Information Fusion 109 (2024) 102424

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Triple-modality interaction for deepfake detection on zero-shot identity

JunHo Yoon [a], Angel Panizo-LLedot [b], David Camacho [b], Chang Choi [a,*]

[a] Department of Computer Engineering, Gachon University, 1342, Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do, Republic of Korea
[b] Computer Systems Engineering Department, Universidad Politécnica de Madrid, Calle Alan Turing s/n, Madrid, 28031, Spain

ARTICLE INFO

Keywords:
Multi-modal
One-shot
Deepfake
Disinformation detection

ABSTRACT

Recent advancements in generative AI technology have created more realistic fake data that are utilized in various fields, such as data augmentation. However, the misuse of deepfake technology has led to increased damage. Consequently, ongoing research aims to analyze modality characteristics and detect deepfakes through AI-based methods. Existing AI-based deepfake-detection techniques have limitations in detecting deepfakes in modalities and identities that are not included in the training data. This study proposes a baseline approach based on zero-shot identity and one-shot deepfake detection for detecting deepfakes in environments with limited data. Additionally, we propose a triple-modality interaction based on a multimodal transformer (TMI-Former) to consider the triple-modality aspects of deepfakes. TMI-Former comprises four stages: vision feature extraction, representation, residual connection, and late-level fusion. It operates in a two-stage manner, extracting visual features and reconstructing them using auditory and linguistic features, thereby allowing for triple-modality interactions. In environments with limited data, such as zero-shot identity and one-shot deepfake scenarios, TMI-Former demonstrated effectiveness, with an accuracy ranging from 18.75% to 19.5% and an f1-score ranging from 0.2238 to 0.3561, compared to unimodal AI. Furthermore, TMI-Former shows superior performance compared to the existing multi-modal AI, with an accuracy ranging from 1.44% to 19.75% and an f1-score ranging from 0.0146 to 0.4169.

1. Introduction

Deepfake, a combination of deep learning and fake, refers to the synthesis of increasingly realistic fake data using advanced AI technologies, such as DALL-E and Chat-GPT. Although deepfake is utilized in various fields such as training data augmentation and virtual fitting, there is growing concern about its misuse, leading to identity fraud, sound spoofing, and spreading fake news [1]. Consequently, research is being conducted to analyze the characteristics of the modalities and detect deepfakes using AI-based approaches. However, AI has limitations in detecting deepfakes in modalities on which it has not been trained, which has led to the development of research using multi-modal AI to consider all available information [2]. Multi-modal AI is a technology that simultaneously learns different modalities such as vision, audio, text, biosignals, and metadata. It extracts features that consider information across modalities or complement the missing information from other modalities. Multi-modal AI-based systems have demonstrated effectiveness in various fields, including autonomous driving, biometric recognition, and medical image analysis, and are not limited to specific domains [3]. Furthermore, multi-modal AI can effectively detect deepfakes in single and multiple modalities by analyzing all the information involved in the deepfakes.

Multi-modal AI is a technology that simultaneously learns the modalities of different dimensions. This includes score-level fusion, feature-level fusion, and multi-modal transformers that utilize attention mechanisms to integrate models. Score-level fusion involves the use of separate AI models for each modality to extract probability values for the labels and then ensemble them in the final layer to select the ultimate label [4]. While score-level fusion integrates probability values to learn multiple modalities simultaneously, it requires sufficient computing resources because individual AI models are required for each modality. Feature-level fusion combines the features extracted from each modality and uses a single AI to select labels [5]. Feature-level fusion requires fewer computing resources than score-level fusion because it is based on a single AI that simultaneously learns multiple modalities. However, it has limitations in extracting a global context that considers information across modalities because it uses individual features. Multi-modal transformers integrate embedding vectors using a single AI and utilize attention mechanisms in the transformer encoder to consider information across modalities [6]. Multi-modal transformers address the limitations of computational resources and global context extraction. However, they require sufficient training data

* Corresponding author.
E-mail addresses: junho6257@gachon.ac.kr (J. Yoon), angel.panizo@upm.es (A. Panizo-LLedot), david.camacho@upm.es (D. Camacho), changchoi@gachon.ac.kr (C. Choi).

https://doi.org/10.1016/j.inffus.2024.102424
Received 11 January 2024; Received in revised form 31 March 2024; Accepted 12 April 2024
Available online 15 April 2024

owing to the lack of inductive bias. Obtaining sufficient training data for deepfake detection can be challenging owing to privacy concerns and potential misuse of personal information.

Existing research on multi-modal AI-based deepfake detection requires sufficient training data to detect deepfakes in both single and multiple modalities effectively and has limitations in detecting deepfakes from unseen identities [7]. To address the limitations of data dependency, models were developed based on the data type, specifically identity, and data augmentation was used. However, this approach requires significant amounts of training data and computing resources [8]. This study proposes a baseline approach based on zero-shot identity and one-shot deepfake detection to detect deepfakes in environments with limited data. Additionally, we propose a triple-modality interaction based on a multimodal transformer (TMI-Former) to consider the triple-modality aspects of deepfakes. Zero-shot identity evaluation assesses the generalization performance of deepfake detection by ensuring that the identities in the training, validation, and test sets do not overlap. The one-shot deepfake evaluation assesses the performance of deepfake detection in environments with limited data using real and fake sets of a single identity. TMI-Former comprises four stages: vision feature extraction, representation, residual connection, and late-level fusion. It operates in a two-stage manner, extracting visual features and reconstructing them using auditory and linguistic features, thereby allowing for triple-modality interactions. In environments with limited data, such as zero-shot identity and one-shot deepfake scenarios, TMI-Former demonstrated effectiveness, with an accuracy ranging from 18.75% to 19.5% and an f1-score ranging from 0.2238 to 0.3561, compared to unimodal AI. Furthermore, TMI-Former shows superior performance compared to the existing multi-modal AI, with an accuracy ranging from 1.44% to 19.75% and an f1-score ranging from 0.0146 to 0.4169.

