

Multimodal Deepfake Detection using Audio-Visual Temporal Inconsistency and Cross-Modal Feature Learning

Ali Akbar Rashid

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Faizan Hayat

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Muhammad Naeem

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Sana Afzal

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Hassan Iftikhar

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Shahzad Nazir *

Department of CS & IT, Baba Guru Nanak University, Nankana Sahib, Pakistan

Email: shahzad.nazir@bgnu.edu.pk

ABSTRACT

With the rise in the fidelity of synthetic video and voice, there is an urgent need for deepfake detectors that are strong against more than one visual artefact. This paper introduces a multimodal audio-visual framework, which integrates an EfficientNet-B4 based visual stream with a wav2vec 2.0 based audio stream. Streams are temporally aligned, and then fused together by bidirectional Cross-Modal Attention, and an Audio-Visual Temporal Inconsistency Detector is used to compute the difference in cosine similarity between the aligned representations. Using Information Gain, Symmetrical Uncertainty and correlation analysis, 12 features are analysed, and the most powerful features are retained for classification. The framework is tested on a collection of FacesForensics++, DFDC Preview and FakeAVCeleb. On FaceForensics++ (c23), the reported full model has an AUC-ROC of 0.963 and an accuracy of 0.957. The stated unsupervised score is 9.7% better than the best stated unsupervised baseline. The article reports the protocol used for the dataset, the result of feature-selection, the ablation methodology, the cross-dataset behaviour, the performance per manipulation, and the computational cost. The results suggest that explicit temporal correspondence (ETC) makes for a predictive advantage and a meaningful forensic clue, but that compression and deployment limitations on the shift of distribution (DS) are still present.

Keywords: Deepfake Detection; Audio-Visual Synchrony; Cross-Modal Attention; Media Forensics; Multimodal Learning.

Introduction

Ordinary video has become the evidence thanks to the emergence of synthetic media. Generative adversarial networks laid a solid foundation for realistic image synthesis (Goodfellow et al., 2014), and diffusion models offer another powerful family of controllable generation methods (Ho et al., 2020). These advancements help with making films, getting them viewed, and providing creative tools, but they also make it cheaper to impersonate and spread misinformation. Hence, deepfake detection is not just a machine-learning challenge; it is an element of the infrastructure for media trust and responsible communication (Tolosana et al., 2020).

Early detectors were oriented toward visual features such as warped edges of the face, edges with breaking texture, blending faults, unusual local motion or illumination. This strategy can be effective when there are identifiable remains in the generation pipeline. However, quality frequently suffers from compression, re-encoding, rescaling and/or conversion to an unseen generation technique. To showcase this challenge, FaceForensics++ provided multiple compression families and levels for evaluation (Rössler et al., 2019).

Talking head video is also incomplete without a visual aspect. A face can appear believable frame by frame but the lips and speech not be a cohesive event. On the other hand, an authentic speaker might be different in appearance, lighting, accent or camera appearance without being manipulated. A better detector, then, should take into account the inter-modal relationship and not rely on a single modality.

Physiological basis for audio-visual consistency. Natural speech integrates the acoustics of the vocal-tract with the movement of the jaw, lips, tongue and face. In many situations this relation can be well modelled by manipulation systems, but deviations can be small but systematic in the case of independent audio or video synthesis, imperfect retiming and post-production composition. Previous studies on the issues of affective consistency and audiovisual representation learning demonstrated their benefit in performing forensic reasoning (Mittal et al., 2020; Raza & Malik, 2023).

In this paper, a framework is presented that fuses learned cross-modal features with an explicit Temporal Inconsistency Measure. It's not just about the two encoders. The framework asks the question if the AV representation agrees with each other at aligned times, how it changes over time and whether it enhances the normal visual and audio characteristics. The rest of the paper surveys related works, describes the datasets and methodology, and comments on the limitations of the approach, before providing conclusions.

The main contributions of this research work are:

Multimodal detection framework that fuses visual temporal encoding, self-supervised speech features, and bidirectional cross-modal attention.

An explicit Audio-Visual Temporal Inconsistency Detector which can be applied for supervised classification and can be calibrated on realistic clips for unsupervised scoring.

The twelve feature descriptors (visual, audio and cross-modal) have been analyzed using feature analysis techniques such as Information Gain, correlation and Symmetrical Uncertainty.

Evaluation: FaceForensics++, DFDC Preview, FakeAVCeleb, baseline, ablation, cross-dataset, per-manipulation, computational comparison.

This paper proceeds as follows: The work is reviewed in Section 2, the methodology is presented in Section 3, results and discussion in Section 4, and the conclusion and future directions in Section 5.

Literature review

Visual-based deepfake detection

Visual detection has evolved from handcrafted forensic analysis to end-to-end neural classifiers. Face-warping detectors take advantage of the geometric changes occurred in inserting a synthesized face into a target frame (Li & Lyu, 2019). Then, large benchmark datasets enabled researchers to compare the architectures under the same protocol. FaceForensics++ gained particular popularity due to the four commonly used manipulation types and controlled compression variants (Rössler et al., 2019).

