Multimodaltrace: Deepfake Detection using Audiovisual Representation Learning

Muhammad Anas Raza Khalid Mahmood Malik
Oakland University
{mraza, mahmood}@oakland.edu

Abstract

By employing generative deep learning techniques, Deepfakes are created with the intent to create mistrust in society, manipulate public opinion and political decisions, and for other malicious purposes such as blackmail, scamming, and even cyberstalking. As realistic deepfake may involve manipulation of either audio or video or both, thus it is important to explore the possibility of detecting deepfakes through the inadequacy of generative algorithms to synchronize audio and visual modalities. Prevailing performant methods, either detect audio or video cues for deepfakes detection while few ensemble the results after predictions on both modalities without inspecting relationship between audio and video cues. Deepfake detection using joint audiovisual representation learning is not explored much. Therefore, this paper proposes a unified multimodal framework, Multimodaltrace, which extracts learned channels from audio and visual modalities, mixes them independently in IntrAmodality Mixer Layer (IAML), processes them jointly in IntErModality Mixer Layers (IEML) from where it is fed to multilabel classification head. Empirical results show the effectiveness of the proposed framework giving state-of-the-art accuracy of 92.9% on the FakeAVCeleb dataset. The cross-dataset evaluation of the proposed framework on World Leaders and Presidential Deepfake Detection Datasets gives an accuracy of 83.61% and 70% respectively. The study also provides insights into how the model focuses on different parts of audio and visual features through integrated gradient analysis.

1. Introduction

The advancement of generative AI algorithms has resulted in voluminous synthetic media, including the emergence of deepfakes, which are audio and/or visual media that can be used by malicious actors to spread disinformation. Visual deepfakes include synthesizing videos by replacing the face in a video with another person (FaceSwap), altering expressions (Expression Swapping), or synchronizing the lip movement with some sound (Puppet Master), while audio deepfakes create cloned voices to depict content that was not spoken. The deepfakes videos and audio have achieved fidelity that now it becomes difficult for humans to distinguish it from forged. This brings up a major privacy and security threat as such media can be used to manipulate voice recognition systems, disseminate fake news, defame an individual, or distribute misinformation by impersonating renowned personnel or celebrities e.g. [1].

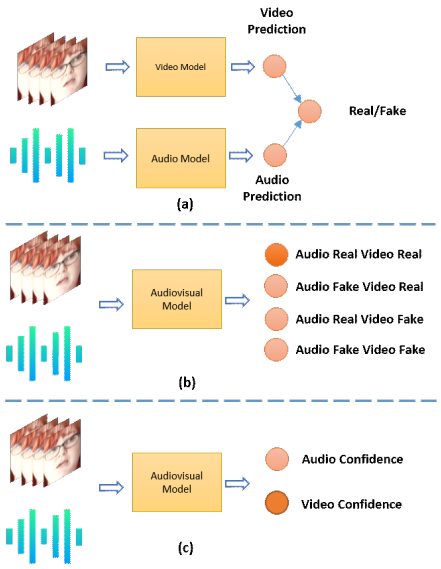

Figure 1. Techniques for multimodal deepfake detection (a) Ensemble models (b) Single Label Multiclass multimodal (c) Multilabel multiclass multimodal (proposed)

Deep learning-based detection methods are increasingly used for deepfake detection as they show better performance [11]. Overall, the algorithms for audiovisual deepfake detection can be classified into three categories depending on the modalities they focus on: unimodal, multimodal, and ensemble learning. Unimodal deal with a single modality for deepfake detection. Existing research has ex-

tensively explored single modality (i.e. image, video, and audio) deepfake detection methodologies [11]. Multimodal methods focus on representing audio and visual deepfakes in the same space [8]. A noticeable work [22] uses attention between audio and visual modalities for deepfake detection. Ensemble methods like [7] fuse decisions at the end of audio and visual networks. Limited research has been done on joint audiovisual representation learning, with most performant methods focusing on either audio or visual cues and ignoring the relationship between the two in a multimodal manner.