2. Related work

Deepfake detection is based on AI and involves analyzing the characteristics of modalities to detect identity fraud, sound spoofing, and fake news. However, AI has limitations in detecting deepfakes in modalities on which it has not been trained, which has led to research using multi-modal AI to consider all available information. Multi-modal AI is a technology that simultaneously learns the modalities of different dimensions. This includes score-level fusion, feature-level fusion, and multi-modal transformers that utilize attention mechanisms to integrate models. Score-level fusion integrates models at the score level, where each modality uses individual AI models to extract probability values for the labels. These probabilities are combined in the final layer to select the final label. Feature-level fusion combines features extracted from each modality and uses a single AI model to select the labels. This approach requires fewer computing resources than score-level fusion; however, it has limitations in extracting a global context that considers information across modalities. Multi-modal transformers integrate embedding vectors using a single AI model and utilize attention mechanisms in the transformer encoder to consider information across modalities. This approach addresses the limitations of computational resources and global-context extraction. Multi-modal AI can effectively detect deepfakes in single and multiple modalities by analyzing all the information involved in deepfakes. The following sections explain AI-based deepfake detection and multi-modal learning methods.

2.1. AI-based deepfake detection

AI-based deepfake detection analyzes the characteristics of modalities based on AI, as listed in Table 1, and detects deepfakes in vision-based identity fraud and audio-based sound spoofing. Artem et al. [9] extracted and analyzed visual features, such as behavior and facial points, from facial images to detect deepfakes, achieving a validation AUC of 0.601. Farman et al. [10] extracted and analyzed auditory features, such as intonation and speaking speed, from audio to detect deepfakes, achieving a validation f1-score of 0.8268. Hasam et al. [11] proposed a deepfake detection system that extracts and analyzes modality features from facial images and audio and performs score-level fusion, achieving an accuracy of 67.4%. Irene et al. [12] detected deepfakes based on the optical flow of facial points extracted from facial videos and achieved an accuracy of 81.61%. AI-based deepfake detection systems can detect deepfakes effectively by extracting features from various modalities. However, there are limitations in detecting deepfakes in modalities that have not been trained, leading to research using multi-modal AI to consider all available information. Additionally, previous research has limitations in detecting deepfakes of untrained identities because they use overlapping identities of train, validation, and test sets. Therefore, research is needed for generalized deepfake detection that detects deepfakes in a zero-shot identity environment where identities are non-overlapping.

Table 1
Conventional research on AI-based deepfake detection method.

Dataset	Identity	Performance	Reference
Face Image	Overlap	AUC: 0.601	[9]
Audio	Overlap	F1-score: 0.8268	[10]
Face Image, Audio	Overlap	Accuracy: 0.674	[11]
Face Video	Overlap	Accuracy: 0.8161	[12]

2.2. Multi-modal learning method

Multi-modal AI is a technology that simultaneously learns the modalities of different dimensions. This includes score-level fusion, feature-level fusion, and multi-modal transformers that utilize attention mechanisms to integrate models, as listed in Table 2. Existing research on multi-modal AI-based deepfake detection requires sufficient training data to detect deepfakes in both single and multiple modalities effectively and has limitations in detecting deepfakes from unseen identities. To address the limitations of data dependency, models were developed based on the data type, specifically identity, and data augmentation was used. However, this approach requires considerable training and computing resources [20].

Score-level fusion involves using separate AI models for each modality to extract and ensemble the probability values for the labels [21]. Kmael et al. [13] proposed a biometric identification system that performs score-level fusion of iris images and fingerprint images, achieving a performance with an accuracy of 95%, which is superior by 9.44% to 18.33% compared to unimodal approaches. Sumeght et al. [14] proposed a biometric recognition system that performs a score-level fusion of 3D faces and 3D fingerprints, achieving a performance with an accuracy of 99.25%, which is superior by 12.89% to 35.81% compared to uni-modal approaches. Score-level fusion integrates probability values to learn multiple modalities simultaneously but requires sufficient computing resources, as individual AI models are required for each modality.

Feature-level fusion combines the features extracted from each modality and uses a single AI model to determine the labels [22]. Additionally, feature-level fusion includes early and late fusion depending on the level at which features are combined, and hybrid fusion involves fusion at both levels. Yikai et al. [15] proposed an early feature fusion-based semantic segmentation system that combines RGB images and depth images before the classification head, and with a performance of IoU 51.2, it is 4.7 to 16.9 better than uni-modal. Yagya Raj et al. [16] proposed a late feature fusion-based emotion recognition system that combines vision and audio after classification, and the performance of f1-score 0.88 is 0.35 to 0.51 superior to uni-modal. Machen et al. [17] proposed a hybrid feature fusion-based emotion recognition system that fuses audio and text at early and late levels, and the performance of

Table 2
Conventional research on multi-modal learning method: Uni-modal performance results from distinguishing data types in order, while multi-modal performance results from training all data types simultaneously.

