x_{12}) captures discrepancies that neither modality alone can reveal, such as a deepfake that synchronizes well but exhibits unnatural articulation patterns or mismatched speech emotion.

To achieve this, we propose **CAD**, a novel multimodal detection framework that unifies modality-shared semantic alignment analysis and modality-specific forensic verification. CAD operates on the principle that *multimodal deepfakes inevitably leave traces in two forms: (1) semantic misalignments between modalities* (*e.g.*, discrepancies between spoken words and lip movements) and **(2) modality-specific inconsistencies** (*e.g.*, residual artifacts from facial swapping or synthetic voice generation). Unlike prior works that prioritize one aspect at the expense of the other, *CAD introduces a dual-path architecture: a cross-modal alignment module to detect semantic disharmony and a distillation mechanism to preserve and harmonize modality-specific forensic signals*. By synergizing these complementary cues, CAD closes the detection gap left by conventional approaches, offering a unified defense against the escalating threat of multimodal synthetic media. Evaluated on the comprehensive IDForge dataset, which spans diverse forgery types, CAD achieves state-of-the-art performance (99.96% AUC), demonstrating its robustness against evolving multimodal attacks.

Our main contributions are summarized as three-fold:

- **Unified Detection of Modality-Shared and Modality-Specific Cues**: We propose CAD, a novel detection framework that integrates both modality-shared semantic alignment analysis (*e.g.*, lip-speech mismatches) and modality-specific forensic verification (*e.g.*, residual artifacts in face-swapped videos or synthetic audio).
- **Dual-Path Architecture Design**: We introduce a cross-modal alignment module to detect semantic disharmony across different modalities, and develop a cross-modal distillation mechanism that preserves forensic artifacts within individual modalities while mitigating conflicts during multimodal feature fusion.

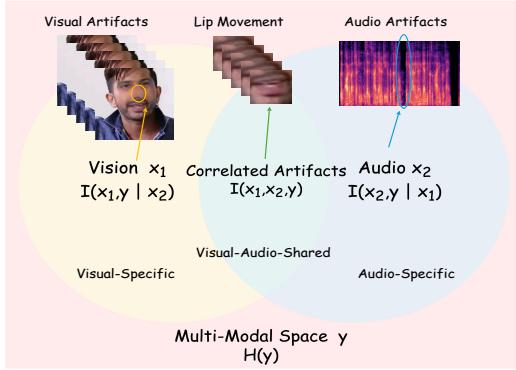


Figure 1: Venn diagram illustrating different types of artifacts in multimodal deepfakes, categorized into *modality-specific* and *modality-shared* cues. Specifically, visual artifacts (x_1) may include blending boundaries of face-swapping, while audio artifacts (x_2) might exhibit spectral anomalies. The shared space (x_{12}) captures semantic cross-modal mismatches, such as inconsistency between lip movements and speech. Ideally, a robust system should integrate both perspectives for improved accuracy.

- **SOTA Performance on Multimodal Deepfake Detection:** We evaluate CAD on the IDForge dataset, which covers a wide range of deepfake manipulation techniques, achieving 99.96% AUC, significantly outperforming existing unimodal and multimodal detection methods, validating the necessity of a holistic approach to multimodal deepfake detection.

2 Related Works

2.1 Deepfake Detection

Unimodal Deepfake Detection. Most existing deepfake video detection methods focus on identifying visual artifacts in image modality [49, 61, 48, 30, 58] and video modality [21, 53, 57, 66, 62]. Notably, early works, such as [21], leveraged lip-reading pre-trained models to detect inconsistencies in lip movements, while deepfake image detection initially focused on blending boundaries [30], a common artifact in face-swapping forgeries. Later, [48, 11] further explored blending-based detection, refining the objective from cross-ID to within-ID forgery analysis [48] and examining the role of blended data in detection [11]. Meanwhile, video-based detection evolved to balance spatial and temporal artifact learning [51, 62], mitigating overfitting to a single modality (*e.g.*, spatial artifacts only) and improving robustness across different forgery types.

Multimodal Deepfake Detection. Recent methods leverage multimodal information for deepfake detection. RealForensics [20] learns video and audio features crosswise, updating student model parameters and training classifiers. [13] and [14] use both cross-entropy and contrastive learning losses to measure similarity between real samples, aiding joint feature learning. [33] focuses on the connection between audio and lip features. Building on MAE [22], [39] masks video and audio embeddings, using additional decoders and reconstruction loss for joint learning. [50] introduces FDMT, utilizing 3D facial information and depth-based attention to improve deepfake detection. However, we find that most previous detection works still fail to consider how to maximize or fully leverage both modality-specific and shared forgeries for detection, leaving vulnerabilities that sophisticated attacks can exploit.

