International Journal of Computer Vision (2024) 132:5255–5269
https://doi.org/10.1007/s11263-024-02128-1

Fine-Grained Multimodal DeepFake Classification via Heterogeneous Graphs

Qilin Yin[1,2,3] · Wei Lu[1,2,3] · Xiaochun Cao[4] · Xiangyang Luo[5] · Yicong Zhou[6] · Jiwu Huang[7]

Received: 27 July 2023 / Accepted: 17 May 2024 / Published online: 6 June 2024

Abstract

Nowadays, the abuse of deepfakes is a well-known issue since deepfakes can lead to severe security and privacy problems. And this situation is getting worse, as attackers are no longer limited to unimodal deepfakes, but use multimodal deepfakes, *i.e.*, both audio forgery and video forgery, to better achieve malicious purposes. The existing unimodal or ensemble deepfake detectors are demanded with fine-grained classification capabilities for the growing technique on multimodal deepfakes. To address this gap, we propose a graph attention network based on heterogeneous graph for fine-grained multimodal deepfake classification, i.e., not only distinguishing the authenticity of samples, but also identifying the forged types, e.g., video or audio or both. To this end, we propose a positional coding-based heterogeneous graph construction method that converts an audio-visual sample into a multimodal heterogeneous graph according to relevant hyperparameters. Moreover, a cross-modal graph interaction module is devised to utilize audio-visual synchronization patterns for capturing inter-modal complementary information. The de-homogenization graph pooling operation is elaborately designed to keep differences in graph node features for enhancing the representation of graph-level features. Through the heterogeneous graph attention network, we can efficiently model intra- and inter-modal relationships of multimodal data both at spatial and temporal scales. Extensive experimental results on two audio-visual datasets FakeAVCeleb and LAV-DF demonstrate that our proposed model obtains significant performance gains as compared to other state-of-the-art competitors. The code is available at https://github.com/yinql1995/Fine-grained-Multimodal-DeepFake-Classification/.

Keywords Multimodal deepfake classification · Audio-visual model · Graph neural networks · Heterogeneous graphs

Communicated by Segio Escalera.

✉ Wei Lu
luwei3@mail.sysu.edu.cn

Qilin Yin
yinqlin@mail2.sysu.edu.cn

Xiaochun Cao
caoxiaochun@mail.sysu.edu.cn

Xiangyang Luo
luoxy_ieu@sina.com

Yicong Zhou
yicongzhou@um.edu.mo

Jiwu Huang
jwhuang@smbu.edu.cn

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, China

[2] Ministry of Education Key Laboratory of Information Technology, Sun Yat-sen University, Guangzhou 510006, China

[3] Guangdong Province Key Laboratory of Information Security Technology, Sun Yat-sen University, Guangzhou 510006, China

[4] School of Cyber Science and Technology, Sun Yat-sen University, Shenzhen 518107, China

[5] State Key Laboratory of Mathematical Engineering and Advanced Computing, Henan 450001, China

[6] Department of Computer and Information Science, University of Macau, Macau, China

[7] Guangdong Laboratory of Machine Perception and Intelligent Computing, Faculty of Engineering, Shenzhen MSU-BIT University, Shenzhen 518116, China

1 Introduction

The combination of artificial intelligence technology and multimedia synthesis technology has led to the popularization of deepfakes in which both video and audio are sufficient to be faked (Juefei-Xu et al., 2022). There are various ways of creating deepfakes, including text-to-speech(TTS) (Jia et al., 2018; Ping et al., 2018), voice conversion(VC) (Arik et al., 2018), face reenactment (Prajwal et al., 2020; Tulyakov et al., 2018; Jamaludin et al., 2019), and face swapping (Korshunova et al., 2017; Nirkin et al., 2019). The new internet communication method, which takes individual media communication as a branch and closely focuses on hotspots, further intensifies the breeding and abuse of deep falsification in cyberspace, and seriously undermines our trust in online media.

To defend against the potential risks of these forged media, numerous efforts have been devoted and achieved promising performances on a single modality forgery i.e., either video or audio, in recent years. For video forgery, detection techniques can be divided into textural feature based and semantic feature based methods. Textural feature based methods focus on capturing discriminative frame-level features, including face blending boundaries (Li et al., 2020a; Zhao et al., 2020) and forgery signals on frequency spectrum (Chen et al., 2021; Qian et al., 2020; Liu et al., 2021a). Semantic feature based methods aim to model temporal inconsistency by extracting trends in irregular facial movements (Haliassos et al., 2021, 2022) or differences over adjacent frames (Li et al., 2020b; Gu et al., 2021; Lu et al., 2023). For audio forgery, most exiting spoofed speech detection methods rely on extracting acoustic representations like MFCC (Muda et al., 2010), STFT and CQCC (Todisco et al., 2016) from raw waveform signals to classify.

However, a convincing deepfake usually require meticulous manipulations of both video and audio channels to more efficiently deliver spurious information and fake news. As shown in Fig. 1, to create different social or political threats, attackers can manipulate audio-visual multimodal samples with different types of audio and video combinations. When faced with such diverse manipulation types, existing unimodal or ensemble-based detection methods struggle to achieve satisfactory authenticity detection performance, let alone accurately distinguish the manipulation types (Khalid et al., 2021). This is because unimodal-based approaches don't consider multimodal forgery scenarios, while ensemble-based approaches don't take full advantage of the complementary information between modalities.

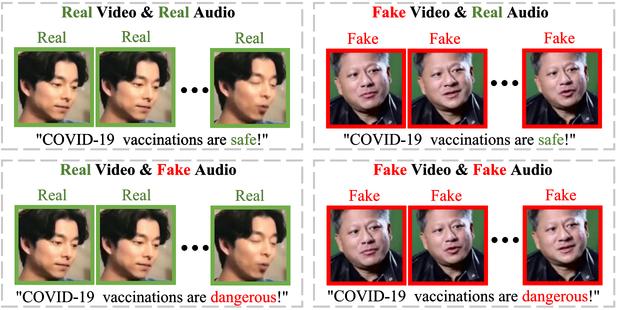

Real Video & Real Audio
Real
Real
Real
"COVID-19 vaccinations are safe!"
Fake Video & Real Audio
Fake
Fake
Fake
"COVID-19 vaccinations are safe!"
Real Video & Fake Audio
Real
Real
Real
"COVID-19 vaccinations are dangerous!"
Fake Video & Fake Audio
Fake
Fake
Fake
"COVID-19 vaccinations are dangerous!"

Fig. 1 Different types of audio-visual manipulation. On the top left is a real video with the subject saying "Vaccinations are safe". On the top right is manipulating the video to get another person to say the same content. On the bottom left is manipulating the audio to change what the person is saying. The bottom right is fabricating misinformation by tampering with video and audio

To overcome the shortcomings in the existing studies, we propose a multimodal deepfake classification model, which can classify deepfakes that are possibly any combination of audio or video. Generally speaking, visual and audio modalities are often intertwined, and a synchronization pattern exists between the two (Zhou and Lim, 2021). The interplay between these two modalities is critical for detecting multimodal deepfakes. Heterogeneous graphs are a compact, efficient and scalable way to represent data involving several different entities and their relationships. Modeling the interaction of entities (including modalities) with heterogeneous graphs is a relatively new paradigm and has been successfully used to solve several tasks in the fields of computer vision and natural language processing, e.g., visual question answering (Saqur and Narasimhan, 2020; Le et al., 2021), multimedia recommendation (Wei et al., 2019; Wang et al., 2021), medical image synthesis (Cao et al., 2023) and audio-visual sentiment analysis (Wu and Li, 2023; Yang et al., 2021). Motivated by the success of graph-based approaches, we propose a heterogeneous graph-based model to detect deepfakes of multiple modalities.

To this end, we first transform an audio-visual sample into a heterogeneous graph with two subgraphs. It is worth noting that positional encoding is introduced into each subgraph to preserve the intra-modal temporal relationships of the audio or video. We next develop a heterogeneous graph attention network with two well-designed modules, namely, cross-modal graph interaction module and the de-homogenization graph pooling module. In combination with adjacency matrices, cross-modal graph interaction module utilizes a self-attention mechanism to establish information interaction between modalities. This allows capturing rich inter-modal complementary information from the synchronization patterns of audio-visual pairs. The de-homogenization graph pooling module enhances the graph-level representation of the model by eliminating homogenized graph nodes and their corresponding adjacency matrices. Our heterogeneous graph model creates a shared space for audio and visual modalities that explores their spatial and temporal relationships explicitly. Overall, the proposed model can take full advantage of multimodal information for fine-grained multimodal deepfake classification.