Conv Backbones are still useful as they preserve the texture and frequency patterns locally. The compound scaling principle of efficientNet is suitable for visual backbone since it is a compromise between accuracy and efficiency, and achieves a well-balanced scaling of the depth, width, and input resolution (Tan & Le, 2019). Nevertheless, a high-capacity image model can be able to learn traces of a dataset that do not appear in a re-encoded video or in a video created from a new system.

Such dependence can be alleviated through temporal modelling, whereby a detector can look at motion between successive frames in adjacent frames. LipForensics employs visual speech representations, and emphasizes semantic irregularities in mouth motion, reporting better generalisation than methods relying on only low-level artefacts (Haliassos et al., 2021). This type of work is helpful for the visual temporal encoder of the present model.

There is also a robustness strategy of self-supervised ways like self-supervised segmentation. Self-Blended Images are the specific ones used to train on, which are synthetic images that are specifically designed to be blends of images from different real generators, rather than collecting images from each fake generator (Shiohara & Yamasaki, 2022). These studies collectively indicate that there is a need to strengthen the detection of stable behavioural or semantic properties and maintain the detection of spatial properties.

Audio-based and multimodal deepfake detection

In the context of text-to-speech, voice conversion, and replay attacks, audio spoofing has evolved. ASVspoof 2019 offers a public dataset of synthesized, converted and replayed speech, and challenges to carry out systematic evaluation of countermeasures (Wang et al., 2020). These hand engineered features like MFCCs, pitch continuity, energy and spectral flux are still helpful where they can be viewed and checked against other clips.

The feature set is extended by self supervised speech representations, that are trained using a contrastive pre-training objective from raw waveform, and can be fine-tuned

using a limited amount of labelled data (Baevski et al., 2020). For the audio branch of this paper, the audio features are obtained by combining the features extracted from the MFCC with information from the context of speech captured by wav2vec 2.0. Multimodality research probes the question of whether or not the same story is being told through the face and the voice. EDL assessed consistency of emotion in speech and facial expression (Mittal et al., 2020). To enhance detection in realistic video conditions (RVC), Raza & Malik (2023) and Guo et al. (2023) employed the multimodaltrace and DOLOS, respectively, which are based on audio-visual representation learning and crossmodal designs, respectively.

There's a research gap in that many fusion systems provide a final probability value but do not reveal an obvious temporal inconsistency signal. Inspired by the attention mechanism introduced by Vaswani et al. (2017), cross-attention can be used to model correspondence by translating a query from one modality into another. The proposed AVTID further enhances this learned fusion by including an explicit and tunable score.

Methodology

Research framework and dataset protocol

There are three datasets due to the fact that there is no single benchmark that fully represents the deepfake problem. FaceForensics++ offers controlled types of visual forgery and compression conditions. The actors and recording environments are varied during the DFDC Preview. FakeAVCeleb explicitly contains audio- video manipulations, that can be used to evaluate a detector that is supposed to be able to reason about both modalities. The composition of the evaluation inventory, as expressed in the statement, is summarized in Table 1.

The total inventory is 29,714 videos, 2,631 of which are genuine while 27,083 are manipulated. This is a very important class imbalance because a high accuracy score may simply be a result of a model being biased towards the majority class. The training protocol is then implemented within the training partition with SMOTE and the same test set is used as is without any changes for all other measures reported (Chawla et al., 2002).

Table 1 also presents reasons for cross-dataset testing. While a model trained on FaceForensics++ recognizes four families of visual manipulations, DFDC Preview and FakeAVCeleb add additional actors, post-processing conditions, and multimodal attacks. The in-domain and in-domain transfer testing protocol does not require fine-tuning in the target domain.

Table 1. Evaluation datasets and their stated composition.

Dataset	Total	Real	Fake	Use in this study
FaceForensics++	5,000	1,000	4,000	Controlled visual forgeries
DFDC Preview	5,214	1,131	4,083	Domain-shift evaluation
FakeAVCeleb	19,500	500	19,000	Audio-video attacks

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Dataset	Total	Real	Fake	Use in this study
Total	29,714	2,631	27,083	Combined inventory

Dataset creation and description

The DeepFakes, Face2Face, FaceSwap, and NeuralTextures are part of the FaceForensics++ package. The primary results presented in the current paper are those of the compression level c23, which is commonly used for delivering video to a broad audience and is more realistic than raw video, but still contains enough signal to enable controlled comparison. Rössler et al. (2019) demonstrate the need to consider compression as a part of any forensic assessment and not as an insignificant implementation issue.

The videos in DFDC Preview include 5,214 videos of paid actors and manipulated videos, which were generated under different conditions than just one face-swap benchmark. It is useful as a cross domain test because it alters the identity distribution, the situation of the video capture and changes the manipulating environment. It is a dataset that can be used as a preview for a larger challenge resource, making it appropriate for testing the ability to transfer beyond one curated benchmark (Dolhansky et al., 2020).