Method	Dataset	Task	Uni-modal	Multi-modal	Reference
Score-level fusion	Iris Image, Fingerprint Image 3D Face, 3D Ear	Biometric identification Biometric recognition	Accuracy: 0.868/0.7667 Accuracy: 0.6344/0.8636	Accuracy: 0.95 Accuracy: 0.9925	[13] [14]
Early feature fusion	RGB Image, Depth Image	Semantic segmentation	IoU: 46.5/34.3	IoU: 51.2	[15]
Late feature fusion	Vision, Audio	Emotion recognition	F1-score: 0.53/0.37	F1-score: 0.88	[16]
Hybrid feature fusion	Audio, Text	Emotion recognition	F1-score: 0.3126/0.543	F1-score: 0.6628	[17]
Multi-modal transformer	Visual, Audio Skin Image, Metadata	Emotion recognition Disease classification	Accuracy: 0.602/0.471 F1-score: 0.716/0.788	Accuracy: 0.629 F1-score: 0.82	[18] [19]

f1-score 0.6628 is 0.1198 to 0.3502 superior to uni-modal. Feature-level fusion requires fewer computing resources than score-level fusion because it is based on a single AI model that simultaneously learns multiple modalities. However, it has limitations in extracting a global context that considers information across modalities because it uses individual features.

Multi-modal transformers integrate embedding vectors using a single AI model and utilize attention mechanisms in the transformer encoder to consider information across modalities [23]. Jian et al. [18] proposed an emotion recognition system based on a multi-modal transformer that learns vision and audio simultaneously, achieving an accuracy of 62.9%, which is superior by 2.7% to 15.8% compared to uni-modal approaches. Gan et al. [19] proposed a disease classification system based on a multi-modal transformer that simultaneously learns skin images and metadata, achieving a performance with an f1-score of 0.82, which is superior by 0.032 to 0.104 compared to uni-modal approaches. Multi-modal transformers address the limitations of computational resources and global context extraction. However, they require sufficient training data owing to the lack of inductive bias. Obtaining sufficient training data for deepfake detection can be challenging owing to privacy concerns and potential misuse of personal information [24].

3. Deepfake detection with TMI-Former

This study proposes TMI-Former, which extracts visual, audio, and text information from videos to detect deepfakes in both single and multiple modalities. TMI-Former comprises four stages: vision feature extraction, representation, residual connection, and late-level fusion, as shown in Fig. 2. It employs a two-stage approach where visual features are extracted and reconstructed through interaction with auditory and textual features. Visual features were extracted from the input videos during the vision-feature extraction stage. These features are then transformed into a common latent space during the representation stage. The residual connection stage helps to preserve the original visual information while incorporating the transformed features. Finally, in the late-level fusion stage, triple-modal features were combined to capture the interactions between vision, audio, and text. The following section explains modality extraction and embedding and the triple-modality interaction on TMI-Former, focusing on their role in detecting deepfakes.

3.1. Modality extract and embedding

This study utilizes multi-modal AI to detect deepfakes in single and multiple modalities and evaluates its performance using vision, audio, and text information extracted from deepfakes. For evaluation, the data used were extracted and embedded into one visual, audio, and text information per video, as shown in Fig. 1. Video interpolation analyzes previous and subsequent frames to generate intermediate frames that maintain the object's temporal characteristics [25]. In this paper, to extract visual information while considering the continuity of the video, the middle frame is extracted and used from T frames (average about 7.8 s), like video interpolation. Auditory information is transformed

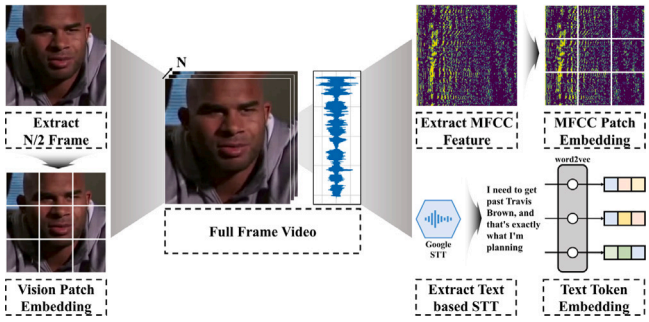

Fig. 1. Modality extract and embedding from video.

from a 1D waveform representing the amplitude over time to a 2D mel-frequency cepstral coefficient (MFCC) representation that reflects frequency information. Unlike the waveform, the MFCC captures amplitude variations over time and frequency, making it invariant to domain and environmental conditions, thus effectively generalizing deepfake detection performance [26]. The language information is extracted using the Google Speech-to-Text API, and data with a similarity score of 0.9 or higher are used, indicating a high level of confidence. The extracted visual and auditory information was divided into 16×16 patches, resulting in 256 patches, which were embedded as 2D representations of size 256×256. Language information is embedded into 256 tokens that preserve the relationships between words based on word2vec [27]. Auditory and language information extracted from the video is co-learned with the visual information of a single frame, allowing the multi-modal AI-based baseline and proposed method to detect temporal information in the video.

3.2. Triple-modality interaction on TMI-Former

The proposed TMI-Former consists of four stages: visual feature extraction, representation, residual connection, and late-level fusion. In the vision feature extraction stage, the vision features extracted from the output layer of the vision transformer are used as multi-modal class tokens, as shown in Eq. (1). The label information is utilized as a multi-modal distillation token, as shown in Eq. (2). The class token generated in the vision feature extraction step performs joint learning of the vision feature with other modalities, and the distillation token distills the label information of the vision during co-learning In the representation stage, the generated class and distillation tokens are reconstructed with the audio data as the class token and embedding vector, respectively, followed by the distillation token, as shown in Eq. (3). Similarly, the text data are reconstructed as shown in Eq. (4). The reconstructed embedding vectors add positional encoding, which is the order information of the input sequence then used to extract the global context through the attention mechanism of the transformer encoder, thereby facilitating the interaction between the triple-modalities. In the triple-modality interaction process, the class token effectively detects