Our contributions are summarized as follows:

- We propose a heterogeneous graph attention network for fine-grained multimodal deepfake classification. The multimodal heterogeneous graph can bring out a closer coupling between audio and video, thereby capturing modality-specific information as well as complementary information between modalities.
- A cross-modal graph interaction module is proposed to promote the interaction between modalities, thereby efficiently capturing rich inter-modal complementary information according to audio-visual synchronization patterns.
- A de-homogenization graph pooling module is devised to measure the degree of homogeneity between graph nodes. The diversity of graph-level features is enhanced by eliminating the homogeneous graph nodes, which is beneficial to improving the fine-grained classification performance of the model.

The rest of this paper is organized as follows: Sect. 2 explains multimodal deepfake datasets and graph neural networks, Sect. 3 elaborates the proposed method, Sect. 4 shows experiments, and Sect. 5 concludes the paper.

2 Related Work

2.1 Multimodal Deepfake Datasets

Benefiting from the continuous development of forgery synthesis technology, there are several public deepfake datasets, such as FaceForensics++ (Rossler et al., 2019), DFDC (Dolhansky et al., 2020), Celeb-DF (Li et al., 2020c), DeeperForensics (Jiang et al., 2020), and WildDeepFake (Zi et al., 2020), etc. However, most of these datasets are designed for the binary task of deepfake classification and focus primarily on video forgery without the corresponding audio forgery. To the best of our knowledge, DFDC is the first dataset containing synthesized audio with the video, but it does not provide respective labels for audio and video to specify if the video is fake or the audio. To overcome this drawback, Khalid et al. (2021) released a novel Audio-Video Deepfake dataset (FakeAVCeleb), which contains not only deepfake videos but also respective synthesized lip-synced fake audios. They used the latest face-swap and face-reenactment methods to manipulate videos and used transfer learning-based real-time voice cloning tools to generate cloned audio. Recently, Cai et al. (2022) proposed a large sized dataset LAV-DF, which includes deepfake videos as well as synthesized fake audios synchronized with the videos. All samples in LAV-DF are partially faked, that is, fake content might constitute only small part of a long real video, which greatly increases the difficulty of detection. LAV-DF is not only suitable for the task of fine-grained multimodal deepfake classification, but also for temporal forgery localization.

2.2 Multimodal Deepfake Detection

Recently, some works consider utilizing the audio-visual information (i.e., the inconsistency of audio and video information) to conduct multimodal deepfake detection. Some works (Chugh et al., 2020; McGurk and MacDonald, 1976) extract the audio and visual information and directly analyze the dissimilarity between two modalities. These methods do not take into account audio forgery scenarios and have limited performance. To address this problem, JAVDD (Zhou and Lim, 2021) and BA-TFD (Cai et al., 2022) considered the interactions between different modalities and performed joint audio-visual learning to capture the effective inconsistency of two modalities to achieve multimodal deepfake detection. Besides, Cheng et al. (2023) used a self-supervised manner to learn the audio-visual synchronization patterns in real videos and applied these pretrained feature to detect multimodal forgery. However, due to the gap between modalities, more intrinsic relationships and subtle inconsistencies between the two modalities need to be further explored. Compared to these methods, our proposed method leverages a novel heterogeneous graph structure to accurately model the local audio-visual correspondence, which facilitates the capture of inter-modal inconsistencies.

2.3 Graph Neural Networks

In recent years, graph neural networks (GNNs) (Scarselli et al., 2008; Brissman et al., 2023) have attracted growing attention, especially variants such as graph convolution networks (GCNs) (Fu et al., 2021) or graph attention networks (GATs) (Veličković et al., 2017). Since traditional neural networks can only handle structured data, they are powerless in face of non-Euclidean data. GNN relies on its powerful points and edges to model non-Euclidean data, efficiently solving the problem of graph-structured data encountered in practical applications.

2.3.1 Homogeneous Graph Based Methods

A number of studies have shown the utility and appeal of unimodal homogeneous graphs for various modality processing tasks, such as text (Peng et al., 2017; Veyseh et al., 2019), audio (Zhang et al., 2019; Tak et al., 2021), and video (Qi et al., 2018; Liu et al., 2021b). Peng et al. (2017) proposed a general relationship extraction framework based on graph long and short term memory networks, which can be easily extended to cross-sentence n-ary relation extraction. Veyseh et al. (2019) proposed a new GCN based approach to integrate semantic and syntactic structures by introducing affinity matrix. Zhang et al. (2019) applied GCNs to a few-shot audio classification task in order to derive attention vectors that help improve the discrimination between different audio instances. Tak et al. (2021) used a spectro-temporal graph to model the relationship between cues spanning different subbands and temporal intervals for speech deepfake detection. Qi et al. (2018) proposed a graph parsing neural network in order to achieve the purpose of detecting and recognizing human-object interactions in videos, which is to infer parsing graphs in an end-to-end manner. Liu et al. (2021b) proposed a graph attention spatio-temporal convolutional network that comprises of interleaved temporal convolutional and graph attention blocks for 3D human pose estimation in video.

2.3.2 Heterogeneous Graph Based Methods

In real-life scenarios, heterogeneous graphs are the most relevant to the actual problem compared to homogeneous graphs. Multimodal heterogeneous graphs have been successfully used to address various problems in computer vision and natural language processing, such as visual question answering (Saqur and Narasimhan, 2020), multimedia recommendation (Wang et al., 2021), and audio-visual sentiment analysis (Yang et al., 2021). Saqur and Narasimhan (2020) proposed multimodal graph networks to learn joint graph-based representation for better solving the problem of compositional generalization for visual question answering. Wang et al. (2021) proposed a dual graph neural network based on the user micro-video bipartite and user co-occurrence graphs for micro-video recommendation. Yang et al. (2021) used multi-channel graph neural networks to learn multimodal representations and then fused multimodal information to predict the sentiment of image-text pairs. Motivated by the success of graph-based methods in multimodal problems, we propose a heterogeneous graph-based approach to capture the synchronization patterns in audio-video sample for fine-grained multimodal deepfake classification.

3 Approach

In this section, the proposed approach for fine-grained multimodal deepfake classification is described. First, we present a detailed description of how to transform an audio-visual sample into a heterogeneous graph without losing the temporal order. Then, we elaborate the framework of the heterogeneous graph-based graph attention network. Finally, we technically introduce the proposed network modules, namely the cross-modal graph interaction module and the de-homogenization graph pooling module, which can enhance the representation ability of graph neural network and facilitate better graph classification.

Algorithm 1 The intra-modality information aggregation process

```
Require: subgraph $\mathcal{G}_v = \{\mathcal{V}^v, \mathcal{E}_{vv}, \mathrm{n}^v, e^v\}$
Ensure: updated subgraph $\widetilde{\mathcal{G}_v}$
$\mathrm{n}^{v_q} = \sigma_q(\mathrm{n}^v)$
$\mathrm{n}^{v_k} = \sigma_k(\mathrm{n}^v)$
$\mathrm{n}^{v_v} = \sigma_v(\mathrm{n}^v)$
for $i = 1 : P$ do
    for $j = 1 : P$ do
        $c_{ij}^v = \frac{\mathrm{n}_i^{v_q}}{\left|\mathrm{n}_i^{v_q}\right|_2} \cdot \frac{\mathrm{n}_j^{v_k}}{\left|\mathrm{n}_j^{v_k}\right|_2}$
    end for
    $\left\{\alpha_{ij}^v\right\}_{j=1}^P \leftarrow \mathrm{softmax}(\left\{\frac{c_{ij}^v}{e_{ij}^v}\right\}_{j=1}^P)$
    $\widetilde{\mathrm{n}_i^v} = \mathrm{n}_i^v + \beta \cdot \sum_{j=1}^P \alpha_{ij}^v \cdot \mathrm{n}_j^{v_v}$
end for
Return: $\widetilde{\mathcal{G}_v} \leftarrow \widetilde{\mathrm{n}^v}$
```

3.1 Positional Coding-Based Heterogeneous Graph Construction Method

In contrast to unimodal data, multimodal data has tight spatio-temporal correlation between its different components. For example, there is a strong correlation between the facial motions (viseme) and the pronounced syllables (phoneme) (McGurk and MacDonald, 1976). Making full use of interaction between audio and video is beneficial for fine-grained classification of multimodal deepfakes. To learn this interaction, we convert an audio-visual sample into a multimodal heterogeneous graph. The powerful coupling ability of the heterogeneous graph helps to explore the subtleties of the intrinsic synchronization patterns between audio and video. However, the heterogeneous graph structure is not naturally defined here, we propose a positional coding-based heterogeneous graph construction method to add inter- and intra-modality edges.