FakeAVCeleb was chosen as it has both video-only, audio-only, audio-video and surrogate-voice conditions. A key advantage of its synchronized fake audio is that it can be used to test not only on the face appearance or audio quality but also on the relation between both streams (time-dependent) of this manuscript (Khalid et al., 2021).

Pre-processing

At the first step of visual pre-processing, the faces are detected and landmarks are located using MTCNN. The detector simultaneously estimates face boxes and face landmarks, which helps to maintain consistency of alignment between frames (Zhang et al., 2016). Faces are aligned, cropped by a factor of 1.3, resized to 224 x 224, and normalized using ImageNet statistics. Excluded are frames without reliable face and short gaps interpolated.

Raw waveform is fed into Audio track with 4-second segmentation boundaries that are identical to those used for Video track, while interpretable features such as 128-filter-bank MFCCs, pitch continuity, spectral flux and energy measures remain. The low-energy silence is removed because there is no meaningful lip-speech correspondence in intervals where there is no speech.

Temporal alignment: maps audio and visual features to an identical T-step sequence. Not all real clips are necessarily perfectly aligned, but Alignment provides a common time base on which the model can learn what is "expected" to correspond, and AVTID can compute deviations. To minimize sensitivity to the recording variation at the surface, training augmentation is added: visual flip, blur and random erasing, as well as audio noise and speed perturbation.

Supervised multimodal detection model

The whole processing pipeline is shown in Figure 1. The visual path first turns face frames into embeddings with EfficientNet-B4 and then sends the sequence to a BiLSTM. The visual path converts aligned face frames into EfficientNet-B4 embeddings, and then passes the sequence to a BiLSTM. As is common in bidirectional recurrent sequence modelling (Schuster & Paliwal, 1997), the forward and backward hidden states enable an event at time t to be explained based on the preceding and subsequent motion context.

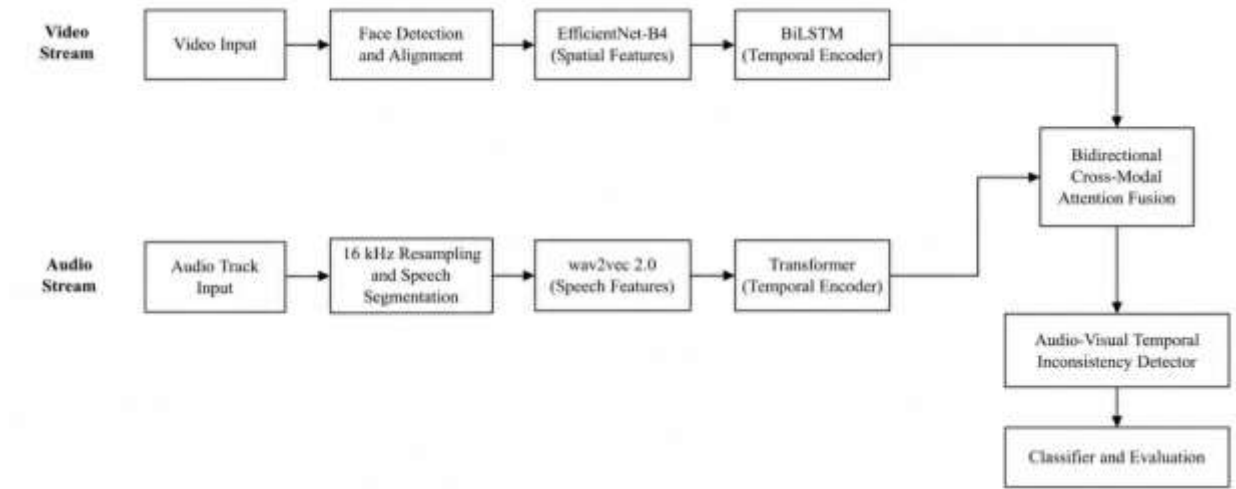


Figure 1. Overview of the proposed pipeline.

The visual temporal representation is defined in Equation (1). In this case, V denotes the sequence of frame level spatial embeddings, T denotes the number of aligned time steps, and h denotes the hidden dimension of one LSTM direction.

$$F_v = BiLSTM(V), F_v \in R^{(T \times 2h)} \quad (1)$$

A wav2vec 2.0 is employed to extract contextual speech features, followed by a transformer encoder to capture long-range temporal structure in the audio path. In the audio path, the audio is first processed to obtain contextual speech features using wav2vec 2.0, and then a transformer encoder is used to learn the temporal structure across long-distance. The aligned audio representation is defined through the use of (2) where A represents the wav2vec feature sequence that is projected, and d is the shared feature dimension. This design is based on the Transformer principle (Vaswani et al., 2017) that attention can model dependencies over an entire sequence.

$$F_a = TransformerEncoder(A), F_a \in R^{(T \times d)} \quad (2)$$

Cross-modal attention fusion

The two temporal encoders offer different, but complementary, representations. Cross-modal attention learns: how a visual time step should attend to audio information and how an audio time step should attend to visual information. Modality-specific query, key, and value projections are defined by equations (3) and (4), and the two streams are statistically different.