2.2 Multimodal Representation Learning

With the advent of multimodal models, CLIP [44] leverages contrastive learning [7] to align semantic information across modalities by learning cross-modal embedding representations of images and text. Building upon this approach, ImageBind [19] and LanguageBind [74] extend contrastive learning [7] to additional modalities, including vision, audio, text, depth, and more. These developments further enhance the capacity of encoders to extract features from diverse modalities, facilitating the learning of richer and more comprehensive representations.

Vision and audio are key modalities in deepfake detection, with audio excelling in action recognition in long videos. Knowledge transfer between modalities has been explored for action recognition [9], and the BAVNet [55] localizes sound within visual scenes. AVoID-DF [63] introduces a multimodal decoder for joint feature fusion in deepfake detection. Depth modality also aids face feature extraction. Depth maps distinguish real and fake faces [52], with PRNet [18] and contrastive depth loss for feature learning. Facial depth maps are used for anti-spoofing [69], with symmetry loss for accurate depth estimation, which is beneficial for forgery detection [50].

3 Method

3.1 Overview

Multimodal deepfake detection can be addressed from **two complementary perspectives**: (1) high-level semantic cues, such as facial structure and expression consistency, and (2) low-level forensic artifacts, such as subtle pixel-level distortions. Semantic inconsistencies often manifest in facial regions involved in expression and speech (*e.g.*, unnatural lip movements), while artifact-based traces typically appear in fine details like the nose or eye contours.

As shown in Figure 2, our proposed **CAD framework** moves beyond traditional approaches that focus on a single aspect of forgery detection. Instead, we unify two key components within one architecture: **modality-shared semantic alignment** and **modality-specific forensic verification**.

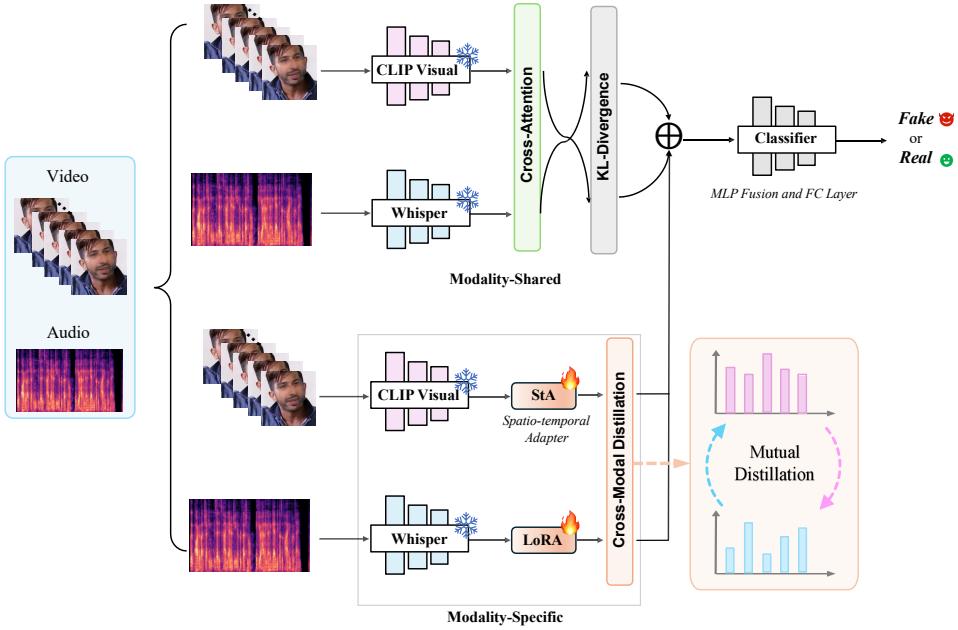


Figure 2: The overview of CAD. Our proposed CAD is designed to maximize and fully mine both modality-specific and modality-shared cues for robust deepfake detection.

This design captures both the high-level semantic alignment between audio and visual modalities and the low-level, modality-specific artifacts. In Section 3.2, we provide a theoretical foundation for this dual-perspective design, based on maximizing mutual information between modalities. Sections 3.3 and 3.4 then detail the implementation of the forensic and semantic modules, each tailored to extract complementary evidence from different modalities for robust deepfake detection.

3.2 Maximizing Multi-modality *Mutual Information* in Multimodal Deepfake Detection

Multimodal deepfake detection should go beyond analyzing individual modalities in isolation. Instead, it should explicitly model potential inconsistencies between the visual and audio streams. As illustrated in Figure 1, different deepfake techniques tend to introduce modality-specific artifacts, leading to subtle mismatches between what is seen and what is heard. To capture these cross-modal inconsistencies, we leverage both joint and disentangled feature representations of vision (V) and audio (A), as formalized in Equation 1 below:

$$H(V) = H(V | A) + I(V, A). \quad (1)$$

Following the principles of information theory [26], this formulation incorporates entropy ($H(X)$) and mutual information ($I(X, Y)$) to guide the model in learning representations that maximize shared information across modalities while preserving modality-specific signals.