As shown in Fig. 2, an audio-visual sample is converted to a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, O, R)$, where $\mathcal{V}$ represents the set of nodes, $\mathcal{E}$ represents the set of edges, O is the set of node types (audio or video), and R is the set of edge types (audio-audio, video-video, and audio-video). First, the video and audio frames are uniformly divided into P and Q clips respectively and then these clips are used for high-level semantic feature extraction. That is to say the heterogeneous graph comprises of the video node sets $\mathcal{V}^v = \{v_j\}_{j=1}^P$ and the audio node sets $\mathcal{V}^a = \{a_i\}_{i=1}^Q$. Each video node $v_j \in \mathcal{V}^v$ is associated with feature vector $\mathrm{n}_j^v \in \mathbb{R}^d$. Similarly, an audio node $a_i \in \mathcal{V}^a$ owns the feature vector $\mathrm{n}_i^a \in \mathbb{R}^d$. Next, according to the predefined parameters, (1) *neighbour* and (2) *overlap*, we add intra- and inter-modality edges

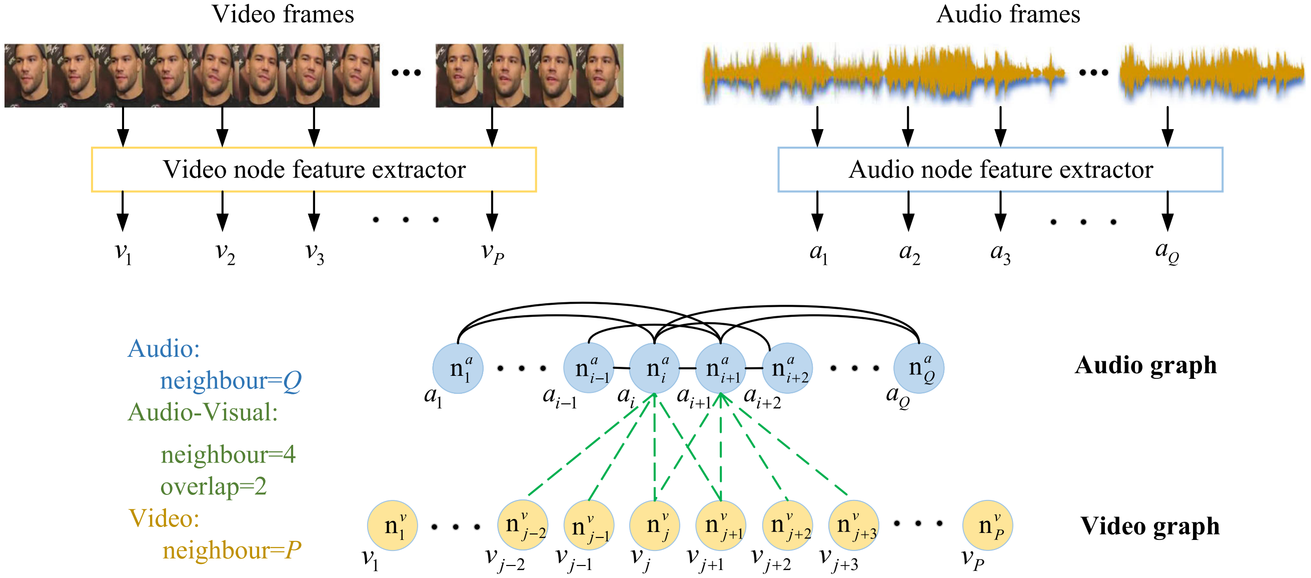

Video frames
Audio frames
Video node feature extractor
Audio node feature extractor
v_1 v_2 v_3 v_P
a_1 a_2 a_3 a_Q
Audio:
neighbour=Q
Audio-Visual:
neighbour=4
overlap=2
Video:
neighbour=P
Audio graph
Video graph
n^a_1 n^a_{i-1} n^a_i n^a_{i+1} n^a_{i+2} n^a_Q
a_1 a_{i-1} a_i a_{i+1} a_{i+2} a_Q
n^v_1 n^v_{j-2} n^v_{j-1} n^v_j n^v_{j+1} n^v_{j+2} n^v_{j+3} n^v_P
v_1 v_{j-2} v_{j-1} v_j v_{j+1} v_{j+2} v_{j+3} v_P

Fig. 2 Positional coding-based heterogeneous graph construction process. For simplicity, the edges are only shown for a_i and a_{i+1}. Similar connections are added for each node

$\mathcal{E} = \{\mathcal{E}_{aa}, \mathcal{E}_{vv}, \mathcal{E}_{av}\}$ and denote the corresponding adjacency matrices A_{aa}, A_{vv}, and A_{av}. The former parameter *neighbour* defines the number of adjacent nodes of each node in intra- or inter-modality. The latter parameter *overlap* defines the number of overlaps between different attribute nodes connected to adjacent nodes.

Since both audio and video are directed sequence data, the constructed heterogeneous graph is undirected. In order to preserve the temporal order of data, we introduce the position coding to assign the temporal weight e_{t_1,t_2} to each edge in intra-modality edge sets ($\mathcal{E}_{aa}$ and $\mathcal{E}_{vv}$). Taking the video-video edge as an example, the temporal weight $e^v_{t_1,t_2}$ can be expressed by:

$$e^v_{t_1,t_2} = \begin{cases} t_1 - t_2, \; 1 \leq t_1, t_2 \leq P \& t_1 \neq t_2 \\ 0.5, \qquad\quad t_1 = t_2 \end{cases} \tag{1}$$

where t_1 and t_2 represent the temporal number of the nodes on each side of an edge. $e^v_{t_1,t_2}$ defines the time interval between two nodes and is specifically set to 0.5 when the node is self-looping. In total, we have four hyperparameters for the heterogeneous graph construction.

3.2 Heterogeneous Graph Attention Network

Combining the abovementioned heterogeneous graph construction method, we propose a graph attention network based on the heterogeneous graph for fine-grained multimodal deepfake classification. As shown in Fig. 3, the proposed model is a two-stage multimodal spatio-temporal feature extractor. In the former stage, we use a video node feature extractor and an audio node feature extractor to extract high-level semantic features from uniformly segmented video and audio clips, respectively. In the latter stage, we construct a heterogeneous graph attention network to capture modality-specific information as well as complementary information between modalities both at spatial and temporal scales.

To this end, the heterogeneous graph attention network comprises of three learning stages: (1) intra-modality information aggregation; (2) cross-modality information interaction; (3) de-homogenization graph pooling. In the first stage, each subgraph captures modality-specific information by updating its own node features through an attention mechanism. Specifically, the feature of each node is updated with the aggregated information from neighboring nodes of the same subgraph by a self-attention mechanism. For video subgraph $\mathcal{G}_v = \{\mathcal{V}^v, \mathcal{E}_{vv}, \mathrm{n}^v, e^v\}$, the intra-modality information aggregation process is illustrated in Algorithm 1, where σ is an affine transform, c^v_{ij} refers to the correlation between node v_i and v_j, α^v_{ij} refers to the attention weight between node v_i and v_j, which reflects how informative one node is of another, where the higher weight implies a higher connective strength, and β is a learnable parameter. The update process of audio subgraph is similar to the video subgraph.

In the second stage, a cross modal graph interaction module is introduced to promote the interaction between audio and video subgraphs for capturing complementary information between modalities. In the third stage, a de-homogenization graph pooling module is designed to compress the scale of the heterogeneous graph, so as to obtain graph-level representations for deepfake classification. The latter two stages are described in detail below.

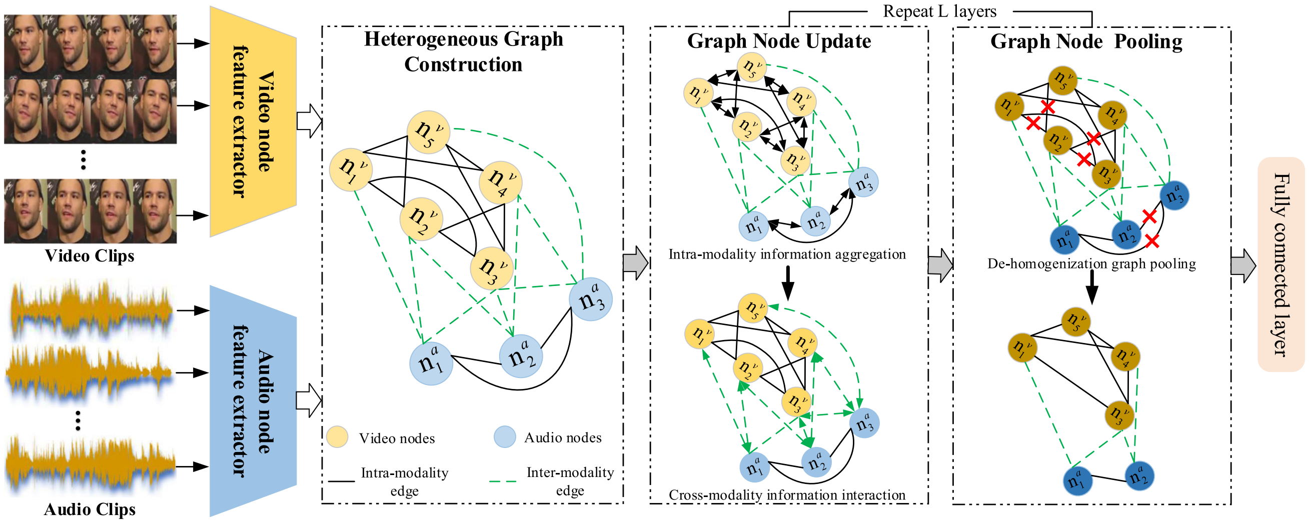

Video Clips
Audio Clips
Video node feature extractor
Audio node feature extractor
Heterogeneous Graph Construction
Video nodes
Audio nodes
Intra-modality edge
Inter-modality edge
Repeat L layers
Graph Node Update
Intra-modality information aggregation
Cross-modality information interaction
Graph Node Pooling
De-homogenization graph pooling
Fully connected layer
n_1^v n_2^v n_3^v n_4^v n_5^v n_1^a n_2^a n_3^a