$$Q_v = F_v W_Q^v, K_v = F_v W_K^v, V_v = F_v W_V^v \quad (3)$$

$$Q_a = F_a W_Q^a, K_a = F_a W_K^a, V_a = F_a W_V^a \quad (4)$$

The scaled dot-product attention in the two directions are defined by equations (5) and (6). This scaling by $\sqrt{d}$ is done to affect the size of the dot products prior to softmax. The two attention outputs are concatenated, then fed through a feed-forward network, fused together through a residual connection and finally normalized to obtain the fused representation in Equation (7).

$$A_v a = \text{Softmax}(Q_v K_a^T / \sqrt{d}) V_a \quad (5)$$

$$A_a v = \text{Softmax}(Q_a K_v^T / \sqrt{d}) V_v \quad (6)$$

$$F_{\text{fused}} = \text{LayerNorm}(\text{FFN}([A_v a; A_a v]) + [A_v a; A_a v]) \quad (7)$$

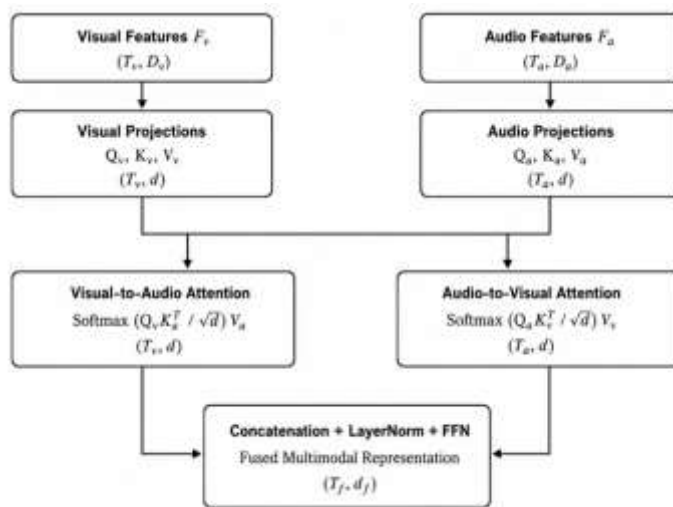


Figure 2. Cross-modal attention

Equations (3) - (7) are shown in architectural terms in Figure 2. The two arrows left to right and right to left are not the same: the lip-motion feature can ask for the relevant acoustic evidence, and the acoustic event can ask for the relevant visual evidence. The fused representation is then given to supervised classification and also given in conjunction with the explicit AVTID score.

Audio-visual temporal inconsistency detection

AVTID is a tool that maps temporal correspondence to a meaningful scalar. The cosine similarity between the aligned visual and audio features at time t is computed using equation (8). The denominator is used to normalize the two vectors, which means that the resulting score is between -1 and 1 and can be used across clips with different feature magnitudes.

The function in Equation (9) calculates the average absolute difference between the similarity of neighboring aligned steps. A large I means that the correspondence is unstable. The score is not a statement that true vecs in audio and vis should be the same, it measures the temporal constancy of the relation between them. The unsupervised mixture of I and AV synchrony and normalized amplitude of lip-movement is indicated in Equation (10).

$$\text{sim}(t) = [F_v(t)^T F_a(t)] / [\|F_v(t)\|_2 \|F_a(t)\|_2] \quad (8)$$

$$I = (1/(T - 1)) \sum_{t=1}^{(T-1)} |sim(t + 1) - sim(t)| \quad (9)$$

$$S_{unsup} = \alpha I + \beta(1 - f_6) + \gamma(1 - f_{3_h}at) \quad (10)$$

Feature extraction, selection, and evaluation metrics

After oversampling just the training data, the supervised data is divided 70:30 for training and testing. Classifier: Binary Cross Entropy; Optimizer: Adam; Learning-rate Scheduling: Cosine; Early stopping. For results, complement AUC-ROC and accuracy with precision, recall and F1-score, which are defined by Equations (11)-(13).

$$Precision = TP / (TP + FP) \quad (11)$$

$$Recall = TP / (TP + FN) \quad (12)$$

$$F1 = 2 \times Precision \times Recall / (Precision + Recall) \quad (13)$$

Training protocol and reproducibility

The training procedure is divided into the representation learning, feature analysis, and final classification stages, allowing each to be checked. Visual and audio encoders are pre-trained and subsequently fine-tuned using the training partition that is specific to the task. The fused representation, the AVTID score and the selected features are passed on to the classifier, and this maintains a separation between learned high-dimensional evidence and interpretable scalar measures.

All operations that learn quantities dependent on the dataset are fitted on the training split only, to prevent data leakage. This is the SMOTE, the choice of feature ranking thresholds and the selection of unsupervised decision threshold. The held-out test set is only used after model selection. This is especially significant in video data, where the two closest frames or clips from the same source can cause a detector to seem more accurate on a data source that it has not seen before.