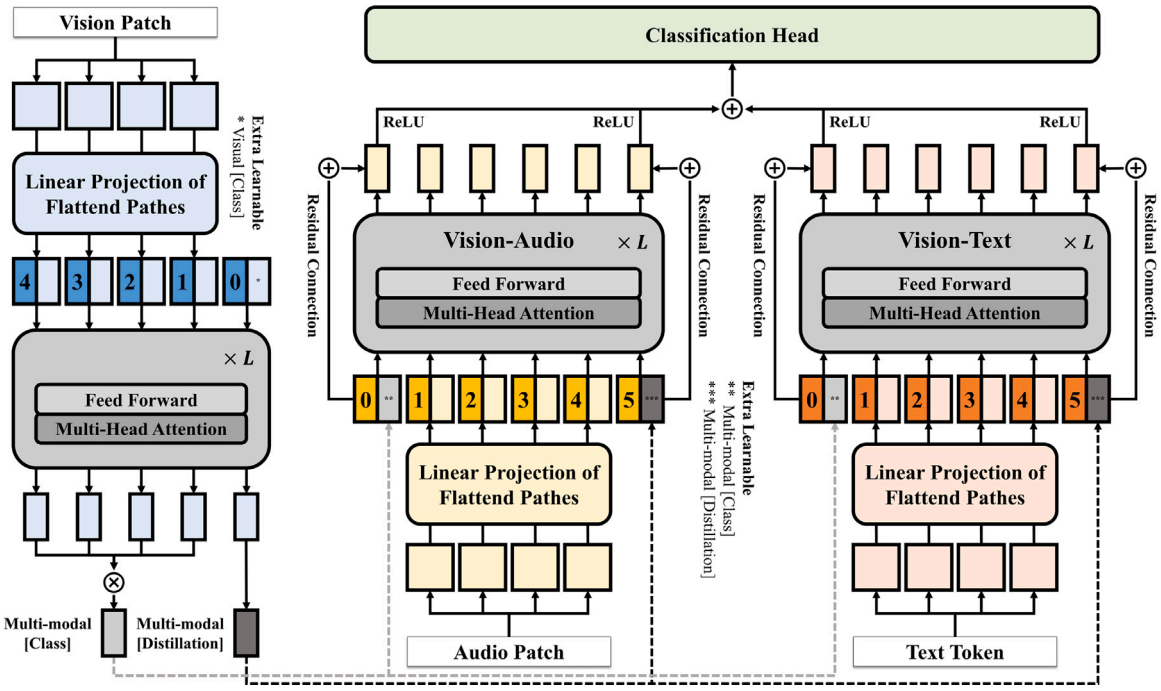

Fig. 2. Architecture of TMI-Former: TMI-Former consists of a vision feature extraction stage on the left, a residual connection stage connecting the input and output of the vision-audio and vision-text transformer, and a late-level fusion stage for classification.

deepfakes by reflecting audio and text's feature values and distillation information

$$M_{Class} = mean(V_L^1, \ldots, V_L^N) \tag{1}$$

$$M_{Distillation} = V_L^0 \tag{2}$$

$$VA_0 = [M_{Class}; A_L^1; A_L^2; \cdots; A_L^N; M_{Distillation}] + E_{pos} \tag{3}$$

$$VT_0 = [M_{Class}; T_L^1; T_L^2; \cdots; T_L^N; M_{Distillation}] + E_{pos} \tag{4}$$

$$\begin{aligned} VA_{Class} &= ReLU(M_{class} + VA_L^0) \\ VA_{Distillation} &= ReLU(M_{Distillation} + VA_L^{N+1}) \end{aligned} \tag{5}$$

$$\begin{aligned} VT_{Class} &= ReLU(M_{class} + VT_L^0) \\ VT_{Distillation} &= ReLU(M_{Distillation} + VT_L^{N+1}) \end{aligned} \tag{6}$$

The residual connection stage aims to prevent the loss of visual features during a triple-modality interaction. To achieve this, the class and distillation tokens from the input and output layers of the vision-audio transformer were connected, as shown in Eq. (5). In addition, rectified linear unit (ReLU) activation was applied. Similarly, the class and distillation tokens from the input and output layers were connected to the vision-text transformer. ReLU activation was applied to prevent the loss of visual features, as shown in Eq. (6). The ReLU returns the input value when it is positive and returns 0 when it is negative, thereby preventing the loss of visual features during the interaction of the triple modalities [28]. In the late-level fusion stage, the multi-modal class and distillation tokens extracted from TMI-Former are fused using a classifier to obtain probability values for detecting deepfakes. This fusion compensates for the lack of information in the class token by utilizing a distillation token, as described in Algorithm 1: Overall, the proposed TMI-Former effectively extracts the global context through the interaction of triple-modalities in the vision feature extraction and representation stages based on a two-stage approach. The residual connection and late-level fusion stages prevent information loss during the interaction of triple modalities and complement the information in the class token with the distillation token, resulting in effective deepfake detection.