Fig. 3 Diagrammatic overview of the proposed fine-grained multimodal deepfake classification framework

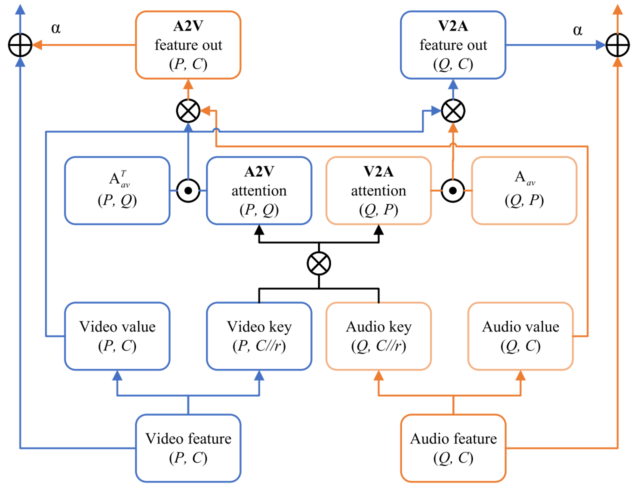


Fig. 4 Cross-modal graph interaction module. ⊙ and ⊗ denote element-wise multiplication and matrix multiplication, respectively

3.3 Cross-Modal Graph Interaction Module

Visual and audio modalities are often intertwined that video and audio can complement each other by providing semantics from different perspectives. Therefore, we propose a cross-modal graph interaction module to model the interaction between video and audio to capture complementary information. The computation process of the interaction module is described in Fig. 4. The proposed module can convert the video subgraph feature F_v and the audio subgraph feature F_a to $\widehat{F_v}$ and $\widehat{F_a}$.

Taking the conversion of F_a as an example, first, $F_a \in \mathbb{R}^{Q\times C}$ is converted into two other representations using different affine transforms. The value representation $F_a^v \in \mathbb{R}^{Q\times C}$ denotes the modality-specific information and the other key representation $F_a^k \in \mathbb{R}^{Q\times C/r}$ measures the correlation between two modalities. Then, the video to audio (**V2A**) attention map $\mathrm{att}_{v2a} \in \mathbb{R}^{Q\times P}$ is measured by the attention mechanism:

$$\mathrm{att}_{v2a} = \mathrm{softmax}(\gamma(F_a^k) \otimes (\gamma(F_v^k))^T), \tag{2}$$

where γ denotes the projection function. att_{v2a} reweights features in video modality according to its correlation with the audio modality. Finally, applying the inter-modality adjacency matrix $\mathrm{A}_{av} \in \mathbb{R}^{Q\times P}$ to **V2A** attention map att_{v2a} and the corresponding value representation $F_v^v \in \mathbb{R}^{P\times C}$, we can get the refined video feature. Adding the refined video feature to F_a yields the enhanced feature $\widehat{F_a}$:

$$\widehat{F_a} = F_a + \alpha \cdot (F_v^v \otimes (\mathrm{att}_{v2a} \odot \mathrm{A}_{av})), \tag{3}$$

where α is a learnable parameter. The calculation for $\widehat{F_v}$ is the same. $\widehat{F_a}$ and $\widehat{F_v}$ embody complementary information and promote the feature learning for each other.

3.4 De-homogenization Graph Pooling Module

Our objective is to classify entire graphs, as opposed to the more common task of classifying each node. Hence, we propose a de-homogenization graph pooling module to eliminate redundant graph nodes and compress the scale of graphs. After multiple iterations, the remaining node-level representations are concatenated to obtain graph-level representations. Denoting the input graph $\mathrm{X} \in \mathbb{R}^{4\times 5}$, the pooling process is described in Fig. 5.

First, we calculate the node similarity matrix $\mathrm{M} \in \mathbb{R}^{4\times 4}$ between the input graph X and its own transpose X^T by self-attention. Then, we sum up each row of the similarity

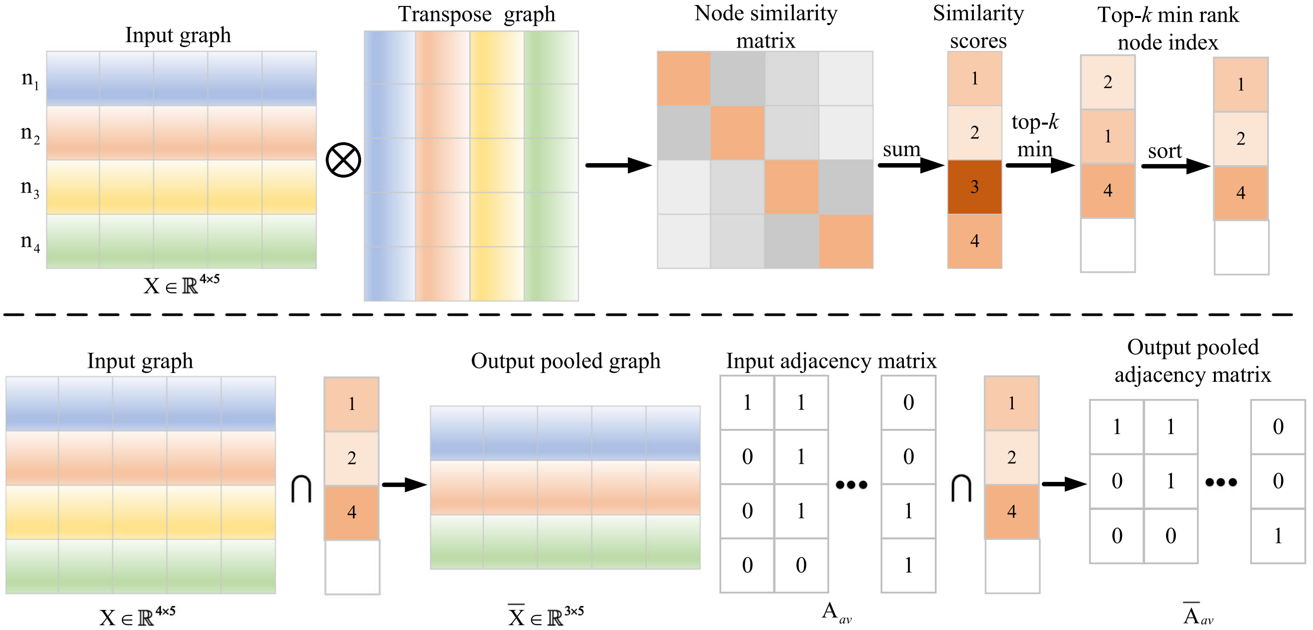


Fig. 5 Diagrammatic of the de-homogenization graph pooling module. ⊗ and ∩ denote matrix multiplication and the aggregation operation based on indexes, respectively

matrix to get similarity score of each node. The k nodes with lower similarity scores are selected and the corresponding indexes are obtained. Since the video contains a large number of duplicate frames, there are some redundant nodes with similar features in the graph. Therefore, using the minimum similarity score as a metric is beneficial to expand the variability of node-level features and enhance the characterization of graph-level features. Finally, we pool the input graph X according to indexes to obtain the output graph $\overline{\mathrm{X}}$. Meanwhile, we also apply the same pooling operation to the corresponding inter-modality adjacency matrix A_{av} to obtain the new adjacency matrix $\overline{\mathrm{A}}_{av}$, thus ensuring the correct correspondence between audio and video in the following operations.

4 Experiments

4.1 Datasets

To evaluate our method, we conduct experiments on two challenging datasets: FakeAV (Khalid et al., 2021) and LAV-DF (Cai et al., 2022).