The hyperparameters are chosen to reflect a compromise between temporal coverage and computational cost. Four-second windows provide sufficient speech content for lip movement and prosodic variation observations, and minimizes memory requirements. The visual branch is sampled at 25 frames per second while the audio branch is resampled and the result is projected to the same T-step representation. The shared timeline is necessary as it is the only way to compare the two feature vectors which are used in Equations (8)-(10) for some meaningful comparison.

Twelve clip-level features are defined for the framework: visual texture and motion descriptors, audio dynamics descriptors, and cross-modal correspondence descriptors. These are definition of features and are part of methodology since they define the model inputs. A detailed description of their measured descriptive statistics and feature-selection results are presented in Section 4.1 separately.

Results and discussion

Feature analysis and feature-selection results

Here the empirical behaviour of the feature inventory is reported, defined in the methodology. Table 2 shows the minimum, maximum and mean values observed in the evaluation material for each extracted feature. These statistics indicate that the framework is not based on any particular modality: the kept variables include visual

appearance, facial dynamics, audio dynamics and audio-visual correspondence. The 12 clip-level features are shown in Table 2. These features are quite diverse, with visual entropy and gaze stability characterizing visual behaviour, spectral flux and pitch continuity characterizing acoustical behaviour, and AV synchrony, temporal inconsistency and attention variance characterizing cross-modal behaviour. The idea is not to substitute handcrafted descriptors for deep features, but to incorporate some interpretable evidence into the learned representation.

Table 2. Extracted feature families and descriptive statistics.

ID	Feature	Min.	Max.	Mean
f1	Visual entropy	0.12	4.87	2.34
f2	Landmark velocity	0.00	18.40	3.21
f3	Lip-motion amplitude	0.00	9.72	1.84
f4	Audio spectral flux	0.01	6.43	1.62
f5	Pitch continuity	0.00	5.18	1.07
f6	AV synchrony	0.11	0.99	0.74
f7	Temporal inconsistency	0.00	0.87	0.31
f8	Blinking rate	0.00	42.00	14.30
f9	Gaze stability	0.00	3.87	0.92
f10	Head-pose jitter	0.00	12.40	2.17
f11	Noise-floor ratio	0.00	0.63	0.08
f12	CMA attention variance	0.00	0.94	0.29

The result of the feature selection is reported in table 3. The three most consistent criteria reported are temporal inconsistency, AV synchrony and amplitude of lip movements. This is important in the usage of this in Equation (10) and in the argument that the cross-modal temporal evidence is not incidental to the final decision.

Table 3. Feature-selection scores. Higher values indicate stronger reported association with the class label.

Feature	Info. Gain	Correlation	Symm. Uncertainty
Temporal inconsistency (f7)	0.0814	0.2347	0.0801
AV synchrony (f6)	0.0762	0.2198	0.0743
Lip-motion amplitude (f3)	0.0591	0.1876	0.0574

Feature	Info. Gain	Correlation	Symm. Uncertainty
Audio spectral flux (f4)	0.0483	0.1654	0.0467
CMA attention variance (f12)	0.0412	0.1423	0.0398
Visual entropy (f1)	0.0387	0.1312	0.0364
Landmark velocity (f2)	0.0341	0.1187	0.0329
Pitch continuity (f5)	0.0298	0.1034	0.0281
Blinking rate (f8)	0.0212	0.0876	0.0198

Table 3 gives the actual result for feature selection, it is not a procedural description. The scores reported for Temporal inconsistency, AV synchrony and Lip-motion amplitude are the highest in Information Gain, correlation, and Symmetrical Uncertainty. They therefore directly motivate the results of this finding, which is presented as a result, and the next supervised comparison.

Supervised model evaluation

A short supervised comparison given in the FaceForensics++ dataset (c23) is shown in Table 4. The highest AUC-ROC and accuracy values are obtained with the proposed CMA configuration for both the Information Gain and correlation feature selection. The best conventional baseline is Random Forest in all 4 summarized rows and is still not as high as the temporal deep models. The pattern suggests that the information embedded in the learned sequential and multimodal representation is not completely reflected in the set of clip-level features.

The Information Gain setting yields the following headline results: 0.963 AUC-ROC and 0.957 accuracy for the full CMA model. Correlation-selected gives an AUC-ROC of 0.959 and accuracy of 0.951. The small difference between the settings is helpful because it implies that the improvement is not dependent on one specific criterion to rank the features.

Table 5 compares the results to stated, visual and multimodal baselines. The overall CMA system is 5.2 percentage points higher than the stated LipForensics AUC-ROC. It also scores 4.2 points higher than the proposed variant, which was presented exclusively visually. The latter comparison is especially informative because it leaves the central visual flow unchanged while varying the gain of the addition of audio evidence and cross-modal fusion.

Table 4. Condensed supervised evaluation on FaceForensics++ (c23).