Specifically, we consider the distributions of video and audio to be x_1 and x_2 , respectively. We also introduce y , a latent space that encapsulates both x_1 and x_2 , ensuring full compatibility with video and audio representations. Based on the definition of mutual information [26], we can derive the following lemma below.

Lemma 1. *The mutual information between two random variables x_1 , x_2 and a universal space distribution y can be expressed in terms of entropy as:*

$$\begin{aligned} I(x_1, x_2, y) &= H(x_1) + H(x_2) + H(y) - H(x_1, y) \\ &\quad - H(x_2, y) - H(x_1, x_2) + H(x_1, x_2, y). \end{aligned} \quad (2)$$

Lemma 2. Since we define a total space distribution y that integrates both modalities, ensuring seamless compatibility with their respective representations, it follows that $x_1 \subset y$ and $x_2 \subset y$. Consequently, the following equation is established:

$$H(x_1, x_2, y) = H(x_1, y) = H(x_2, y) = H(y). \quad (3)$$

Substitute into formula (1):

$$I(x_1, x_2, y) = H(x_1) + H(x_2) - H(x_1, x_2). \quad (4)$$

When we perform a single-modal task, we can consider maximizing its mutual information part, $I(x_1, y)$, which is the visual channel goal. According to the definition of mutual information, we have the following Lemma.

Lemma 3. Maximizing the mutual information between the input x_1 and the target variable y can be formulated as optimizing the information flow within a specific modality, which represents the objective for the visual channel.

$$I(x_1, y) = I(x_1, x_2, y) + H(x_1, y | x_2). \quad (5)$$

Previous studies have extensively demonstrated the benefits of joint learning for multimodal feature representations. However, these approaches primarily focus on the shared joint distribution characteristics across different modalities—specifically, $I(x_1, x_2, y)$ in Equation 4, while neglecting the modality-specific components $H(x_1, y | x_2)$ that do not align with the multimodal representation. In the context of audio-visual feature learning, a more detailed analysis can be conducted. Let s denote a single sample, while v and a represent the extracted video and audio features, respectively. Furthermore, let V and A denote the distributions of video and audio features across all samples. Based on this formulation, we derive the following theorem.

Theorem 1. For the representation learning component, when the input is encoded into an embedding vector by the encoder, the mutual information between the video and audio embedding representations of a single sample and the overall distribution is formulated as:

$$I(v, a | V, A) = E_{p(v)}[H(p(v | V))] - E_{p(s, V, A, a)}[H(p(v | V), p(v | a, V, A))]. \quad (6)$$

Proof. We can derive as follows:

$$\begin{aligned} & I(v, a | V, A) \\ &= \int p(V, A) \int [p(v, a | V, A) \log \frac{p(v, a | V, A)}{p(v | V, A)p(a | V, A)}] dv da dV dA \\ &= \int p(V, A) \int [p(v, a | V, A) \log \frac{p(v | a, V, A)}{p(v | V, A)}] dv da dV dA \\ &= \int p(V, A) \int [p(v, a | s, V, A) p(s | V, A) \log \frac{p(v | a, V, A)}{p(v | V, A)}] dv da dV dA \\ &= \int p(s, V, A) \int [p(v | V, s) p(a | A, s) \log \frac{p(v | a, V, A)}{p(v | V, A)}] dv da dV dA \\ &= E_{p(s, V, A, a)} \left[\int p(v | V, s) \log p(v | a, V, A) dv \right] - E_{p(s, V, A, a, v)} [\log p(v | V, A)] \\ &= -E_{p(s, V, A, a)} [H(p(v | V), p(v | a, V, A))] + E_{p(v)} [H(p(v | V))]. \end{aligned}$$

□

For the first term in Equation 6, the target representation must exhibit high entropy to prevent collapse and ensure adaptability to downstream tasks. The second term necessitates that the distribution of v closely approximates that of a , thereby minimizing the distribution distance of visual and audio modalities. Notably, the second term corresponds to $I(x_1, x_2, y)$ in Equation 5, which explains the effectiveness of joint learning in multimodal feature representation. In contrast, the first term

aligns with traditional unimodal video authentication, as it excludes the objective of maximizing the expected joint information with audio.

Inspired by the derivation, CAD comprehensively accounts for the influence of both components, enabling the modeling of multimodal interactions within the framework of multimodal representation learning, capturing both modality-specific characteristics and collaborative dependencies.

3.3 Cross-Modal Alignment for Modality-Specific Forgery Learning

In Equation 5, for $I(x_1, x_2, y)$, highlights the need for learning modality-shared representations. Previous research has thoroughly validated the enhancement of feature representation through the joint learning of multi-modality. Similarly, to maximize the $I(x_1, x_2, y)$ term in Equation 4, CAD effectively captures the shared information embeddings across different modalities by aligning the feature representations of audio-visual signals.