- FakeAV is a large-scale multimodal forgery dataset containing 500 real videos and 19,500 fake videos. There are four forgery types in FakeAVCeleb, namely, ‘real-real’(video is real and audio is real), ‘fake-fake’, ‘fake-real’, and ‘real-fake’. Since some unimodal competitors only focus on visual information or audio information, we additionally prepared a video subset of FakeAV, excluding those with audio-only modifications and an audio subset of FakeAV, excluding those with video-only modifications.
- LAV-DF is another large multimodal forgery dataset, consisting of 36,431 real videos and 99,873 fake videos. Since the fake videos are partially forged and the duration of fake segment is in the range of [0–1.6] sec, it is a very challenging dataset. Similarly, we also obtained the video subset and the audio subset of LAV-DF.

4.2 Baseline Methods

We compare our method with some SOTA unimodal video methods (e.g., UIA-ViT (Zhuang et al., 2022), RECCE (Caoet al., 2022), FTCN (Zheng et al., 2021), and Real-Forensics (Haliassos et al., 2022)) and SOTA unimodal audio methods (e.g., TSSDN (Hua et al., 2021), RawGAT (Tak et al., 2021), and SSLAS (Tak et al., 2022)) to show how multimodal information can be helpful for deepfake detection. Meanwhile, we also select the unimodal video methods and unimodal audio method with better performance from the above methods to form the ensemble methods for comparison to demonstrate the difficulty of the fine-grained classification task. Moreover, to the best of our knowledge, there is no multimodal forgery detection method that not only distinguish the authenticity of samples, but also point out the forgery types of samples. To demonstrate the effectiveness of the proposed method on the fine-grained classification task, we compare it with several methods that use multimodal information as an aid to detect unimodal forgery. MDS (Chugh et al., 2020) and JAVDD (Zhou and Lim, 2021) mainly focused on detect video-only forgery. BA-TFD (Cai et al., 2022) was designed for temporal forgery localization. Since both datasets, FakeAVCeleb and LAV-DF, are new datasets that have just been made publicly available, no baseline methods have been previously experimented on these datasets. Therefore, we reproduce all baseline methods for comparison. It is worth noting that the code for JAVDD is not publicly available. As a result, we implemented it and tried to match the experimental setup of the original paper as much as possible to ensure the fairness of the comparison.

4.3 Feature Encoders

Audio encoder: to extract audio node feature, each audio is segmented into non-overlapping 1000 ms clips. For each audio clip, a log-mel spectrogram is computed by the short-time Fourier transform. The dimension of each log-mel spectrogram is 64 × 100. We use the 2D ResNet to extract 512-dimensional features for each log-mel spectrogram.

Video encoder: each video is divided into non-overlapping 400 ms video clips. Each clip comprises of 10 frames of size 128 × 128 × 3. Due to the high repetition of frame contents, we randomly sample 4 frames from each clip as input to the Spatio-temporal model proposed by (Yin et al., 2023). This creates 512-dimensional features for each video clip.

4.4 Implementation Details

Each video produces a heterogeneous graph with $Q = 4$ audio and $P = 10$ video nodes, where each node corresponds to a 1000 ms audio clip or a 400 ms video clip. In general, multiple temporal slices of audio can map to a single video frame due to the redundancy of frames, temporally. Therefore, different time ranges of audio and video nodes in localized segments of the whole video do not lead to synchronization issue. The graph construction hyperparameters are set to *audio neighbour* = 4, *audio-visual neighbour* = 4, *audio-visual overlap* = 2, and *video neighbour* = 10, for all experiments. To verify the rationality of the graph construction parameter setting, we repeat our experiments with different *audio-visual neighbour* values and report results in Sect. 4.7.2. During the training phase, the whole network is initialized randomly and is optimized by Adam optimizer with a learning rate of 1e−3, a batch size of 12, betas of 0.9 and 0.999, and epsilon of 1e−8.

4.5 Performance Comparisons

4.5.1 Intra-dataset Comparisons

Unimodal Methods We compare the proposed method with some unimodal methods on the video subset or audio subset of two popular multimodal deepfake datasets, FakeAV and LAV-DF. As shown in Tables 1 and 2, our proposed method can achieve the best detection performance in all settings, both for the video subsets and audio subsets. For the experimental results on video subsets, all unimodal video methods perform poorly on FakeAV dataset because of the sample imbalance problem in FakeAV, i.e., the number of real samples is much less than the number of fake samples. In addition, the detection performance of UIA-ViT and RECCE has a large gap compared to FTCN and RealForensics on the LAV-DF dataset. This is because UIA-ViT and RECCE are frame-level methods, while FTCN and RealForensics are video-level methods. To achieve video-level results for the frame-level methods, we average the model predictions for each frame across the entire video. However, in LAV-DF dataset, the fake videos are partially forged and the duration of the fake segment is small, which makes it unfriendly to the frame-level methods.

For the experimental results on audio subsets, all unimodal audio methods do not suffer from the sample imbalance problem of FakeAV dataset and also achieve comparable performance on the more challenging LAV-DF dataset. This is due to the fact that audio content is purer and clearer compared to complex video content. Audio features are simpler and more effective than video features. In contrast, our method is also immune to sample imbalance problem and outperforms all compared opponents due to the learning and exploiting of multimodal complementary information by the heterogeneous graph model. Multimodal complementary information can capture more tampering traces in the deepfakes and effectively mitigate the effects of sample imbalance problem with the help of audio information.

Ensemble Methods Further, two unimodal video methods (FTCN and RealForensics) and one unimodal audio method (RawGAT) are chosen to compose the ensemble models for four classification comparison based on their performance on the video and audio subsets. The results presents in Table 3. The proposed method shows a promising results and outperforms the strong ensemble method by about 5.7% under ACC and 19.34% under AUC on FakeAV dataset. Even on the challenging LAV-DF dataset, our method can also outperform all ensemble methods. This is because the ensemble method ignores the complementary information between modalities and is simply a superimposition of unimodal approaches. Furthermore, the ensemble method is not an end-to-end model, and training video branch and audio branch separately often brings extra computational and storage costs while not being conducive to unleashing the power of deep learning. This can also lead to ensemble models that are vulnerable to the performance of a single unimodal branch, resulting in poor stability. Compared with the ensemble models, our method can bring a close coupling between modalities and take full advantage of multimodal information, which allows

Table 1 Quantitative results (video-level ACC and AUC) for binary classification of unimodal methods on the video subset of FakeAV and LAV-DF datasets

Methods	FakeAV		LAV-DF	
	ACC	AUC	ACC	AUC
UIA-ViT (Zhuang et al., 2022)	0.9680	0.5793	0.6577	0.7550
RECCE (Caoet al., 2022)	0.9648	0.6013	0.6974	0.7738
FTCN (Zheng et al., 2021)	0.9716	0.5976	0.7786	0.8576
RealForensics (Haliassos et al., 2022)	0.9687	0.5986	0.7949	0.8766
Ours	**0.9977**	**0.9997**	**0.9942**	**0.9999**

The bold results are the best

Table 2 Quantitative results (video-level ACC and AUC) for binary classification of unimodal methods on the audio subset of FakeAV and LAV-DF datasets

Methods	FakeAV		LAV-DF	
	ACC	AUC	ACC	AUC
TSSDN (Hua et al., 2021)	0.9916	0.9748	0.8127	0.8519
RawGAT (Tak et al., 2021)	0.9964	0.9913	0.8421	0.8885
SSLAS (Tak et al., 2022)	0.9922	0.9796	0.8376	0.8726
Ours	**0.9970**	**0.9999**	**0.9983**	**0.9999**

The bold results are the best

Table 3 Quantitative results (video-level ACC and AUC) for four classification of ensemble methods on FakeAV and LAV-DF datasets

Methods	FakeAV		LAV-DF	
	ACC	AUC	ACC	AUC
FTCN+RawGAT	0.9421	0.8063	0.7063	0.8726
RealForensics+RawGAT	0.9349	0.7919	0.7197	0.8868
Ours	**0.9991**	**0.9997**	**0.7213**	**0.8956**

The bold results are the best

for good stability and is suitable for the deployment in real applications.

Multimodal Methods We conduct comprehensive experiments on the full set of FakeAV and LAV-DF under both binary classification and four classification scenarios and report comparisons against state-of-the-art works in Tables 4 and 5. It is clear that our proposed method outperforms all compared methods, especially under AUC metrics by a large margin. In term of ACC metrics, all methods achieve comparable results. This is reasonable that the number of positive and negative samples in the two datasets is unbalanced, and the number of real samples is much smaller than the number of faked samples. In order to better analyze the experimental results, we also plotted the confusion matrices of the classification results of each method on different datasets. As shown in Fig. 6, both the FakeAV and LAV-DF datasets suffer from some degree of data imbalance. The sample imbalance problem can have a huge impact on the performance of different detection models that none of the baseline methods learn the features of those categories with low sample sizes well. however, our proposed model is almost immune to this problem. The reason behind is that de-homogenization graph pooling module designed by us can enhance the diversity of graph-level features, which facilitates mining the specific features for those categories with low sample sizes.