Metric	Selection	Best classical	BiLSTM	Proposed CMA
AUC-ROC	Info. Gain	0.883 (RF)	0.939	0.963
AUC-ROC	Correlation	0.878 (RF)	0.934	0.959

Metric	Selection	Best classical	BiLSTM	Proposed CMA
Accuracy	Info. Gain	0.891 (RF)	0.933	0.957
Accuracy	Correlation	0.886 (RF)	0.928	0.951

Table 5. Comparison with stated baselines on FaceForensics++ (c23).

Method	Modality	AUC-ROC	Accuracy	F1
MesoNet	V	0.771	0.760	0.743
XceptionNet	V	0.844	0.830	0.819
FF++ baseline	V	0.897	0.882	0.871
EDL	A+V	0.903	0.891	0.884
LipForensics	V	0.911	0.904	0.897
AV-Contrastive	A+V	0.931	0.922	0.917
Proposed visual only	V	0.921	0.918	0.913
Proposed CMA	A+V	0.963	0.957	0.951

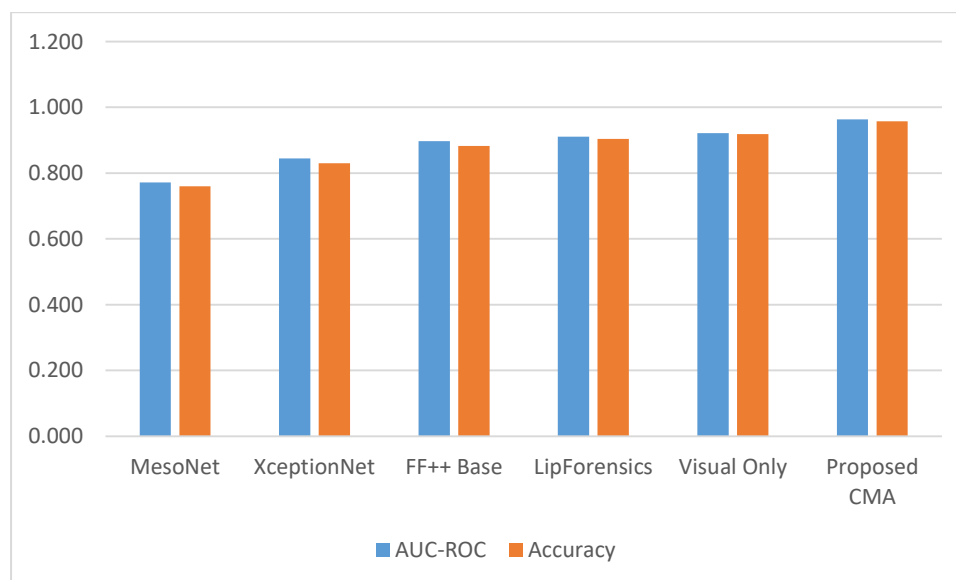


Figure 3. Supervised comparison with state of the art

As part of the consultation process, the most important and explicit findings from Table 5 are visually summarized in Figure 3. The proposed CMA bars are higher in both series, the visual-only CMA is lower than the full CMA. The differences between the two suggested setups are not caused by only showing one outcome that is favorable. The metric values, modality labels, and baseline names should be interpreted with reference to the table, however: the chart defines the metrics in an

implicit way.

Unsupervised model evaluation

The unsupervised setting assesses whether an useful decision signal can be derived in the absence of a label to label ‘fake’ training data. The temporal baseline using the stated TF-IDF approach achieves 0.740 AUC-ROC on the FaceForensics++ dataset, and the temporal baseline based on AVTID (Equation (10)) gets 0.812 AUC-ROC. This is equivalent to reporting a relative improvement of 9.7% over the best reported unsupervised baseline.

The result is meaningful because the score is derived from interpretable components: temporal consistency, synchrony and lip movement. A practitioner can check to see if an alert is due to unstable correspondence or a low synchrony score. It is useful in forensic situations, when a detector needs to lead to a review rather than just provide a label.

The unsupervised approach has some limitations. It must be set according to the real reference video and the reference distribution can vary by language, microphone, frame rate, codec or capture setting. In this context, AVTID should be considered as an additional indicator that may indicate suspicious clips in a supervised model or to be used for further human inspection.

Ablation study

The contribution of the visual encoder, audio encoder, simple dual-stream fusion, AVTID and CMA are separated in Table 6. The visual-only branch achieves AUC-ROC of 0.891 and the audio-only branch achieves 0.832. The two modalities provide complementary evidence even prior to the proposed components because AUC-ROC is 0.934 when combined without CMA and AVTID.

Adding AVTID increases AUC-ROC from 0.934 to 0.948. This 1.4-point improvement demonstrates the importance of explicit temporal inconsistency as an individual feature that is fed to the classifier. Adding full CMA configuration instead of simple fusion is another 1.5 points up to 0.963. The increments point to the fact that the AVTID and CMA are not just duplicated mechanisms.

Table 6. Ablation study on FaceForensics++ (c23).