As illustrated in the joint representation in Figure 2, we employ feature alignment using CLIP [44] and Whisper [45] as the frozen pre-trained encoders for video and audio. After acquiring high-level semantic representations via the pre-trained model, we extract their high-level semantic features using a cross-attention mechanism, and subsequently minimize the distributional discrepancy between modalities via Kullback–Leibler (KL) divergence. Such efficient operations enable the modality-shared component to learn common semantic features across modalities, such as the correlation between lip movements and vocal signals. Let x_{v2a} and x_{a2v} represent the outputs after the cross-attention mechanism. And $P(x), Q(x)$ represent the potential distributions of $v2a$ and $a2v$ respectively. We compute the modality-specific alignment loss as follows:

$$\begin{aligned} P(x) &= \text{Softmax}(x_{v2a}), \\ Q(x) &= \text{Softmax}(x_{a2v}), \\ L_{\text{KL}} &= \sum_x P(x) \log \frac{P(x)}{Q(x)}. \end{aligned} \tag{7}$$

3.4 Cross-Modal Distillation for Modality-Shared Forgery Learning

In Equation 5, for $H(x_1, y | x_2)$, aligns with the video modality-specific features. During the process of video feature extraction, it is crucial to consider not only the spatial representation of static frames but also to effectively capture the temporal dynamics between frames to obtain comprehensive spatiotemporal features. We draw inspiration from [62] in the encoder design to optimize the spatiotemporal feature aggregation strategy. This ensures that the model can effectively integrate the local detail information of individual video frames with the global temporal patterns, thereby enhancing the overall performance of video understanding tasks. We employ the audio pre-trained frozen model with Low-Rank Adaptation (LoRA) [23], allowing efficient fine-tuning through a trainable low-rank decomposition matrix. This approach enables effective adaptation to task-specific requirements with minimal computational overhead.

As illustrated in the specific representation in Figure 2, to leverage the audio modality to enhance video representation learning and improve cross-modal consistency, we apply distillation between the two unimodal channels. As the distillation loss, we introduce the SimSiam [8] loss, which offers two key advantages. First, SimSiam mitigates the introduction of erroneous negative samples by directly comparing representations, thereby addressing potential distributional discrepancies between different modalities. Second, it prevents feature collapse by ensuring that representations from different modalities do not converge into identical vectors but instead retain modality-specific information, preserving the unique characteristics of each unimodal representation. Denote x_v^u and x_a^u as the unimodal embeddings for vision and audio, respectively, and z_v^u and z_a^u as the SimSiam projections for vision and audio. \odot represents element-wise multiplication and Norm represents L2 normalization. We compute the modality-shared cross-modal distillation loss as:

$$\begin{aligned} \text{Dist}_1 &= -\text{Norm}(x_v^u) \odot \text{Norm}(z_a^u), \\ \text{Dist}_2 &= -\text{Norm}(x_a^u) \odot \text{Norm}(z_v^u), \\ L_{\text{KD}} &= (\text{Dist}_1 + \text{Dist}_2)/2. \end{aligned} \tag{8}$$

4 Experiments

4.1 Implementation Settings

Dataset. To comprehensively assess the performance of CAD from multiple perspectives, we conducted experiments on a diverse range of audio-visual and video-only deepfake datasets. Specifically, we utilized the following datasets: (1) **FakeAVCeleb** [28]. This dataset comprises both video and audio deepfakes, featuring precisely synchronized lip movements and fine-grained annotations. (2) **IDForge-v2** [56]. IDForge-v2 improves upon IDForge-v1 by adding compression and super-resolution to better simulate real-world conditions. Each video retains 16 frames in four evenly spaced groups, along with corresponding audio and text. (3) **FaceShifter** [46] and **Celeb-DF** [32]. FaceShifter and Celeb-DF consist exclusively of visual deepfakes and enhance various forgery approaches to generate synthetic videos.

Training details. During training, 16 frames per sample are used without truncation to align with Whisper’s 30-second audio feature extraction, preserving temporal continuity. For datasets with extensive backgrounds, facial landmarks are detected and cropped beforehand. We trained on 8×NVIDIA-H100 for 4 hours.

Model details. For the pre-trained models, we use CLIP ViT-Base-16 [44] and Whisper-Small [45] to balance prior knowledge and computational efficiency. Videos are preprocessed via the CLIP processor, resizing to 224×224 and normalizing RGB channels. The LoRA [24] decomposition in the audio encoder defaults to $r = 8$ and $\alpha = 16$ unless specified.

4.2 Comparison with Existing Methods

We evaluate our model’s performance against state-of-the-art algorithms using multiple criteria, including average accuracy (Acc), Average Precision (AP), and Area Under the Curve (AUC), across diverse datasets. For audio-visual algorithms, a video is classified as fake if either modality, or both, is manipulated. Unimodal models consider a video fake only when the visual modality is forged.