For FakeAV dataset, the performance of JAVDD which is specifically designed to fuse multimodal information at each layer is better than MDS and BA-TFD under four classification scenario. The reason behind is that both MDS and BA-TFD only fuse multimodal information in late fusion stage. However, different modalities have different convergence trends and simple information fusion does not yield more effective representation features, which causes subpar results. Although JAVDD can learn the correspondence between the audio and video over time, it only focuses on the overall correspondence of audio and video and ignores the local correspondence, which is insufficient to learn the subtleties of the intrinsic synchronisation patterns. This is also reflected in the fact that JAVDD does not perform well on the LAV-DF dataset, where all fake videos in the LAV-DF dataset are partially manipulated. In contrast, our method uses a heterogeneous graph model to model intra- and inter-modality relationships, and further promotes the local regional information interaction between the audio and video. Therefore, our method outperforms all compared opponents on all settings.

For LAV-DF dataset, all methods don't perform very well in four classification scenario. This is because the duration of the fake segment in each fake video is in the range of [0.8, 1.6] seconds. For fake videos with an average total duration of 4 s, the relatively small proportion of fake segments is a huge challenge for the detection method. Even then, our method still suppresses all compared competitors and achieve 5.51% and 3.04% performance gains than state-of-the-art results in terms of ACC and AUC metrics.

4.5.2 Inter-dataset Comparisons

In this section, we train the model on one dataset and test on another dataset to evaluate the model generalization. Comparisons under AUC metrics are shown in Table 6. We achieve 63.8% AUC on LAV-DF and 65.52% AUC on FakeAV dataset, exceeding the competitors by about 8% on average.

Table 4 Quantitative results (video-level ACC and AUC) for binary classification of multimodal methods on FakeAV and LAV-DF datasets

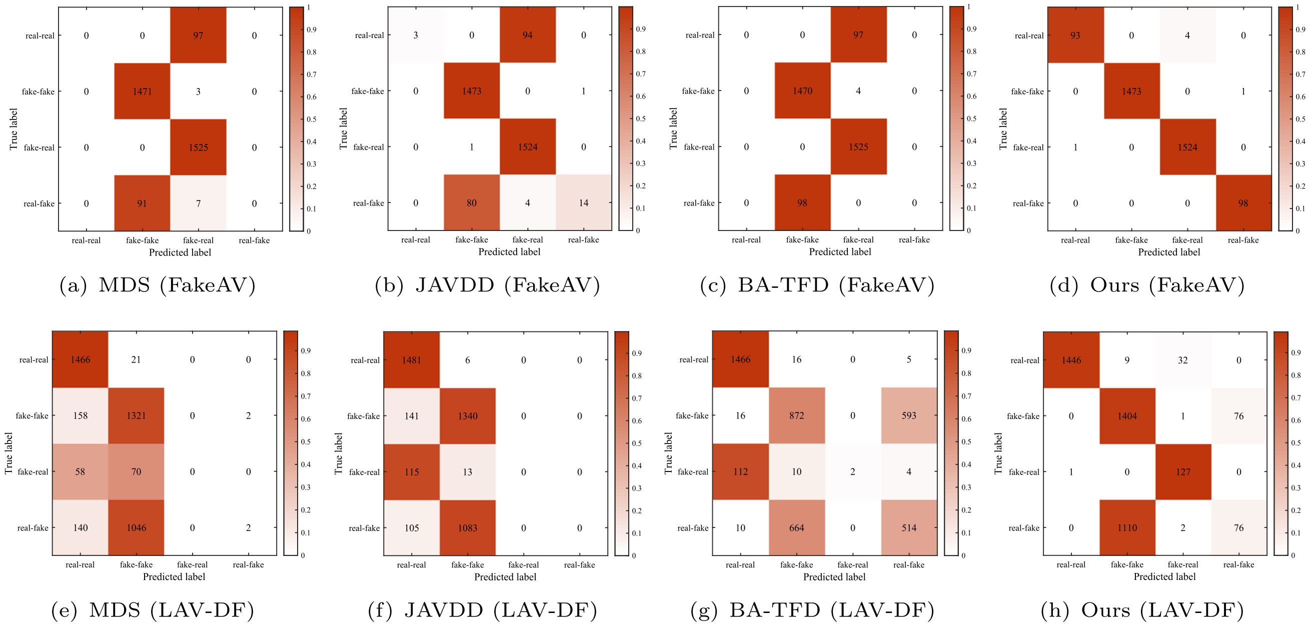
Methods	FakeAV		LAV-DF	
	ACC	AUC	ACC	AUC
MDS (Chugh et al., 2020)	0.9697	0.5193	0.9342	0.9783
JAVDD (Zhou and Lim, 2021)	0.9697	0.5204	0.9442	0.9691
BA-TFD (Cai et al., 2022)	0.9696	0.5431	0.6529	0.5097
Ours	**0.9984**	**0.9994**	**0.9981**	**0.9998**

The bold results are the best

Table 5 Quantitative results (video-level ACC and AUC) for four classification of multimodal methods on FakeAV and LAV-DF datasets

Methods	FakeAV		LAV-DF	
	ACC	AUC	ACC	AUC
MDS (Chugh et al., 2020)	0.9371	0.7318	0.6510	0.8091
JAVDD (Zhou and Lim, 2021)	0.9428	0.7778	0.6585	0.8271
BA-TFD (Cai et al., 2022)	0.9377	0.7757	0.6662	0.8652
Ours	**0.9991**	**0.9997**	**0.7213**	**0.8956**

The bold results are the best



(a) MDS (FakeAV) (b) JAVDD (FakeAV) (c) BA-TFD (FakeAV) (d) Ours (FakeAV)

(e) MDS (LAV-DF) (f) JAVDD (LAV-DF) (g) BA-TFD (LAV-DF) (h) Ours (LAV-DF)

Fig. 6 Confusion matrices of different model tested on different datasets under four classification scenarios. Each element in the confusion matrix represents the number of samples predicted to be in that class

Table 6 Comparisons (video-level AUC) on cross-dataset generalization under four classification scenario

Methods	LAV-DF	FakeAV
MDS (Chugh et al., 2020)	0.5522	0.5178
JAVDD (Zhou and Lim, 2021)	0.5811	0.6214
BA-TFD (Cai et al., 2022)	0.5511	0.5609
Ours	**0.6380**	**0.6552**

The bold results are the best

Since the videos in LAV-DF are partially forged, the detectors present relatively better generalization on FakeAV dataset compared to the LAV-DF dataset. Larger performance gains of 13.74% is obtained on FakeAV dataset. This is reasonable as our model explicitly takes advantage of spatial and temporal relationships between audio and visual modalities, which allows for a certain robustness of the proposed model.

Table 7 Robustness to common corruptions

Method	Clean	Saturation	Contrast	Block	Noise	Blur	JPEG
MDS (Chugh et al., 2020)	0.7318	0.7164	0.7114	0.7251	0.6662	0.6789	0.6655
JAVDD (Zhou and Lim, 2021)	0.7778	0.7469	0.7614	0.7768	0.7158	0.7315	0.7547
BA-TFD (Cai et al., 2022)	0.7757	0.7662	0.7584	0.7728	0.7312	0.7404	0.7387
Ours	0.9997	**0.9091**	**0.9107**	**0.9995**	**0.8598**	**0.8544**	**0.8703**

The bold results are the best

Average AUC scores (%) across five intensity levels for each corruption type

Table 8 Ablation study on key components

Pos	Cross	Pool	Binary classification		Four classification	
			ACC	AUC	ACC	AUC
×	×	×	0.9696	0.5502	0.9502	0.8161
✓	×	×	0.9696	0.5507	0.9512	0.8141
×	✓	×	0.9878	0.9216	0.9577	0.9010
×	×	✓	0.9809	0.9045	0.9549	0.8795
✓	✓	×	0.9912	0.9537	0.9840	0.9556
✓	×	✓	0.9904	0.9451	0.9853	0.9696
×	✓	✓	0.9937	0.9982	0.9906	0.9860
✓	✓	✓	**0.9984**	**0.9994**	**0.9991**	**0.9997**

The bold results are the best

4.6 Robustness to Common Corruptions

Given the ubiquity of post-processing operations on social media, it is critical that deployed multimodal forgery detectors are not easily subverted by common perturbations. We investigate the robustness of the detectors by training on original uncompressed FakeAV and then testing on FakeAV samples that were exposed to various unseen corruptions. The set of perturbations, proposed in Jiang et al. (2020), are changes in saturation and contrast, block-wise occlusions, Gaussian noise and blur, and JPEG compression (JPEG). Each perturbation type is applied at five different intensity levels. Table 7 presents the average AUC across all intensity levels for each corruption type. The proposed method still outperforms all competitors. The smaller performance degradation of all competitors is due to the fact that they originally initially failed to correctly distinguish between the different types of multimodal forgeries, as detailed in Fig. 6.