We propose a unified, multimodal audio-visual channel mixer for deepfake Detection analyzing audio and visual streams simultaneously. Our framework exploits the spatiotemporal characteristics of the input video along with spectral features in the input audio. The proposed framework has six blocks named input block, learned channel extraction block, IntrAmodality Mixer Layers (IAML), IntErmodality Mixer Layers (IEML), Channel Fusion, and a multi-label prediction head. The input block consists of preprocessing of both audio and visual modalities while the output block contains a multi-layered perceptron network with a multi-label classification head. Feature extraction block comprises Resnet-3D and Resnet1-D architectures for visual and acoustic modalities, respectively. Each modality is processed independently in IntrAmodality Mixer Layer (IAML) through MLP-mixer layers and jointly in IntErmodality Mixer Layer (IEML) before finally feeding to the output block. Our contributions are listed as follows:

1. We propose a novel Audiovisual Patch Mixer for DeepFake Detection, Multimodaltrace, which fuses learned channels from audio and visual modalities independently in IntrAmodality Mixer Layer (IAML) and jointly in IntEr-Modality Mixer Layer (IEML).

2. We also propose to reformulate the problem of audiovisual deepfake detection as a multi-label classification while previously it has been studied binary or multi-class classification which essentially means our model gives predictions about each modality while processing in the same space.

3. We evaluate the proposed framework on the FakeAVCeleb dataset and perform cross-dataset evaluation on World Leaders Dataset(WLD) and Presidential Deepfake Dataset(PDD) giving state-of-the-art performance among all unimodal and multimodal frameworks.We also perform a cross-dataset evaluation of the proposed framework on other publicly available datasets. We also perform an integrated gradient analysis to study the effectiveness of the proposed framework based on predictions in multi-label classification heads.

2. Related Work

Video Deepfake Detection. In the last few years, a lot of work has been done that uses Deep Neural Network architectures for deepfake detection using CNNs, and patches based-based architectures. [4, 9, 15].

Fake Audio Detection. Traditional approaches for detecting spoofed or fake audio employed hand-crafted features such as Cochlear Filter Cepstral Coefficients Instantaneous Frequency (CFCCIF) [14], Linear Frequency Cepstral Coefficients (LFCC) [16], and Constant-Q Cepstral Coefficients (CQCC) [17], combining with the Gaussian Mixture Model (GMM) for classification [18]. Several works use deep learning-based techniques for detecting fake audio, utilizing CNNs, RNNs, and a combination of both CNN and RNN models [13, 20].

Multi-modal Deepfake Detection The current literature lacks exploration in joint audiovisual representation learning for detecting deepfakes. Although multimodal architectures have been developed for other vision tasks, not much work has been done for deepfake detection. [12] developed an emotional embedding-based audio-visual deepfake detection method. The framework is evaluated on publicly available datasets not including FakeAVCeleb making the framework unexplored on jointly forged audiovisual content. [3] proposed a deepfake video detector that identifies fake videos based on the discrepancy between audio and visual components. The authors do not evaluate its effectiveness on deepfakes with audio and visual forgeries. [8] evaluated multiple frameworks including ensembles and multimodal architectures transferred from other domains, but the multimodal frameworks fail to generalize well on test data. [7] use audio and visual ensemble transformers for detecting audiovisual deepfakes. The major disadvantage of ensemble approaches is ignoring audiovisual cues during processing. Lastly, the approach in [6] ensemble audio, video, and audiovisual model showing increased performance on test data. Their major disadvantage, is redundancy of audio, video, and audiovisual models making it complex. Besides facial deepfakes, other domains have also been explored like video inpainting detection [19, 21].

3. Proposed Framework

The proposed architecture consists of six blocks with learned unimodal channel extractors, intramodality feature mixers, fusion, and joint intermodality feature mixers in the same space, and finally classifying both audio and video as a multi-label classification head. An overview of the proposed method is shown in Figure 2 a).

Problem Formulation

We formulate the deepfake detection as a binary classification task by utilizing both real and fake audiovisual modalities. To represent a video that includes human speech, we use the notation X = $\{x_a, x_v\}$, where x_a and x_v denote the audio and video channels, respectively. Both channels consist of sequences of sampled waveform digits

and video frames. The network that makes prediction is denoted as $\mathcal{F}_\theta(X)$, which includes three parts: the audio channel extractor $\mathcal{F}_{\phi a}$ maps input audio to a feature representation in $\mathbb{R}^{T_a \times d}$ with T_a and d respectively being the length of the audio sequence and feature dimension; While $\mathcal{F}_{\phi v}$ maps input video to a feature representation in $\mathbb{R}^{T_v \times d}$ with T_v and d respectively being the length of video sequence and feature dimension. $\mathcal{F}_{\psi av}$ processes maps the audiovisual feature representation to feature representation in $\mathbb{R}^{(T_a+T_v)\times d}$. The classification layer $\mathcal{F}$ maps feature representations to labels. Consider $D = \{(a_i, v_i, y_{a_i}, y_{v_i})\}i = 1^N$ be the dataset, where $y \in 0, 1$ is label representing whether the input is real or fake. y_{v_i} and y_{a_i} labels for the video and audio channels respectively and are independent of each other.
Given audio a and video v, audio and visual features are calculated as in Equation 1.