Intra-manipulation evaluations. We follow the methodology outlined in AVoID-DF [63], randomly selecting 70% of FakeAVCeleb [28] as the training set for model training, while the remaining 30% serves as the evaluation set. As shown in Table 1, our method demonstrates significant improvements over two audio-visual deepfake detection methods, AVoID-DF [63] and AVFF [39]. Table 2 further presents a comparative analysis between CAD and multiple open-source state-of-the-art forgery detection methods on the IDForge dataset [56], encompassing both unimodal and multimodal approaches.

For all models except VFD, additional classification heads are incorporated during experiments to facilitate classification. The results indicate that CAD outperforms RealForensics significantly,

Table 1: **Intra-manipulation evaluation** on **FakeAVCeleb** [28]. Following [63, 39], select 70% as the training set and the remaining 30% as the test set. We report accuracy (Acc) and Area Under the Curve (AUC) scores (%).

Method	Modality	Acc	AUC
MesoNet [1]	V	57.3	60.9
Capsule [37]	V	68.8	70.9
Head Pose [64]	V	45.6	49.2
VA-MLP [34]	V	65.0	67.1
Xception [12]	V	67.9	70.5
LipForensics [21]	V	80.2	82.4
DeFakeHop [5]	V	68.3	71.6
CViT [54]	V	70.5	72.1
Multiple-Attention [68]	V	77.6	79.3
SLADD [6]	V	70.5	72.1
AVN-J [43]	A-V	73.2	77.6
MDS [13]	A-V	82.8	86.5
Emotion Don’t Lie [35]	A-V	78.1	79.8
AVFakeNet [25]	A-V	78.4	83.4
VFD [10]	A-V	81.5	86.1
BA-TFD [3]	A-V	80.8	84.9
AVoID-DF [63]	A-V	83.7	89.2
AVFF [39]	A-V	98.6	99.1
CAD (Ours)	A-V	99.0	99.6

Table 2: **Intra-manipulation evaluation on IDForge** [56]. We conducted experiments based on the training and test sets divided by IDForge-v2. IDForge contains 11 different forgery subsets and we conduct multiple tests and compute the average for robustness. We report Area Under the Curve (AUC) scores (%) and Average Precision (AP) scores (%).

Method	Modality	AUC	AP
MesoI4 [1]	V	74.72	38.39
I3D [4]	V	77.74	45.56
CADMM [15]	V	80.35	59.92
UFD [38]	V	89.12	70.00
RealForensics [20]	V	93.21	88.18
CDCN [27]	A-V	87.43	71.17
VFD [10]	A-V	90.70	-
IDForge [56]	A-V	98.12	86.67
CAD (Ours)	A-V	99.96	99.63

Table 3: **Cross-manipulation evaluation** on **FakeAVCeleb** [28]. The performance is evaluated using a leave-one-category-out approach, where one category is reserved for testing while the remaining categories are used for training. We report Average Precision (AP) scores (%) and Area Under the Curve (AUC) scores (%). The average performance category is provided in AVG-FV.

Method	Modality	RVFA		FVRA-WL		FVFA-FS		FVFA-GAN		FVFA-WL		AVG-FV	
		AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
Xception [46]	V	-	-	88.2	88.3	92.3	93.5	67.6	68.5	91.0	91.0	84.8	85.3
LipForensics [21]	V	-	-	97.8	97.7	99.9	99.9	61.5	68.1	98.6	98.7	89.4	91.1
FTCN [70]	V	-	-	96.2	97.4	100	100	77.4	78.3	95.6	96.5	92.3	93.1
RealForensics [20]	V	-	-	88.8	93.0	99.3	99.1	99.8	99.8	93.4	96.7	95.3	97.1
AV-DFD [73]	A-V	74.9	73.3	97.0	97.4	99.6	99.7	58.4	55.4	100	100	88.8	88.1
AVAD(LRS2) [17]	A-V	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) [17]	A-V	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 [39]	A-V	93.3	92.4	94.8	98.2	100	100	99.9	100	99.4	99.8	98.5	99.5
CAD (Ours)	A-V	99.9	99.9	100									

with a 6.75% increase in AUC and an 11.45% improvement in AP. Additionally, compared to the R-MFDN network proposed in IDForge, CAD demonstrates notable superiority, achieving a 12.96% increase in AP.

Cross-manipulation evaluations. The cross-manipulation evaluation examines model performance on a dataset with an unknown distribution in Table 3. emphasizing the necessity for deepfake detection algorithms to exhibit robust generalization capabilities against previously unseen forgery techniques to ensure adaptability across diverse scenarios. To systematically evaluate this aspect, we categorize subsets of FakeAVCeleb [28] based on different forgery techniques, including RVFA,

To assess the generalization ability of our model, we conduct cross-category evaluation on four distinct forgery types: FVRA-WL, FVFA-FS, FVFA-GAN, and FVFA-WL, each representing a unique data generation paradigm. In each experiment, one forgery category is excluded during training and used exclusively for testing, while the model is trained on the remaining categories. As shown in Table 3, CAD exhibits strong generalization capabilities, consistently outperforming baseline methods when facing unknown forgery types. This robustness highlights the effectiveness of our joint modeling of semantic alignment and modality-specific forensic cues. Furthermore, results across Table 1 and Table 2 reinforce a clear trend: multimodal deepfake detection methods consistently surpass unimodal approaches in both accuracy and robustness.