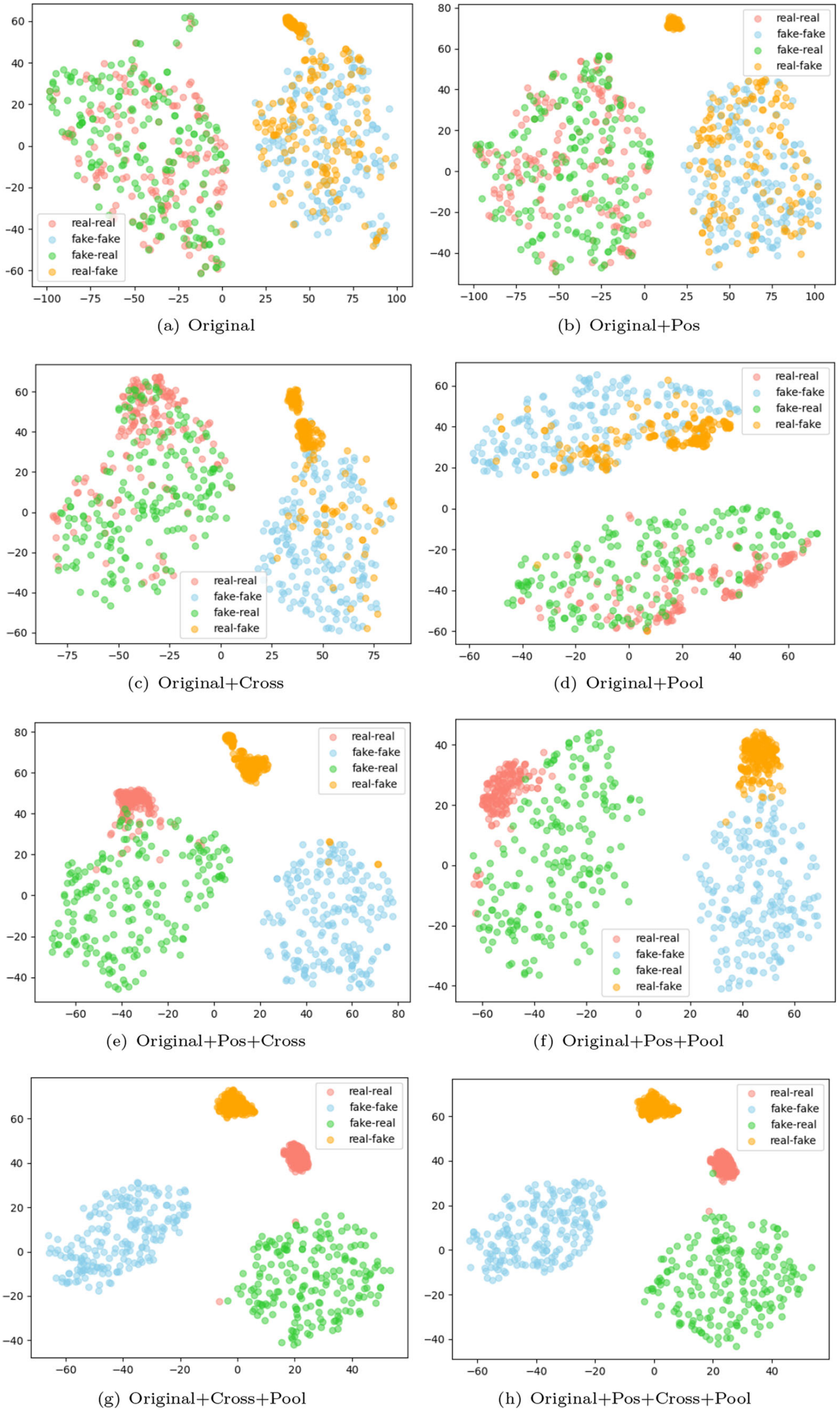
4.7 Ablation Study

4.7.1 Study on Module Effectiveness

We conduct comprehensive ablation studies on FakeAV dataset to further explore the effectiveness of the proposed method and modules, i.e., positional coding (Pos), cross-modal graph interaction module (Cross), and de-homogenization graph pooling module (Pool), as listed in Table 8. We observe from Table 8 that without introducing any proposed modules or methods, the original model has the poor performance. Inserting only Cross module or Pool module already improves the performance a lot. Obviously, both inter- and intra-modality information are vital and combination of them boosts the performance. The Pos method is an effective auxiliary approach that can further improve model performance. Note that the inter-modality information contributes to the improvement more, which again demonstrates the important of complementary information between modalities for fine-grained multimodal deepfake classification. We also use t-SNE (Van der Maaten and Hinton, 2008) to visualize the graph-level representations of the models with different key components. As shown in Fig. 7, the more key components a model has, the more discriminating its graph-level representations are. This also illustrates the effectiveness of the proposed method and modules.

4.7.2 Study on Graph Construction Parameters

To investigate the effect of the graph construction hyperparameter (*Audio-Visual neighbor*), we test our model with different hyperparameter value under four classification scenario on FakeAV dataset. As shown in Fig. 8, different *Audio-Visual neighbor* parameters will lead to different structure of the heterogeneous graph. It is noted that the *Audio-Visual overlap* changes as the *Audio-Visual neighbor* changes, aiming to follow the global correspondences of the audio and video nodes. As shown in Fig. 9, the performance of the proposed model increases as hyperparameter value increases. Performance improves up to 4 and then levels off. *Audio-Visual neighbor* = 4 means that 1000ms audio clip corresponding to 1600ms video clip. In other words, an appropriate increase in the inter-modal receptive field not only does not cause audio-visual synchronization problems, but also incorporates more inter-modal contextual information, which is good for performance. Meanwhile, due to the de-homogenization graph pooling operation, redundant inter-modal connections are removed and do not cause performance degradation.



(a) Original

(b) Original+Pos

(c) Original+Cross

(d) Original+Pool

(e) Original+Pos+Cross

(f) Original+Pos+Pool

(g) Original+Cross+Pool

(h) Original+Pos+Cross+Pool

Fig. 7 t-SNE visualization on feature space of graph-level representations

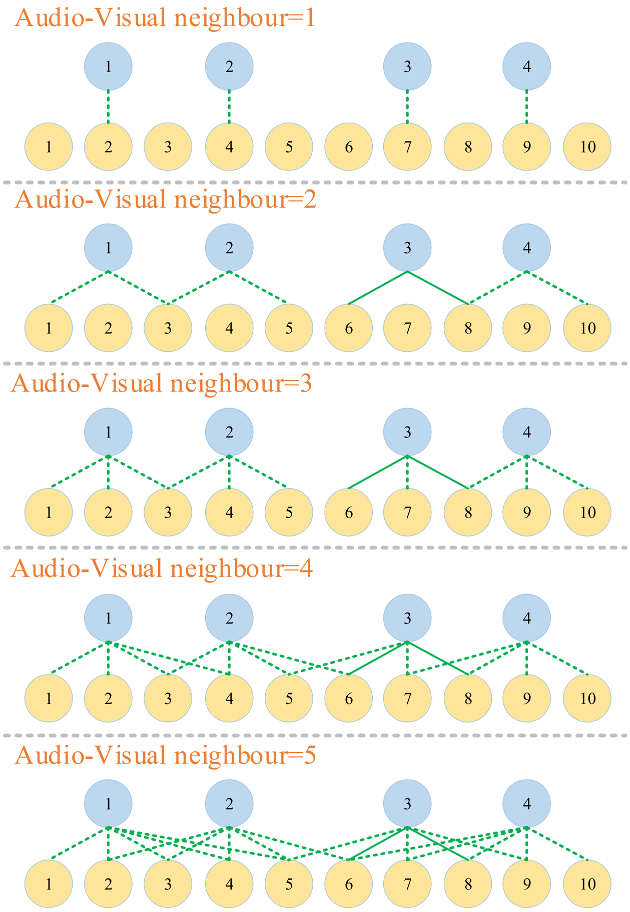


Fig. 8 Diagrammatic of the different structure of the heterogeneous graph caused by different *Audio-Visual neighbor* parameters