$$f^{*a}, f^{*v} = \mathcal{F}\phi^a(a), \mathcal{F}\phi^v(v) \tag{1}$$

f^{*a} and f^{*v} are audio and video feature representations calculated by $\mathcal{F}\phi^a$ and $\mathcal{F}\phi^v$ with a and v being corresponding audio and videos. The audio and visual feature representations are are fed to $\mathcal{F}\psi^{av}$ for joint feature representation as in Equation 2

$$f^{*av} = \mathcal{F}\psi^{av}(f^{*a}, f^{*v}) \tag{2}$$

f^{*av} is joint feature representations calculated by $\mathcal{F}\psi^{av}$ which is then fed to final classification function $\mathcal{F}$ for audio and visual prediction as in Equation 3

$$y^{*a}, y^{*v} = \mathcal{F}(f^{*av}) \tag{3}$$

Deepfake detection is a binary classification task for audio and visual modalities as in Equation 4.

$$L_{cls} = \sum_{(a,v,y^a,y^v)\in D} \mathcal{C}[(y^a, y^v), (y^{*a}, y^{*v}))] \tag{4}$$

Where L_{cls} is the overall loss that is backpropagated.

3.1. Input Block

To preprocess raw audio and videos from the dataset to be fed to subsequent layers of Multimodaltrace, input block is used. In this block, all audios are normalized and resampled at a constant sample rate, spectral features are computed using a discrete Fourier transform over the innermost dimension of the audio. The absolute value of the first half of the fft is taken which represents the positive frequencies. Videos are resized to a fixed size with a stack of frames and then standardized in the range of 0 to 1.
Given audio x_A with 3200 samples and video x_V comprising 5 frames at a frame rate of 25 FPS. fft is a function calculating Fast Fourier Transform of audio samples and ϖ truncates negative features as in Equation 5. γ operation resizes the sequence of frames to the fixed dimensions after which they are standardized using Υ as in Equation 6.

$$\rho_A = \varpi(fft(x_A)) \tag{5}$$

$$\rho_V = \Upsilon(\gamma(x_V)) \tag{6}$$

3.2. Learned Channel Extraction (LCE)

To extract deep features from both audio and video to be fed to mixer layers, unimodal learned channels are extracted for both modalities. For audio, we use ResNet-1D architecture to extract deep features. Learned deep channels for Videos, which are processed as a sequence of frames, are extracted using a multi-layered 3D Resnet architecture. We take 3D tublets from these learned channels which are then projected using dense neurons before finally feeding them to IntrAmodality Mixer Layer (IAML).
The preprocessed audio ρ_A and video ρ_V are fed to channel extraction layers χ_A and χ_V as shown in Equations 7 and 8.

$$\xi_A = \chi_A(\rho_A) \tag{7}$$

$$\rho'_V = \chi_V(\rho_V), \rho'_V \in \mathbb{R}^{T\times H\times W\times C} \tag{8}$$

For audio, 1-D patches are used to be fed to the IntrAmodality Mixer Layer(IAML) while 3D channels learned in the Equation 8 of shape (T, H, W, C) transposed to 3D tubelets with shape $(T \cdot H \cdot W, C)$ as shown in Equation 9.

$$\xi_V = (\rho'_V)^{\Xi}, \xi_V \in \mathbb{R}^{N_V \times C} \tag{9}$$

Ξ operation rearranges visual embeddings ρ'_V for feeding tubelets to the IntrAmodality Mixer Layer and,

$$N_V = T \cdot H \cdot W$$

An illustration of Audio and Visual Channel extractor is shown in Figure 2 b) and c) respectively.