Algorithm 1 Late-level Fusion step of the TMI-Former

Require: VA_{Class}, $VA_{Distillation}$, VT_{Class}, $VT_{Distillation}$
VA Class = Classifier(VA_{Class})
VA Distillation = Classifier($VA_{Distillation}$)
Vision-Audio Output = mean(VA Class, VA Distillation)

VT Class = Classifier(VT_{Class})
VT Distillation = Classifier($VT_{Distillation}$)
Vision-Text Output = mean(VT Class, VT Distillation)

Output = mean(Vision-Audio Output, Vision-Text Output) **return** Output

4. Experiment result

In this study, TMI-Former evaluates the deepfake detection performance in a data-limited environment, specifically zero-identity and one-shot deepfake scenarios. The experimental data were based on FakeAVCeleb, where the training, validation, and test sets had nonoverlapping identities, and the real and fake sets within each identity were used as a single set. The TMI-Former was compared and evaluated against uni-modal and multi-modal AI models by leveraging the interaction between modalities in videos in the zero-shot identity and one-shot deepfake baselines. The AI models used in the experiments were tested using the model weights that achieved the highest validation accuracy over 100 epochs. The following section explains the baseline and experimental environments and deepfake detection with modality interaction.

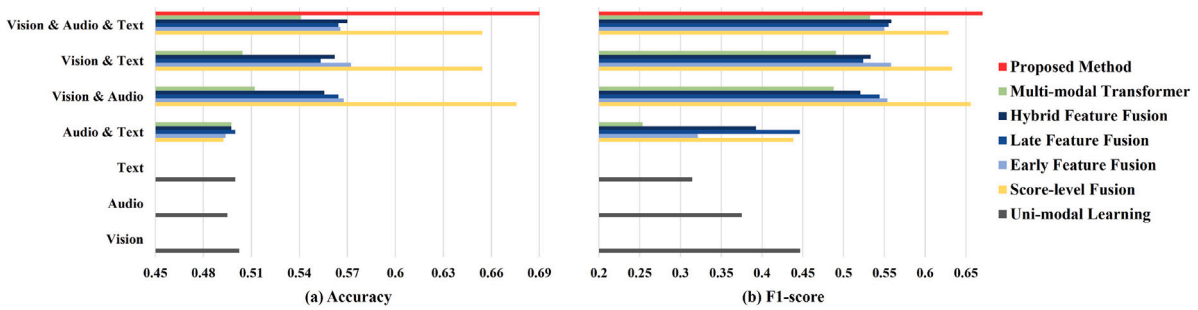

Fig. 3. Performance of deepfake detection according to uni-modal and multi-modal AI.

Table 3
Label definition and number on dataset: The number of modalities and identities for vision, audio, and text is the same as the number of datasets in each task.

Label	Vision type	Audio type	Train	Validation	Test
Real	Real	Real	495	45	45
Fake	Real	Fake	135	15	15
Fake	Fake	Real	135	15	15
Fake	Fake	Fake	135	15	15

4.1. Baseline and experiment environment

In this study, to demonstrate the effectiveness of TMI-Former in a data-limited environment, a baseline is proposed to evaluate the generalized deepfake detection performance without relying on specific training data. The baseline focuses on zero-shot identity and one-shot deepfake scenarios. At baseline, TMI-Former is compared and evaluated against the backbone transformer-based uni-modal AI and existing multi-modal AI models, such as score-level fusion, early, late, and hybrid feature fusion of feature-level fusion, and multi-modal transformer. The AI models used in the experiments were trained and validated for 100 epochs in a Python 3.8 environment, utilizing an Intel Core i9 processor and an NVIDIA TITAN RTX @ 24.0 GB memory. The model weights that achieved the highest validation accuracy were used for the testing. The evaluation metrics for assessing deepfake detection performance are accuracy, defined in Eq. (7), and f1-score, defined in Eq. (8), based on the four cases. These metrics are commonly used to evaluate the performance of classification models [29].

- TP(True Positive) : Predict samples with a positive label as positive
- FP(False Positive) : Predict samples with negative label as positive
- TN(True Negative) : Predict samples with negative label as negative
- FN(False Negative) : Predict samples with a positive label as negative

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$F1\ Score = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{8}$$

The deepfake dataset used in the experiments is based on FakeAVCeleb [33], which consists of deepfakes created from 500 speech videos. The dataset was reconstructed to evaluate the deepfake detection performance in zero-shot identity and one-shot deepfake scenarios. When extracting language information, real and fake data from 495 individuals with a confidence score of 0.9 or higher are utilized, as listed in Table 3, and if a specific modality is determined to be a deepfake, the label is assigned as fake. In the zero-shot identity scenario, the training, validation, and test sets had nonoverlapping identities, allowing for the evaluation of deepfake detection for unseen identities. In the one-shot deepfake scenario, the real and fake sets within each identity were used as a single set, enabling the evaluation of deepfake detection in a data-limited environment. To evaluate the generalized deepfake detection performance, k-fold cross-validation was employed because the dataset used in this study had nonoverlapping identities for each task. The performance of the generalized deepfake detection was evaluated based on the 10 cases, aiming to prevent underfitting owing to limited training data [34]. Of the 10 sets in which identities are divided non-overlap, 9 are used as train sets, the remaining 1 is used as validation sets, and the average and standard deviation according to 10 cases is measured. The 10 k-folded cases evaluate generalized deepfake detection performance by intersecting the identities of train and validation in a non-overlapping manner and testing the deepfake detection performance of the fixed identity.