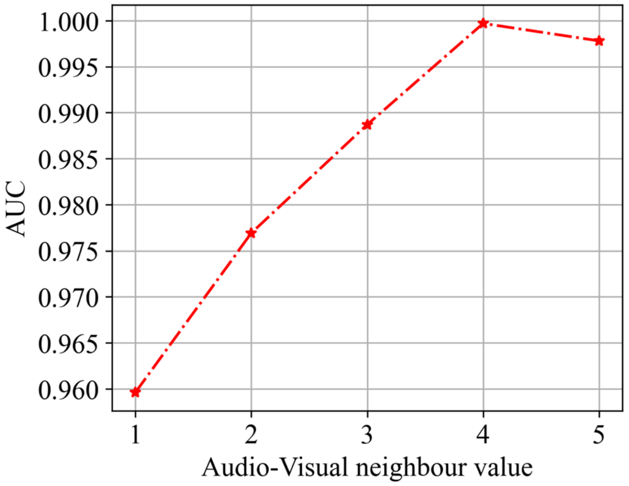


Fig. 9 Effect of using different graph construction hyperparameter values (*Audio-Visual neighbor*) on model performance

5 Conclusions

In this paper, we propose a novel task about fine-grained multimodal deepfake classification. This task is important because we don't have any prior knowledge about whether it is the video or audio that has been manipulated in practice. To this end, we propose a heterogeneous graph attention network that makes full use of intra- and inter-modality relationships. The graph attention network consists of a cross-modal graph interaction module for capturing complementary information between modalities and a de-homogenization graph pooling module for extracting modality-specific information. The proposed model presents superior performance and generalization on several benchmarks.

Acknowledgements This work was jointly supported by the National Natural Science Foundation of China (Grant No. U2001202, No. 62072480, No. 62172435).

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

Arik, S., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural voice cloning with a few samples. In *Advances in Neural Information Processing Systems*, (pp. 10040–10050).

Brissman, E., Johnander, J., Danelljan, M., & Felsberg, M. (2023). Recurrent graph neural networks for video instance segmentation. *International Journal of Computer Vision, 131*(2), 471–495.

Cai, Z., Stefanov, K., Dhall, A., & 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 (DICTA)* (pp. 1–10).

Cao, B., Bi, Z., Hu, Q., Zhang, H., Wang, N., Gao, X., & Shen, D. (2023). Autoencoder-driven multimodal collaborative learning for medical image synthesis. *International Journal of Computer Vision*, *131*(8), 1995–2014.

Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., & Yang, X. (2022). End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4113–4122).

Chen, S., Yao, T., Chen, Y., Ding, S., Li, J., & Ji, R. (2021). Local relation learning for face forgery detection. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 1081–1088).

Cheng, H., Guo, Y., Wang, T., Li, Q., Chang, X., & Nie, L. (2023). Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications and Applications, 20*(3), 1–22.

Chugh, K., Gupta, P., Dhall, A., & Subramanian, R. (2020). Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 439–447).

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

Fu, X., Qi, Q., Zha, Z. J., Ding, X., Wu, F., & Paisley, J. (2021). Successive graph convolutional network for image de-raining. *International Journal of Computer Vision, 129*, 1691–1711.

Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Huang, F., & Ma, L. (2021). Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3473–3481).

Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5039–5049).

Haliassos, A., Mira, R., Petridis, S., & Pantic, M. (2022). Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14950–14962).

Hua, G., Teoh, A. B. J., & Zhang, H. (2021). Towards end-to-end synthetic speech detection. *IEEE Signal Processing Letters, 28*, 1265–1269.

Jamaludin, A., Chung, J. S., & Zisserman, A. (2019). You said that?: Synthesising talking faces from audio. *International Journal of Computer Vision, 127*, 1767–1779.

Jia, Y., Zhang, Y., Weiss, R., Wang, Q., Shen, J., Ren, F., Nguyen, P., Pang, R., Lopez Moreno, I., & Wu, Y. (2018). Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems* (pp. 4485–4495).

Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C. (2020). Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2889–2898).

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2022). Countering malicious deepfakes: Survey, battleground, and horizon. *International Journal of Computer Vision, 130*(7), 1678–1734.

Khalid, H., Tariq, S., Kim, M.,& Woo, S. S. (2021). Fakeavceleb: A novel audio-video multimodal deepfake dataset. arXiv preprint arXiv:2108.05080.

Korshunova, I., Shi, W., Dambre, J., & Theis, L. (2017). Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 3677–3685).

Le, T. M., Le, V., Venkatesh, S., & Tran, T. (2021). Hierarchical conditional relation networks for multimodal video question answering. *International Journal of Computer Vision, 129*(11), 3027–3050.

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & and Guo, B. (2020a). Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5001–5010).

Li, X., Lang, Y., Chen, Y., Mao, X., He, Y., Wang, S., Xue, H., & Lu, Q. (2020b). Sharp multiple instance learning for deepfake video detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 1864–1872).

Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020c). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207–3216).

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., & Yu, N. (2021a). Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 772–781).

Liu, J., Rojas, J., Li, Y., Liang, Z., Guan, Y., Xi, N., & Zhu, H. (2021b). A graph attention spatio-temporal convolutional network for 3d human pose estimation in video. In *2021 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3374–3380). IEEE.

Lu, W., Liu, L., Zhang, B., Luo, J., Zhao, X., Zhou, Y., & Huang, J. (2023). Detection of deepfake videos using long-distance attention. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14. https://doi.org/10.1109/TNNLS.2022.3233063

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746–748.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2823–2832).

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint arXiv:1003.4083.

Peng, N., Poon, H., Quirk, C., et al. (2017). Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics, 5*, 101–115.

Peng, N., Poon, H., Quirk, C., et al. (2017). Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics, 5*, 101–115.

Ping, W., Peng, K., & Chen, J. (2018). Clarinet: Parallel wave generation in end-to-end text-to-speech. arXiv preprint arXiv:1807.07281.

Prajwal, K., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020). A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 484–492).

Qi, S., Wang, W., Jia, B., Shen, J., & Zhu, S. C. (2018). Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 401–417).

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision* (pp. 86–103). Springer.

Saqur, R., & Narasimhan, K. (2020). Multimodal graph networks for compositional generalization in visual question answering. *Advances in Neural Information Processing Systems, 33*, 3070–3081.

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., & Monfardini, G. (2008). The graph neural network model. *IEEE Transactions on Neural Networks, 20*(1), 61–80.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., & Saurous, R. A. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4779–4783). IEEE.

Tak, H., Jung, J. W., Patino, J., Kamble, M., Todisco, M., & Evans, N. (2021). End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. In *Proc. 2021 edition of the automatic speaker verification and spoofing countermeasures challenge* (pp. 1–8).

Tak, H., Todisco, M., Wang, X., Jung, J. W., Yamagishi, J., & Evans, N. (2022). Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. In *The speaker and language recognition workshop*.

Todisco, M., Delgado, H., & Evans, N. W. (2016). A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Odyssey* (pp. 283–290).

Tulyakov, S., Liu, M. Y., Yang, X., et al. (2018). Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1526–1535).

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research, 9*(11), 2579–2605.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. arXiv preprint arXiv:1710.10903.

Veyseh, A. P. B., Nguyen, T. H., & Dou, D. (2019). Graph based neural networks for event factuality prediction using syntactic and semantic structures. arXiv preprint arXiv:1907.03227.

Wang, Q., Wei, Y., Yin, J., Wu, J., Song, X., & Nie, L. (2021). Dualgnn: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia*, *25*, 1074–1084.

Wei, Y., Wang, X., Nie, L., He, X., Hong, R., & Chua, T. S. (2019). Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 1437–1445).

Wu, X., & Li, T. (2023). Sentimental visual captioning using multimodal transformer. *International Journal of Computer Vision, 131*(4), 1073–1090.

Yang, X., Feng, S., Zhang, Y., & Wang, D. (2021). Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 328–339).

Yin, Q., Lu, W., Li, B., & Huang, J. (2023). Dynamic difference learning with spatio-temporal correlation for deepfake video detection. *IEEE Transactions on Information Forensics and Security, 18*, 4046–4058.

Zhang, S., Qin, Y., Sun, K., & Lin, Y. (2019). Few-shot audio classification with attentional graph neural networks. In *Interspeech* (pp. 3649–3653).

Zhao, T., Xu, X., & Xu, M. (2020). Learning to recognize patch-wise consistency for deepfake detection. arXiv preprint arXiv:2012.09311.

Zheng, Y., Bao, J., Chen, D., Zeng, M., & Wen, F. (2021). Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 15044–15054).

Zhou, Y., & Lim, S. N. (2021). Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14800–14809).

Zhuang, W., Chu, Q., Tan, Z., Liu, Q., Yuan, H., Miao, C., Luo, Z., & Yu, N. (2022). Uia-vit: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In *European conference on computer vision* (pp. 391–407). Springer.

Zi, B., Chang, M., Chen, J., Ma, X., & Jiang, Y. G. (2020). Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 2382–2390).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.