3.3. IntrAmodality Mixer Layer (IAML)

To mix and learn correlations and patterns within the learned audio and visual channels, the outputs of Learned Channel Extractors (LCE) are transformed through IntrAmodality Mixer Layer (IAML) for both audio and visual modalities as in Equations 10 and 11 respectively.

$$\xi^*_A = \eta_A(\xi_A), \xi^*_A \in \mathbb{R}^{N_A \times d} \tag{10}$$

$$\xi^*_V = \eta_V(\xi_V), \xi^*_V \in \mathbb{R}^{N_V \times d} \tag{11}$$

Where η_A and η_V are independent MLP mixer layers for audio and visual modalities.

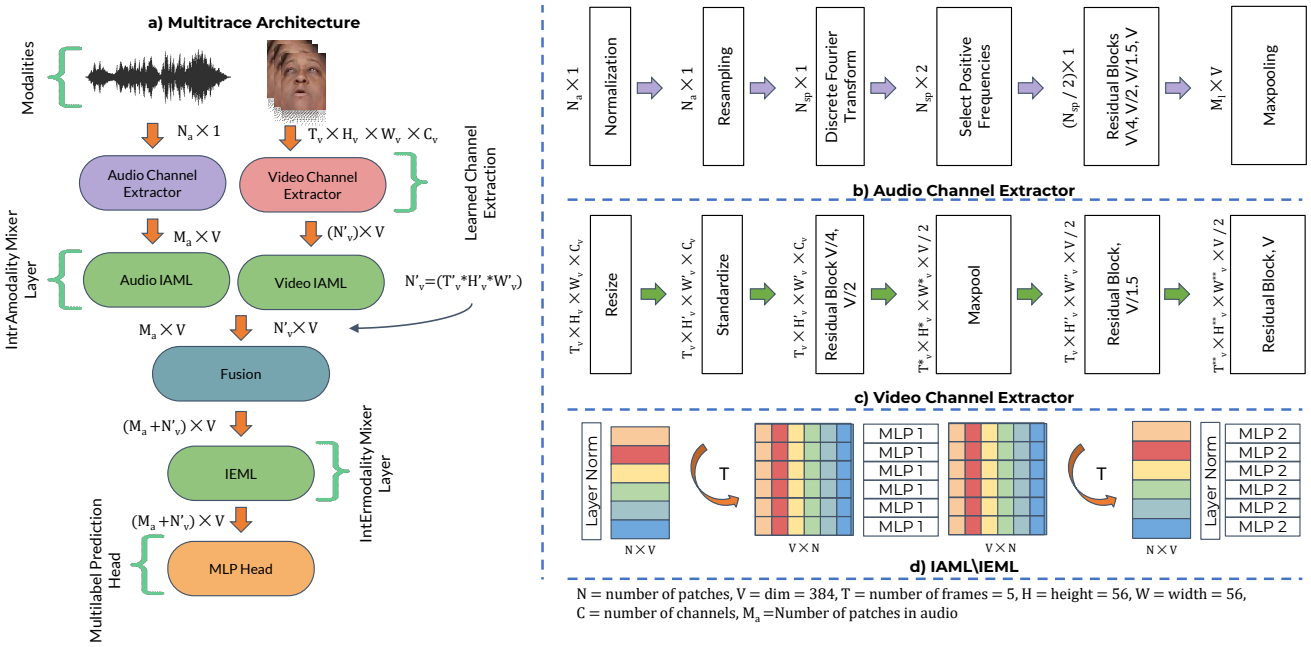

Figure 2. a) Multimodaltrace : Fuses channels from audio and visual modalities independently using IntrAmodality Mixer Layer (IAML) and jointly in IntErModality Mixer Layer (IEML). b) Audio Channel Extractor : Extracts spectral features with resampling and learn features using residual blocks. c) Video Channel Extractor: Resizes sequence of frames, standardizes and extracts features using multiple residual blocks and pooling layers. d) IAML/IEML: Mixer layers which mix across patch dimensions and pass through a shared mlp block to learn patterns between patches after which patterns within channels are learned.

The architecture of MLP Mixer is primarily based on multilayer perceptrons, which comprise two distinct types of layers. The first type involves the application of MLPs independently to individual channels, which facilitates the combination of per-channel features. The second type of layer applies MLPs across patches, enabling the mixing of cross-channel features. Importantly, the MLP Mixer design does not require sophisticated computations beyond basic matrix multiplication operations, scalar nonlinearities, and simple data layouts manipulations such as reshaping and transposition.