4.2. Deepfake detection with modality interaction

The results of comparing and evaluating TMI-Former with existing methods using zero-shot identity and one-shot deepfake are shown in Fig. 3 and Table 4. The accuracy of the unimodal AI ranged from 49.5% to 50.25%, with an f1-score of 0.3143 to 0.4466, indicating poor performance in detecting deepfakes. The existing multi-modal AI shows a higher accuracy ranging from 0.19% to 18.06% and an f1-score ranging from 0.0412 to 0.3415 compared to uni-modal AI, demonstrating better deepfake detection performance. However, existing multi-modal AI fails to correctly detect vision-based deepfakes because it does not consider visual features, resulting in a degradation of deepfake detection performance compared to uni-modal AI. Score-level fusion exhibits the highest deepfake detection performance among the existing multi-modal AI methods, as it individually analyzes the probability values for real and fake images for each modality. However, the multi-modal transformer shows the lowest deepfake detection performance among existing multi-modal AI methods, as it does not consider modalities individually, indicating a lack of inductive bias. However, as the number of modalities increases, the deepfake detection performance improves. Feature-level fusion, which extracts individual features, shows superior deepfake detection performance compared with multi-modal transformers. However, it has limitations in terms of performance compared to score-level fusion because it integrates and analyzes real and fake detections at a later stage. Ultimately, TMI-Former demonstrated effective deepfake detection performance. Compared to unimodal AI, it achieves an accuracy of 18.75% to 19.5% and an f1-score of 0.2238 to 0.3561. Compared to existing multi-modal AI, it achieves an accuracy of 1.44% to 19.75% and an f1-score of 0.0146 to 0.4169, showing superior performance.

4.3. Ablation study

The proposed TMI-Former comprises four stages: visual feature extraction, representation, residual connection, and late-level fusion. It operates in a two-stage manner in which visual features are extracted and reconstructed using auditory and linguistic features, enabling

Table 4
Performance of deepfake detection according to uni-modal and multi-modal AI: In deepfake detection performance, the first row represents the average when k-fold is performed, and the value in parentheses represents the standard deviation.

Method	Modality	Accuracy	F1-score	Method	Modality	Accuracy	F1-score
Uni-modal learning	Vision [30]	0.5025 (±0.0175)	0.4466 (±0.2926)	Multi-modal transformer [18]	Vision & Audio	0.5122 (±0.0304)	0.4878 (±0.0493)
	Audio [31]	0.495 (±0.017)	0.375 (±0.3067)	Score-level fusion [13]	Vision & Text	0.6544 (±0.0202)	0.644 (±0.0325)
	Text [32]	0.5 (±0.0177)	0.3143 (±0.3149)	Early feature fusion [15]		0.5722 (±0.0437)	0.558 (±0.0451)
Score-level fusion [13]	Audio & Text	0.4925 (±0.016)	0.4385 (±0.2874)	Late feature fusion [16]		0.5533 (±0.0557)	0.524 (±0.0769)
Early feature fusion [15]		0.4938 (±0.0151)	0.3215 (±0.3217)	Hybrid feature fusion [17]		0.5622 (±0.0439)	0.5334 (±0.0754)
Late feature fusion [16]		0.5 (±0.0194)	0.4463 (±0.2927)	Multi-modal transformer [18]		0.5044 (±0.0315)	0.4905 (±0.0331)
Hybrid feature fusion [17]		0.5075 (±0.017)	0.3927 (±0.3208)	Score-level fusion [13]	Vision & Audio & Text	0.6544 (±0.0175)	0.6288 (±0.0191)
Multi-modal transformer [18]		0.4975 (±0.0192)	0.2535 (±0.3108)	Early feature fusion [15]		0.5656 (±0.0363)	0.5499 (±0.0319)
Score-level fusion [13]	Vision & Audio	0.6756 (±0.0227)	0.6558 (±0.0224)	Late feature fusion [16]		0.5644 (±0.0333)	0.5554 (±0.0366)
Early feature fusion [15]		0.5678 (±0.0356)	0.5537 (±0.0371)	Hybrid feature fusion [17]		0.57 (±0.0258)	0.5586 (±0.0421)
Late feature fusion [16]		0.5644 (±0.0276)	0.5443 (±0.0332)	Multi-modal transformer [18]		0.5411 (±0.233)	0.5325 (±0.0264)
Hybrid feature fusion [17]		0.5556 (±0.0436)	0.5208 (±0.0756)	**TMI-Former**		0.69* (±0.0279)	0.6704* (±0.0339)

Table 5
Performance of TMI-Former according to modality interaction.

Method	Modality	Accuracy	F1-score
w/o Vision-Text	Vision & Audio	0.6622 (±0.0254)	0.6373 (±0.0335)
w/o Vision-Audio	Vision & Text	0.6722 (±0.0224)	0.6516 (±0.0288)
TMI-Former	Vision & Audio & Text	0.69* (±0.0279)	0.6704* (±0.0339)

Table 6
Performance of TMI-Former according to residual connection.

Layer	Residual connection	Accuracy	Difference (Accuracy)	F1 score	Difference (F1-score)
2	False	0.64 (±0.0431)	0.0233	0.61 (±0.0543)	0.0277
	True	0.6633 (±0.0322)		0.6377 (±0.0458)	
4	False	0.6478 (±0.0254)	0.0089	0.6174 (±0.0419)	0.0124
	True	0.6567 (±0.0328)		0.6298 (±0.0443)	
6	False	0.6111 (±0.0362)	0.0533	0.5721 (±0.0476)	0.0663
	True	0.6644 (±0.028)		0.6384 (±0.0374)	
8	False	0.6144 (±0.0244)	0.0756*	0.5694 (±0.0417)	0.101*
	True	0.69* (±0.0279)		0.6704* (±0.0339)	

triple-modality interactions. The deepfake detection performance when the vision-text interaction and vision-audio interaction of TMI-Former were removed is shown in Table 5. When all triple-modalities interacted, the deepfake detection performance was superior, with an accuracy ranging from 1.78% to 2.78% and an f1-score ranging from 0.0188 to 0.0331. The evaluation of the residual connection stage in TMI-Former, which prevents the loss of visual features, is presented in Table 6. As the layers deepened, the accuracy ranged from 0.89% to 7.56%, and the f1-score ranged from 0.0124 to 0.101, indicating the prevention of information loss.