4.3 Ablation Study

To investigate the rationality of the CAD architecture for video detection, we conducted a series of self-module ablation experiments. As presented in Table 4 and Table 5, the model was trained on IdForge [56] and evaluated on FaceShifter of FaceForensics++ [46] and CelebDF [32], which both lack audio channels. The results in Table 4 and Table 5 indicate that the performance of the model utiliz-

Table 4: Ablation experiments of the proposed **cross-modal alignment** with training on IDForge-v2, we report Area Under the Curve (AUC) scores (%) and Average Precision (AP) scores(%) and test on FaceForensics++ and Celeb-DF.

Method	FaceShifter		Celeb-DF		Avg	
	AUC	AP	AUC	AP	AUC	AP
only video	100	100	74.9	80.5	87.5	90.3
w/o alignment	98.6	98.8	93.5	96	96.1	97.4
w/o cross attention	95.7	95.9	85.1	90.1	90.4	93
w/o frozen model	99.9	99.9	90.8	95.0	95.4	97.5
CAD (Ours)	99.6	99.6	94.2	96.7	96.9	98.2

Table 5: Ablation experiments of the proposed **cross-modal distillation** with training on IDForge-v2, we report Area Under the Curve (AUC) scores (%) and Average Precision (AP) scores(%) and test on FaceForensics++ and Celeb-DF.

Method	FaceShifter		Celeb-DF		Avg	
	AUC	AP	AUC	AP	AUC	AP
only video	100	100	74.9	80.5	87.5	90.3
w/o distillation	97.8	97.5	87.9	91.6	92.9	94.6
Substitute SimSiam with KL	99.2	99.6	93.2	95.8	96.2	97.7
CAD (Ours)	99.6	99.6	94.2	96.7	96.9	98.2

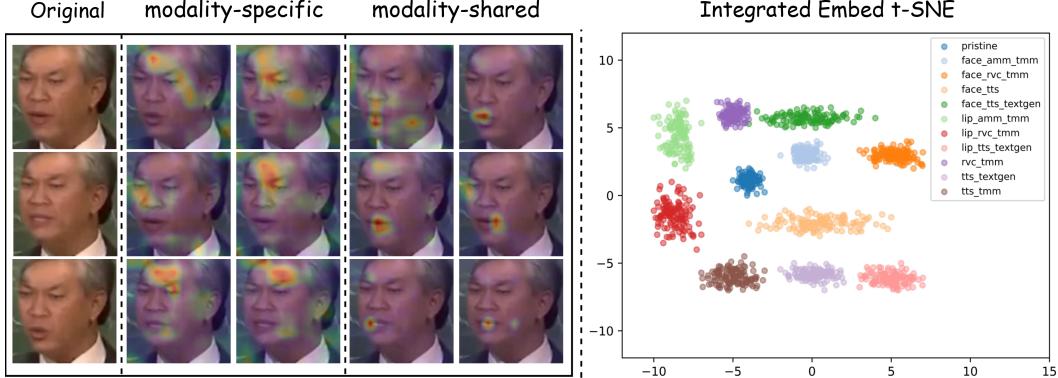


Figure 3: Visual illustrations of our method. **Left:** The visualization results of modality-specific learning and modality-shared learning by CAM [71]. **Origin** denotes the original video input. Modality-specific shows the attention distribution within the vision unimodal encoder, where attention focuses mainly on visual artifacts. Instead, Modality-shared illustrates the attention distribution when both modalities are aligned, with attention primarily on the lips and surrounding musculature. **Right:** t-SNE visualization on integrated embeddings.

ing only modality-specific features is significantly inferior to that with dual modality inputs, highlighting the critical role of multimodal information complementarity in enhancing single-modality learning. The absence of cross-modal distillation results in a decrease of 4% and 3.6% in average AUC and AP, respectively, indicating a clear drop in detection performance. This highlights the effectiveness of cross-modal distillation in alleviating distributional discrepancies between modalities and enhancing the discriminability of the learned feature representations. In addition, substituting the distillation strategy with a KL divergence loss leads to a further decline in performance, suggesting that KL divergence may be suboptimal for capturing complex multimodal relationships in this context.

4.4 More Visualization and Analysis

We further conducted a visualization experiment and performed more analyses. As depicted in Figure 3, we extracted the attention weights from the model and overlaid them onto the input image to examine the model’s attention distribution across different input regions. The term **modality-specific** refers to the encoder weights corresponding to the video unimodality in the CAD. While **modality-shared** denotes the attention weight distribution during joint learning. Observations with training on IDFForge-v2, specific vision modal focuses on pixel-level forgery traces around the eyes and nose, and these forgeries are typically the result of manipulations in the video source. The joint learning component places greater emphasis on the lips and the surrounding muscle tremors, and these tremors highlight the regions where general features capture shared high-level semantics across modalities.