Given channels $p \in \mathbb{R}^{N_p \times dim}$, the MLP Mixer normalizes it using $\hbar$ after which they are transposed as in Equation 12,

$$p^* = (\hbar^*(p))^T, p^* \in \mathbb{R}^{dim \times N_p} \tag{12}$$

and projected using a shared dense network MLP_1 along channel dimension as in Equation 13,

$$p_1^* = MLP_1(p^*) \tag{13}$$

after which they are transposed along with a skip connection as in Equation 14,

$$p_1^{**} = (p_1^*)^T + p, p_1^{**} \in \mathbb{R}^{N_p \times dim} \tag{14}$$

the result of which is normalized after which another dense network MLP_2 is applied to patch independently as in Equation 15,

$$p_2 = MLP_2(p_1^{**}), p_2 \in \mathbb{R}^{N_p \times dim} \tag{15}$$

After which skip connection is applied as in Equation 16,

$$p_2^* = \hbar^{**}(p_2 + p), p_2^* \in \mathbb{R}^{N_p \times dim} \tag{16}$$

The outputs p_2^* are fed to subsequent MLP mixer layers or classification heads. An illustration of IAML architecture is shown in Figure 2 d)

3.4. Fusion

For joint learning of audiovisual modalities, the patches transformed through IntrAmodality Mixer Layer (IAML) are fused in a single vector, after which they are normalized to be fed to IEML in a joint space as in Equation 17.

$$\xi_{AV} = \hbar(\xi_A^* \parallel \xi_V^*), \xi_{AV} \in \mathbb{R}^{N_{AV} \times d} \tag{17}$$

Where $N_{AV} = N_A + N_V$, $\parallel$ operation concatenates audio embeddings ξ_A^* and visual embeddings ξ_A^*, while $\hbar$ brings dual modality features on a similar scale, helping to stabilize the gradient descent step.

3.5. IntErmodality Mixer Layer (IEML)

IEML combines and processes audio and visual patterns using an MLP-mixer on mixed and normalized tokens from

IAML layers. It uses channel-mixing and token-mixing MLP layers to communicate between channels and spatial locations, respectively. The layers are interleaved to enable interaction between audio and visual dimensions.

$$\xi^*_{AV} = \eta_{AV}(\xi_{AV}), \xi^*_{AV} \in \mathbb{R}^{N_{AV} \times d} \tag{18}$$

Where η_{AV} represents MLP Mixer architecture for mixing audio and visual channels.

3.6. Multi-label Classification Head

Multimodaltrace's classification head uses multiclass multilabel classification to predict both audio and visual modalities. The mixed and processed channels obtained from IEML are fed to an MLP Head from where we get multi-label classification output. The final logits are given by a fully connected network as in Equation 19

$$y_f = MLP(\xi^*_{AV}), y_f \in \mathbb{R}^2 \tag{19}$$

We apply sigmoid function to the logits for obtaining confidence across both audio and visual modalities as in Equation 20

$$y^*_{av} = \sigma(y_f), y_{av} \in \mathbb{R}^2 \tag{20}$$

Where,

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{21}$$

4. Experiments and Results

This section covers the experimental setup, datasets, and results for evaluating the Multimodaltrace model, including a comparison with state-of-the-art methods and performance on different datasets.

4.1. Datasets

The datasets used in this study include the FakeAVCeleb dataset, which comprises 500 real videos of celebrities and over 20,000 fake videos, including lip-synced fake videos with synthesized audio. The PDD dataset features videos of US presidents with half being real and the other half manipulated using lip synchronization, impersonated audio, and misleading content. The WLD dataset contains real and deepfake videos of prominent US politicians and presidents created using GANs and comedic impersonators. The deepfake videos include face-swapped and impersonating videos. Only FakeAVCeleb Dataset was used for training the proposed framework as it contained multimodal deepfakes including forged audios and videos.

4.2. Training and Hyperparameter Setting

The proposed model is optimized using the Gradient Centralized Adam optimizer with weight decay and a binary cross entropy loss. To find the optimal hyperparameters, extensive experimentation is conducted, including tuning the learning rate, batch size, label smoothing, and weight decay, with the optimal values found to be: learning rate = 0.001, batch size = 32, label smoothing = 0.1, and weight decay = 0.0001. The model is trained in a distributed manner, with the best model weights stored using early stopping. For optimal feature extraction, experiments are performed with different embedding dimensions and ResNet encoder-like convolution layers, with the best-performing model found to have an embedding size of 384. To tackle the problem of data imbalance, data augmentations and samples from other datasets like VoxCeleb are used for both audio and visual modalities. Additionally, samples from ASVSpoof 2021 with black frames as inputs for training resulted in the gradient exploding.