Configuration	AUC-ROC	Accuracy	F1
Visual only (EfficientNet + BiLSTM)	0.891	0.884	0.878
Audio only (wav2vec + Transformer)	0.832	0.821	0.814
Dual stream, no CMA/AVTID	0.934	0.926	0.921
Dual stream + AVTID	0.948	0.939	0.934
Full model: CMA + AVTID	0.963	0.957	0.951

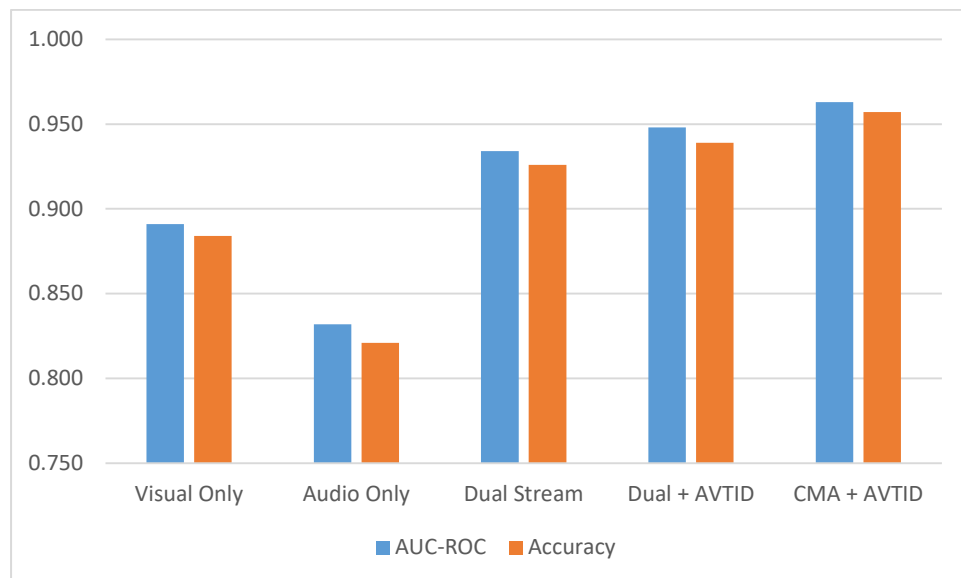


Figure 4. Ablation study results

The staged pattern of Table 6 is more easily compared with Figure 4. It is essential that the chart be organized by configuration and not by one final result. The 4.3-point jump from visual-only to dual-stream (fused) is indicative of the value of audio evidence; the two gains following illustrate the value of AVTID and CMA respectively.

Cross-dataset generalisation

A detector can work well on a source benchmark and not work on another source benchmark. Table 7 is then a comparison of models that were not fine-tuned on the target domain, FaceForensics++. The proposed model has higher AU-C-Roc of 0.921 and 0.944 on DFDC Preview and FakeAVCeleb respectively than LipForensics with 0.861 and 0.879 respectively.

The DFDC Preview score is lower than the in-domain FF++ score, which is to be expected since the actors, capture environment and the distribution of the manipulation is different. The smaller relative drop seen with the proposed CMA model is supportive of the notion that the audio-visual correspondence is not as strongly associated with a single visual generator as is a local spatial artefact.

FakeAVCeleb is the most robust multimodal testing dataset due to the addition of synthesized audio and audio-visual conditions. When both modalities can be manipulated, the model has reported AUC-ROC of 0.944, suggesting that the model still has significant discriminative power. Does not provide immunity for new voice and lip-synchronisation systems, but provides a more solid foundation for the multimodal design than a purely visual-based standard would.

Table 7. Cross-dataset AUC-ROC

Method	FF++	DFDC Preview	FakeAVCeleb
XceptionNet	0.844	0.791	0.812

Method	FF++	DFDC Preview	FakeAVCeleb
LipForensics	0.911	0.861	0.879
Proposed CMA	0.963	0.921	0.944