The attention distribution in Figure 3 demonstrates the necessity of modality-shared semantic alignment analysis and modality-specific forensic verification, consistent with the conclusion of Theorem 1. In Figure 3, we present t-SNE visualizations of various forgery subsets from the IDFForge test set. The embeddings represent the final integrated vector from CAD, encompassing both the specific and shared components. It is evident that each subset forms a distinct cluster, suggesting that CAD effectively captures features corresponding to different forgery methods.

5 Conclusion

In this paper, we introduce CAD, a novel method that leverages mutual information maximization to enhance multimodal alignment and trans-membrane state distillation. Our approach integrates two key components: modality-shared semantic alignment analysis and modality-specific forensic verification, enabling the extraction and complementation of information across different states. This provides a fresh perspective for the design of face authentication models. CAD is the first framework to jointly

perform both coupled and decoupled feature learning in the domain of audio-visual authentication. It incorporates KL divergence and distillation loss, achieving state-of-the-art performance on the IDForge benchmark. Our findings not only demonstrate the effectiveness of CAD, but also offer a promising solution to mitigate the escalating risks associated with face forgery.

Limitations and Future Work: While CAD effectively integrates modality-shared semantics and modality-specific forensics, its current fusion and alignment mechanisms are relatively constrained in their modeling capacity. In future work, we plan to explore much larger models and integrate auto-regressive LLMs to enable more expressive and context-aware modeling of semantic consistency and cross-modal relationships. This could potentially further enhance the framework’s ability to reason over complex temporal cues and subtle modality inconsistencies in challenging forgery scenarios.

References

- [1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7, 2018.
- [2] Zaynab Almutairi and Hebah Elgibreen. A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*, 15(5):155, 2022.
- [3] Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–10, 2022.
- [4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.
- [5] Hong-Shuo Chen, Mozhdeh Rouhsedaghat, Hamza Ghani, Shuowen Hu, Suya You, and C.-C. Jay Kuo. Defakehop: A light-weight high-performance deepfake detector. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021.
- [6] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18689–18698, 2022.
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *ArXiv*, abs/2002.05709, 2020.
- [8] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15745–15753, 2021.
- [9] Yanbei Chen, Yongqin Xian, A. Sophia Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7012–7021, 2021.
- [10] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20:1 – 22, 2022.
- [11] Jikang Cheng, Zhiyuan Yan, Ying Zhang, Yuhao Luo, Zhongyuan Wang, and Chen Li. Can we leave deepfake data behind in training deepfake detector? *arXiv preprint arXiv:2408.17052*, 2024.
- [12] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1800–1807, 2017.

- [13] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other- audio-visual dissonance-based deepfake detection and localization. In Pradeep K. Atrey and Zhu Li, editors, *Proceedings of the 28th ACM International Conference on Multimedia*, pages 439–447, United States of America, 2020. Association for Computing Machinery (ACM). Publisher Copyright: © 2020 ACM. Copyright: Copyright 2021 Elsevier B.V., All rights reserved.; ACM International Conference on Multimedia 2020, MM 2020 ; Conference date: 12-10-2020 Through 16-10-2020.
- [14] Davide Cozzolino, Alessandro Pianese, Matthias Nießner, and Luisa Verdoliva. Audio-visual person-of-interest deepfake detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 943–952, 2023.
- [15] S. Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3994–4004, 2022.
- [16] FaceSwap. www.github.com/MarekKowalski/FaceSwap Accessed 2021-04-24.
- [17] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10491–10503, 2023.
- [18] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. *ArXiv*, abs/1803.07835, 2018.
- [19] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind one embedding space to bind them all. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023.
- [20] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14930–14942, 2022.
- [21] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5037–5047, 2021.
- [22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15979–15988, 2022.
- [23] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.
- [24] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [25] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection. *Appl. Soft Comput.*, 136:110124, 2023.
- [26] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.
- [27] Hasam Khalid, Minhan Kim, Shahroz Tariq, and Simon S. Woo. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. *Proceedings of the 1st Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection*, 2021.
- [28] Hasam Khalid, Shahroz Tariq, and Simon S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *ArXiv*, abs/2108.05080, 2021.