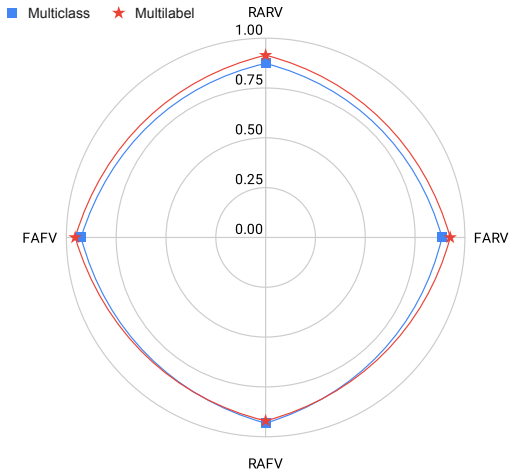

Figure 3. Comparison of Multilabel and Multiclass prediction heads on Multimodaltrace. RARV: Real Audio Real Video, FARV: Fake Audio Real Video, RAFV: Real Audio Fake Video, FAFV: Fake Audio Fake Video.

4.3. Performance Analysis

Our proposed architecture showed 92.9% accuracy which is state-of-the-art on FakeAVCeleb Dataset. We compare the proposed framework with unimodal, multimodal and ensemble baselines as shown in Table 1. The previous state-of-the-art model is a combination of ensemble and multimodal [6] gives 3.9% lesser accuracy on FakeAVCeleb. They ensemble audio, video, and audiovisual model to give predictions, with a major disadvantage, being the redundancy of audio, video, and audiovisual modalities making the entire architecture more complex. The multimodal approaches released by [3, 8] give maximum accuracy of 69%. While other multimodal frameworks are based on the fusion of audio and visual modalities and show below 70% accuracy on the same dataset. Additionally, Unimodal models [5,8] showed lower performance with maximum accuracy being 76.26%. Ensemble mod-

Table 1. Comparative Analysis of the Multimodaltrace.

Method	Model	Modality	Accuracy
Unimodal [5]	LipForensics	$\mathcal{V}$	0.76
Unimodal [8]	Xception	$\mathcal{A}$	0.7626
Unimodal_a [8]	VGG16	$\mathcal{V}$	0.8103
HRNet-18-BI100K [2, 10]	Face X-ray	$\mathcal{V}$	0.7288
HRNet-18-BI500K [2, 10]	Face X-ray	$\mathcal{V}$	0.7565
HRNet-32-BI100K [2, 10]	Face X-ray	$\mathcal{V}$	0.7675
Ensemble (Soft-Voting) [8]	VGG16	$\mathcal{AV}$	0.7804
Ensemble (Hard-Voting) [8]	VGG16	$\mathcal{AV}$	0.7804
Multimodal-1 [8]	Multimodal-1	$\mathcal{AV}$	0.5
Multimodal-2 [8]	Multimodal-2	$\mathcal{AV}$	0.674
Multimodal-3 [8]	CDCN	$\mathcal{AV}$	0.515
Multimodal-4 [3]	Not-made-for-each-other	$\mathcal{AV}$	0.69
VFD [2]	Multimodal Alignment	$\mathcal{AV}$	0.85
Multimodal-5 [6]	Ensemble + Multimodal	$\mathcal{AV}$	0.89
Multimodaltrace (Multi-class Head)	Modality Mixing	$\mathcal{AV}$	0.90325
Multimodaltrace (Multilabel - Proposed)	Modality Mixing	$\mathcal{AV}$	**0.929**

Table 2. Performance Analysis of the Multimodaltrace.

Class	Precision	Recall	F1-Score
ARVR	0.892	0.853	0.872
AFVR	0.921	0.895	0.908
ARVF	0.95	0.941	0.945
AFVF	0.914	0.921	0.917

Table 3. Cross-dataset Evaluation: **MT_ml**:MultiModalTrace with Multilabel head, **MT_mc**:MultiModalTrace with Multiclass head

Testing Subset	MT_ml	MT_mc	AVFNet [7]
WLD - FaceSwap	**75.84**	60.00	73.98
WLD- LipSync	**83.33**	66.66	69.32
WLD - Imposter	**91.66**	79.16	61.74
PDD - full	70.00	62.00	**78.12**

els [8] with soft and hard voting showed accuracy values of 78.04%. The unimodal approaches [8] with the VGG16 network showed an accuracy of 81.03% which is better when compared with ensemble approaches. The Multimodaltrace with a multiclass head showed better accuracy of 90.325%. The proposed Multimodaltrace with a multilabel prediction head outperformed all other models, including the previous state-of-the-art model.