5. Conclusion

With recent advancements in generative AI technologies, such as DALL-E and Chat-GPT, the production of highly realistic deepfakes has become more prominent. Deepfakes are utilized in various fields, such as data augmentation and virtual fitting. However, they also cause increased identity fraud, sound spoofing, and the spread of fake news, leading to harmful consequences. Therefore, research is being conducted to analyze the characteristics of the modalities and detect deepfakes using AI-based approaches [35]. However, AI-based deepfake detection methods that rely solely on training data have limitations in detecting deepfakes in modalities that have not been learned. To address this issue, research is being conducted using multi-modal AI, which considers all the available information [36]. Multi-modal AI encompasses techniques such as score-level fusion, which integrates models at the scoring level; feature-level fusion, which integrates features from different modalities; and multi-modal transformers, which employ attention mechanisms. Existing research on deepfake detection based on multi-modal AI detects deepfakes using single and multiple modalities. However, to achieve high deepfake detection performance, a sufficient amount of training data is required, and there are limitations to detecting deepfakes of identities that have not been trained [37]. Obtaining deepfake data is challenging because of privacy concerns and potential misuse; developing models based on specific identities and utilizing data augmentation to overcome the limitations of data dependency ultimately requires ample training data and computing resources [38].

In this study, we propose a baseline for deepfake detection in environments with limited data by introducing zero-shot identity and one-shot deepfake approaches. The zero-shot identity evaluates the generalized deepfake detection performance by ensuring non-overlapping identities, whereas the one-shot deepfake identity evaluates the performance in environments with limited data by using a single set of real and fake identities. To consider a triple modality, we propose TMI-Former. TMI-Former consists of vision feature extraction, representation, residual connection, and late-level fusion stages, enabling the interaction of the triple modalities involved in deepfakes in a two-stage manner in the vision feature extraction and representation stages. The residual connection and late-level fusion stages prevent information loss that may occur during the interaction of triple

modalities and effectively detect deepfakes by complementing missing information through class tokens and distillation tokens. We evaluated TMI-Former using zero-shot identity and one-shot deepfake against existing methods. The results show that TMI-Former achieves higher accuracy (18.75% to 19.5%) and f1-score (0.2238 to 0.3561) than unimodal AI. TMI-Former demonstrated superior accuracy (1.44% to 19.75%) and f1-score (0.0146 to 0.4169) compared to the existing multi-modal AI. In the future, TMI-Former can be extended to downstream tasks through pretraining on large-scale video datasets such as UCF101 and Kinetics. It can also be applied to explainable AI by utilizing semantic relationships between modalities for tasks such as scene graph generation and situational awareness [39].

CRediT authorship contribution statement

JunHo Yoon: Writing – original draft, Visualization, Validation, Methodology, Investigation, Data curation, Conceptualization. **Angel Panizo-LLedot:** Writing – review & editing, Validation. **David Camacho:** Writing – review & editing, Validation. **Chang Choi:** Writing – review & editing, Supervision, Software, Resources, Project administration, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Chang Choi reports financial support was provided by Gachon University. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (2021R1A2B5B02087169). This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (RS-2023-00259004) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation)