- [29] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3697–3705, 2017.
- [30] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
- [31] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. Audio anti-spoofing detection: A survey. *arXiv preprint arXiv:2404.13914*, 2024.
- [32] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3204–3213, 2020.
- [33] Weifeng Liu, Tianyi She, Jiawei Liu, Boheng Li, Dongyu Yao, Ziyou Liang, and Run Wang. Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [34] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92, 2019.
- [35] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. *Proceedings of the 28th ACM International Conference on Multimedia*, 2020.
- [36] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.
- [37] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, 2019.
- [38] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24480–24489, 2023.
- [39] Trevine Oorloff, Surya Koppisetty, Nicolò Bonettini, Divyarat Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27092–27102, 2024.
- [40] OpenAI. Gpt-4: A large multimodal model. <https://openai.com/research/gpt-4>, 2023. Accessed: 2023-01-21.
- [41] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*, 2024.
- [42] K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *MM '20: Proceedings of the 28th ACM International Conference on Multimedia*, pages 484–492, 2020.
- [43] Yanmin Qian, Zhengyang Chen, and Shuai Wang. Audio-visual deep neural network for robust person verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1079–1092, 2021.
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

- [45] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, 2022.
- [46] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. Faceforensics++: Learning to detect manipulated facial images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.
- [47] Jia Wen Seow, Mei Kuan Lim, Raphaël CW Phan, and Joseph K Liu. A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing*, 513:351–371, 2022.
- [48] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.
- [49] Chengrui Wang and Weihong Deng. Representative forgery mining for fake face detection. In *CVPR*, 2021.
- [50] Haoyue Wang, Sheng Li, Ji He, Zhenxing Qian, Xinpeng Zhang, and Shaolin Fan. Exploring depth information for detecting manipulated face videos. *CoRR*, abs/2411.18572, 2024.
- [51] Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T. Tan, and Haizhou Li. Seeing what you said: Talking face generation guided by a lip reading expert. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14653–14662, 2023.
- [52] Zezheng Wang, Zitong Yu, Chenxu Zhao, Xiangyu Zhu, Yunxiao Qin, Qiusheng Zhou, Feng Zhou, and Zhen Lei. Deep spatial gradient and temporal depth learning for face anti-spoofing. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5041–5050, 2020.
- [53] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4129–4138, 2023.
- [54] Deressa Wodajo and Solomon Atnafu. Deepfake video detection using convolutional vision transformer. *ArXiv*, abs/2102.11126, 2021.
- [55] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. Binaural audio-visual localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):2961–2968, May 2021.
- [56] Junhao Xu, Jingjing Chen, Xue Song, Feng Han, Haijun Shan, and Yu-Gang Jiang. Identity-driven multimedia forgery detection via reference assistance. In *ACM Multimedia 2024*, 2024.
- [57] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *ICCV*, pages 22658–22668, 2023.
- [58] Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8984–8994, 2024.
- [59] Zhiyuan Yan, Taiping Yao, Shen Chen, Yandan Zhao, Xinghe Fu, Junwei Zhu, Donghao Luo, Chengjie Wang, Shouhong Ding, Yunsheng Wu, et al. Df40: Toward next-generation deepfake detection. *arXiv preprint arXiv:2406.13495*, 2024.
- [60] Zhiyuan Yan, Junyan Ye, Weijia Li, Zilong Huang, Shenghai Yuan, Xiangyang He, Kaiqing Lin, Jun He, Conghui He, and Li Yuan. Gpt-imgeval: A comprehensive benchmark for diagnosing gpt4o in image generation. *arXiv preprint arXiv:2504.02782*, 2025.
- [61] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22412–22423, 2023.

- [62] Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. *arXiv preprint arXiv:2408.17065*, 2024.
- [63] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023.
- [64] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265, 2019.
- [65] Jiangyan Yi, Chenglong Wang, Jianhua Tao, Xiaohui Zhang, Chu Yuan Zhang, and Yan Zhao. Audio deepfake detection: A survey. *arXiv preprint arXiv:2308.14970*, 2023.
- [66] Daichi Zhang, Zihao Xiao, Shikun Li, Fanzhao Lin, Jianmin Li, and Shiming Ge. Learning natural consistency representation for face forgery video detection. In *European Conference on Computer Vision*, pages 407–424. Springer, 2024.
- [67] Yi Zhang, Weize Gao, Changtao Miao, Man Luo, Jianshu Li, Wenzhong Deng, Zhe Li, Bingyu Hu, Weibin Yao, Wenbo Zhou, et al. Inclusion 2024 global multimedia deepfake detection: Towards multi-dimensional facial forgery detection. *arXiv preprint arXiv:2412.20833*, 2024.
- [68] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2185–2194, 2021.
- [69] Wei Zheng, Mengyuan Yue, Shuhuan Zhao, and Shuaiqi Liu. Attention-based spatial-temporal multi-scale network for face anti-spoofing. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):296–307, 2021.
- [70] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15024–15034, 2021.
- [71] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2015.
- [72] Yang Zhou, Xintong Han, Eli Shechtman, Jose Echevarria, Evangelos Kalogerakis, and Dingzeyu Li. Makelttalk: speaker-aware talking-head animation. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2023.
- [73] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14780–14789, 2021.
- [74] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Cai Wan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *The Twelfth International Conference on Learning Representations*, 2024.