Table 2 shows the performance analysis of the Multimodaltrace evaluated on four different classes: 1) ARVR: Real Audio, Real Video, 2) AFVR: Fake Audio, Real Video, 3) ARVF: Real Audio, Fake Video, 4) AFVF: Fake Audio, Fake Video. For each class, the precision, recall, and F1-score metrics are reported. From Table 2, we can see that the proposed Multimodaltrace performs well on all four classes, with F1-scores ranging from 87.2% to 94.5%. The best performance is achieved on the ARVF class, with a precision of 95% and recall of 94.1%, resulting in an F1-score of 94.5%. The relatively low performance is on the AFVR class, with an F1-score of 90.8%, which is still quite high. This indicates that the proposed framework can effectively detect deepfakes in audiovisual content.

4.4. Multi-label Effectiveness and Cross-dataset Evaluation

We compare the performance of multiclass and multilabel prediction heads in Figure 3 demonstrating the improved accuracy with multilabel prediction in Table 1. The effectiveness of the multilabel approach is further confirmed by the cross-dataset evaluation on subsets of the WLD dataset and the complete PDD dataset, as shown in Table 3. The proposed multi-label approach outperforms multiclass prediction on all subsets and compares favorably to an ensemble-based approach by [7] on most subsets, except for the Full Presidential Deepfakes Dataset. The multilabel approach outperforms the multiclass approach in cross-dataset evaluation across all testing subsets, with the highest accuracy of 91.66% achieved in the Imposter subset. Integrated gradient analysis is performed to better explain the relationship between each pipeline's predictions and input sequences and audio. The analysis shows the model's focus on the particular modality of the input, with gradients computed with respect to both audio and visual modalities for multimodal inputs. Results are presented in Figure 4.

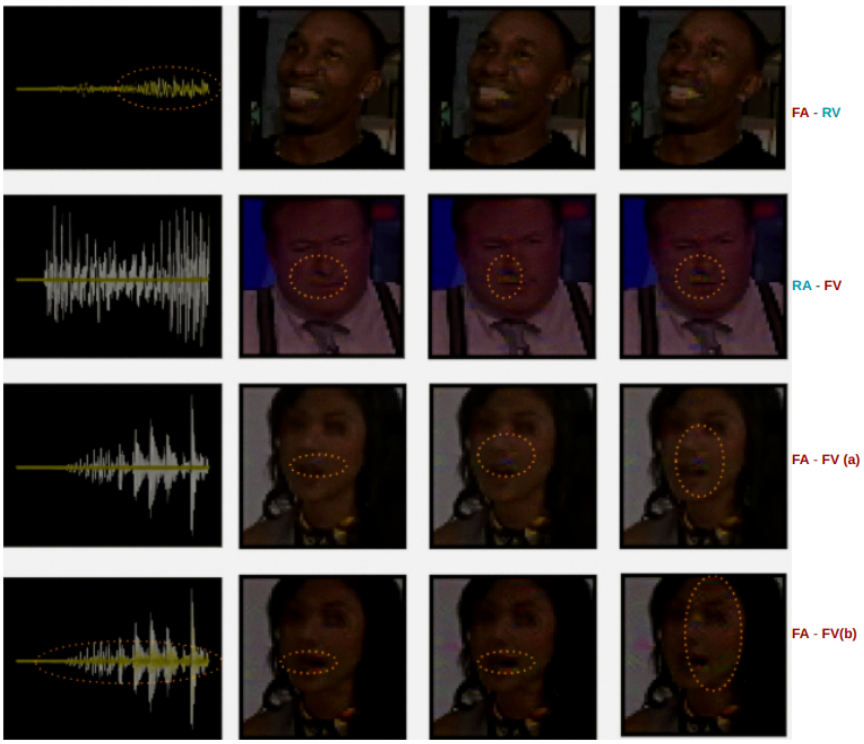

Figure 4. Integrated Gradient Analysis on all types of forged samples in FakeAVCeleb dataset. FA - RV: Fake Audio Real Video (gradient w.r.t audio modality), RA - FV: Real Audio Fake Video (gradient w.r.t video modality), FA - FV (a): Fake Audio Fake Video (gradient w.r.t. video), FA - FV (b): Fake Audio Fake Video (gradient w.r.t. audio)

5. Future Work

In the future, we plan to investigate the robustness of our proposed architecture by conducting an analysis of its performance under various types of noise perturbations and laundering scenarios. Additionally, we will consider alternative methods for integrating audio and visual information into the model, such as attention mechanisms or graph neural networks, which may offer improved performance and interpretability in the learned representations.