References

[1] M.S. Rana, M.N. Nobi, B. Murali, A.H. Sung, Deepfake detection: A systematic literature review, IEEE Access 10 (2022) 25494–25513.
[2] M. Lomnitz, Z. Hampel-Arias, V. Sandesara, S. Hu, Multimodal approach for deepfake detection, in: 2020 IEEE Applied Imagery Pattern Recognition Workshop, AIPR, IEEE, 2020, pp. 1–9.
[3] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, A. Hussain, Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Inf. Fusion 91 (2023) 424–444.
[4] G.H. Supreetha, H.G. Kumar, M. Imran, Multimodal biometric verification system: Evaluation of various score level fusion rules, in: 2019 IEEE International Conference on Electrical, Computer and Communication Technologies, ICECCT, IEEE, 2019, pp. 1–4.
[5] H. Cai, Z. Qu, Z. Li, Y. Zhang, X. Hu, B. Hu, Feature-level fusion approaches based on multimodal EEG data for depression recognition, Inf. Fusion 59 (2020) 127–138.
[6] Y. Khare, V. Bagal, M. Mathew, A. Devi, U.D. Priyakumar, C. Jawahar, Mmbert: Multimodal bert pretraining for improved medical vqa, in: 2021 IEEE 18th International Symposium on Biomedical Imaging, ISBI, IEEE, 2021, pp. 1033–1036.
[7] T.T. Nguyen, Q.V.H. Nguyen, D.T. Nguyen, D.T. Nguyen, T. Huynh-The, S. Nahavandi, T.T. Nguyen, Q.-V. Pham, C.M. Nguyen, Deep learning for deepfakes creation and detection: A survey, Comput. Vis. Image Underst. 223 (2022) 103525.
[8] S. Das, S. Seferbekov, A. Datta, M.S. Islam, M.R. Amin, Towards solving the deepfake problem: An analysis on improving deepfake detection using dynamic face augmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3776–3785.
[9] A.A. Maksutov, V.O. Morozov, A.A. Lavrenov, A.S. Smirnov, Methods of deepfake detection based on machine learning, in: 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, EIConRus, IEEE, 2020, pp. 408–411.
[10] F. Hassan, A. Javed, Voice spoofing countermeasure for synthetic speech detection, in: 2021 International Conference on Artificial Intelligence, ICAI, IEEE, 2021, pp. 209–212.
[11] H. Khalid, M. Kim, S. Tariq, S.S. Woo, Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors, in: Proceedings of the 1st Workshop on Synthetic Multimedia-Audiovisual Deepfake Generation and Detection, 2021, pp. 7–15.
[12] I. Amerini, L. Galteri, R. Caldelli, A. Del Bimbo, Deepfake video detection through optical flow based cnn, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.
[13] K. Aizi, M. Ouslim, Score level fusion in multi-biometric identification based on zones of interest, J. King Saud Univ. Comput. Inf. Sci. 34 (1) (2022) 1498–1509.
[14] S. Tharewal, T. Malche, P.K. Tiwari, M.Y. Jabarulla, A.A. Alnuaim, A.M. Mostafa, M.A. Ullah, et al., Score-level fusion of 3D face and 3D ear for multimodal biometric human recognition, Comput. Intell. Neurosci. 2022 (2022).
[15] Y. Wang, F. Sun, M. Lu, A. Yao, Learning deep multimodal feature representation with asymmetric multi-layer fusion, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 3902–3910.
[16] Y.R. Pandeya, J. Lee, Deep learning-based late fusion of multimodal information for emotion classification of music video, Multimedia Tools Appl. 80 (2) (2021) 2887–2905.
[17] M. Luo, H. Phan, J. Reiss, Cross-modal fusion techniques for utterance-level emotion recognition from text and speech, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2023, pp. 1–5.
[18] J. Huang, J. Tao, B. Liu, Z. Lian, M. Niu, Multimodal transformer fusion for continuous emotion recognition, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2020, pp. 3507–3511.
[19] G. Cai, Y. Zhu, Y. Wu, X. Jiang, J. Ye, D. Yang, A multimodal transformer to fuse images and metadata for skin disease classification, Vis. Comput. 39 (7) (2023) 2781–2793.
[20] T. Zhang, Deepfake generation and detection, a survey, Multimedia Tools Appl. 81 (5) (2022) 6259–6276.
[21] J. Amin, M. Sharif, M. Yasmin, T. Saba, M.A. Anjum, S.L. Fernandes, A new approach for brain tumor segmentation and classification based on score level fusion using transfer learning, J. Med. Syst. 43 (2019) 1–16.
[22] M. Amini, M. Nazari, I. Shiri, G. Hajianfar, M.R. Deevband, H. Abdollahi, H. Arabi, A. Rahmim, H. Zaidi, Multi-level multi-modality (PET and CT) fusion radiomics: prognostic modeling for non-small cell lung carcinoma, Phys. Med. Biol. 66 (20) (2021) 205017.
[23] Y.-H.H. Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.-P. Morency, R. Salakhutdinov, Multimodal transformer for unaligned multimodal language sequences, in: Proceedings of the Conference. Association for Computational Linguistics. Meeting, Vol. 2019, NIH Public Access, 2019, p. 6558.
[24] P. Xu, X. Zhu, D.A. Clifton, Multimodal learning with transformers: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2023).
[25] Z. Huang, T. Zhang, W. Heng, B. Shi, S. Zhou, Real-time intermediate flow estimation for video frame interpolation, in: European Conference on Computer Vision, Springer, 2022, pp. 624–642.
[26] A. Hamza, A.R.R. Javed, F. Iqbal, N. Kryvinska, A.S. Almadhor, Z. Jalil, R. Borghol, Deepfake audio detection via MFCC features using machine learning, IEEE Access 10 (2022) 134018–134028.
[27] D. Jatnika, M.A. Bijaksana, A.A. Suryani, Word2vec model analysis for semantic similarities in english words, Procedia Comput. Sci. 157 (2019) 160–167.
[28] Y. Yu, K. Adu, N. Tashi, P. Anokye, X. Wang, M.A. Ayidzoe, Rmaf: Relu-memristor-like activation function for deep learning, IEEE Access 8 (2020) 72727–72741.
[29] P. Korshunov, S. Marcel, Subjective and objective evaluation of deepfake videos, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2021, pp. 2510–2514.
[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.
[31] Y. Khasgiwala, J. Tailor, Vision transformer for music genre classification using mel-frequency cepstrum coefficient, in: 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies, GUCON, IEEE, 2021, pp. 1–5.
[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Ro{BERT}a: A robustly optimized {BERT} pretraining approach, 2020, URL https://openreview.net/forum?id=SyxS0T4tvS.

[33] H. Khalid, S. Tariq, M. Kim, S.S. Woo, FakeAVCeleb: A novel audio-video multimodal deepfake dataset, in: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021.
[34] T.-T. Wong, P.-Y. Yeh, Reliable accuracy estimates from k-fold cross validation, IEEE Trans. Knowl. Data Eng. 32 (8) (2019) 1586–1594.
[35] A. Chadha, V. Kumar, S. Kashyap, M. Gupta, Deepfake: an overview, in: Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020, Springer, 2021, pp. 557–566.
[36] J.K. Lewis, I.E. Toubal, H. Chen, V. Sandesera, M. Lomnitz, Z. Hampel-Arias, C. Prasad, K. Palaniappan, Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning, in: 2020 IEEE Applied Imagery Pattern Recognition Workshop, AIPR, IEEE, 2020, pp. 1–9.
[37] P. Swathi, S. Sk, Deepfake creation and detection: A survey, in: 2021 Third International Conference on Inventive Research in Computing Applications, ICIRCA, IEEE, 2021, pp. 584–588.
[38] P. Neekhara, B. Dolhansky, J. Bitton, C.C. Ferrer, Adversarial threats to deepfake detection: A practical perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 923–932.
[39] W. Saeed, C. Omlin, Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities, Knowl.-Based Syst. 263 (2023) 110273.