6. Conclusion

This paper proposes a state-of-the-art framework, Multimodaltrace, for audiovisual deepfake detection, which outperforms existing methods and achieves an accuracy of 92.9% on the FakeAVCeleb dataset. The proposed framework uses a combination of channel extractors and mixers to effectively analyze both audio and visual modalities using IntrAmodality Mixer Layers (IAML) and IntErmodality Mixer Layers (IEML). We also demonstrate the effectiveness of the multilabel classification approach in audiovisual deepfake detection, providing more detailed information to the framework. The study further shows the generalizability of the Multimodaltrace framework for detecting deepfakes by evaluating it on other datasets like Presidential Deepfakes Dataset and World Leaders Dataset.

Acknowledgement

This material is based upon work supported by the National Science Foundation (NSF) under award number 1815724 and Michigan Transnational Research and Commercialization (MTRAC), Advanced Computing Technologies (ACT) award number 292883.

References

[1] Zuckerberg deepfake appears in damning ad with toast to democrats, Dec 2022. 1

[2] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Tao Ye, and Liqiang Nie. Voice-face homogeneity tells deepfake. *arXiv preprint arXiv:2203.02195*, 2022. 6

[3] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia*, pages 439–447, 2020. 2, 5, 6

[4] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva. Forensictransfer: Weakly-supervised domain adaptation for forgery detection. *arXiv preprint arXiv:1812.02510*, 2018. 2

[5] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. 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*, pages 5039–5049, 2021. 5, 6

[6] Ammarah Hashmi, Sahibzada Adil Shahzad, Wasim Ahmad, Chia Wen Lin, Yu Tsao, and Hsin-Min Wang. Multimodal forgery detection using ensemble learning. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1524–1532. IEEE, 2022. 2, 5, 6

[7] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio-visual deepfakes detection. *Applied Soft Computing*, page 110124, 2023. 2, 6

[8] Hasam Khalid, Minha Kim, Shahroz Tariq, and Simon 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*, pages 7–15, 2021. 2, 5, 6

[9] Hasam Khalid and Simon S Woo. Oc-fakedect: Classifying deepfakes using one-class variational autoencoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 656–657, 2020. 2

[10] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020. 6

[11] Momina Masood, Mariam Nawaz, Khalid Mahmood Malik, Ali Javed, Aun Irtaza, and Hafiz Malik. Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, pages 1–53, 2022. 1, 2

[12] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832, 2020. 2

[13] Joao Monteiro, Jahangir Alam, and Tiago H Falk. Generalized end-to-end detection of spoofing attacks to automatic speaker recognizers. *Computer Speech & Language*, 63:101096, 2020. 2

[14] Tanvina B Patel and Hemant A Patil. Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech. In *Sixteenth annual conference of the international speech communication association*, 2015. 2

[15] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019. 2

[16] Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. 2015. 2

[17] Massimiliano Todisco, Héctor Delgado, and Nicholas WD Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Odyssey*, volume 2016, pages 283–290, 2016. 2

[18] Longbiao Wang, Yohei Yoshida, Yuta Kawakami, and Seiichi Nakagawa. Relative phase information for detecting human speech and spoofed speech. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. 2

[19] Bingyao Yu, Wanhua Li, Xiu Li, Jiwen Lu, and Jie Zhou. Frequency-aware spatiotemporal transformers for video inpainting detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8188–8197, 2021. 2

[20] Chunlei Zhang, Chengzhu Yu, and John HL Hansen. An investigation of deep-learning frameworks for speaker verification antispoofing. *IEEE Journal of Selected Topics in Signal Processing*, 11(4):684–694, 2017. 2

[21] Peng Zhou, Ning Yu, Zuxuan Wu, Larry S Davis, Abhinav Shrivastava, and Ser-Nam Lim. Deep video inpainting detection. *arXiv preprint arXiv:2101.11080*, 2021. 2

[22] Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14800–14809, 2021. 2