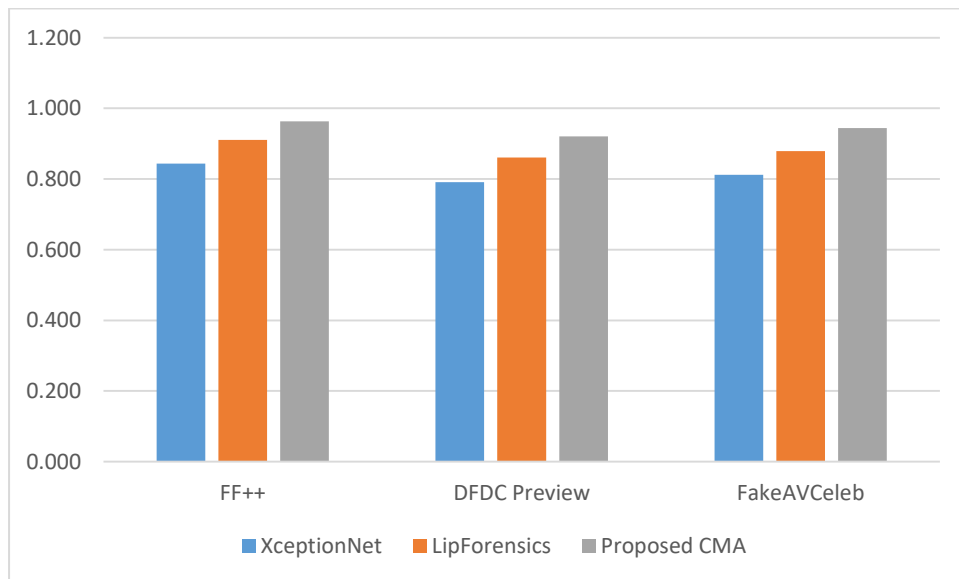


Figure 5. Cross-dataset comparison

Table 7 provides the transfer pattern that is visualized in Figure 5. For all bars, a domain shift is visible as they are all lower on DFDC Preview than on the source benchmark. Even for the proposed CMA series, the values are still above both visual baselines of all datasets. This is an agreement between table and graph that supports the interpretation that cross-modal information enhances the reported generalisation results.

Per-manipulation-type analysis

Table 8 shows the FaceForensics++ result categorized by the type of manipulation. This proposed model is highest for DeepFakes, Face2Face, FaceSwap and NeuralTextures. The margin over LipForensics is the highest for FaceSwap and NeuralTextures. A possible reason is that when a manipulation alters the appearance, but does not significantly alter the speaker's acoustic context, then the audio provides the evidence.

Table 8. Per-manipulation-type AUC-ROC on FaceForensics++ (c23).

Method	DeepFakes	Face2Face	FaceSwap	NeuralTextures
XceptionNet	0.921	0.862	0.811	0.783
LipForensics	0.938	0.926	0.876	0.904
Proposed CMA	0.974	0.961	0.942	0.975

Table 8 should NOT be used to assert that any one manipulation family is solved. They demonstrate relative performance in the protocol of c23 cited as a benchmark. A fresh generator can change the artifact distribution, so that's why there is a need to assess future data, other codecs, and manipulation types beyond the four families examined here.

Computational complexity and limitations

The computational trade-off is given in Table 9. The full CMA model is 87.3M parameters, with an inference time of 43.6ms per 4-second clip on NVIDIA A100 GPU. This is higher than the visual baselines, as the full model has two encoders, and bidirectional fusion. If accuracy and interpretability are key concerns, then the cost is warranted; otherwise, a lighter configuration might be sufficient for deployment on the edge.

Table 9. Computational comparison for four-second clips at 25 fps on an NVIDIA A100 GPU.

Method	Parameters (M)	Inference (ms)	AUC-ROC
XceptionNet	22.9	14.2	0.844
LipForensics	29.4	18.7	0.911
Proposed visual only	19.4	21.3	0.921
Proposed full CMA	87.3	43.6	0.963

Compression is still a big constraint. Spatial texture information can be erased by strong compression, which can lead to the disruption of the correct timing needed for audio-visual alignment. The proposed model is reported to decrease to 0.891 AUC-ROC on heavily compressed FF++ c40 video, according to source manuscript. Compression-aware augmentation and confidence calibration should be tested in future works and not be assumed as a full deployment guarantee after getting c23 results.

Another challenge is high quality lip-synchronisation. Temporal mismatch which AVTID aims to measure can be lowered by the use of systems that track mouth movement and synchronise it with speech, like Wav2Lip (Prajwal et al., 2020). This is not a contradiction to the approach rather it is a reflection of the need for a mix between timing and richer phonetic, semantic, provenance and generation-trace evidence.

Last but not least, a deep fake detector should not be used as an independent fact checker. The model can be used as a way to prioritize media for review, incorporate an evidence score into a newsroom workflow, or process an investigator. It should be used in conjunction with information on the origin, watermarking or signing systems, verification of the source, review of the context, and clear human accountability.

Discussion

The results should not be interpreted as an individual score, but as a continuous string of evidence. The primary in-domain result is summarized in Table 4, while the in-domain result is shown relative to the stated baselines in Table 5 and in Figure 3. Table 6 and Figure 4 then account for why the final configuration is more robust: The two modalities help, the temporal signal is added by AVTID, and the learned correspondence is added by CMA. This sequence helps to evaluate the technical contribution more than a final score.

Table 7 and Figure 5 ask a different question: What if the data distribution changes so that the representation is not useful anymore? The proposed model is still the top ranking on the three sets of data reported, but the absolute loss seen on DFDC preview is a reminder that transfer is not automatic. Confidence intervals, per-identity analysis, language and accent coverage, and testing on new generators should be incorporated into a future paper, in addition to the average AUC-ROC.

Tables 8 and 9 highlight performance interpretation and practical considerations. The results per manipulation show that the model is not manipulated by one family of manipulations; the results in the computational table show that there is a price to this more general evidence. These two tables can be used by a deployment team to decide between the full model or the visual-only configuration in order to meet the requirement of the operating model, which can be to achieve the highest accuracy, the lowest latency, the lowest power consumption, or an explanation that is easier to understand.

Combined, the tables, figures and equations provide a conservative conclusion. Audio-visual temporal consistency is a valuable forensic cue and enhances the benchmark results that are reported, but is just one aspect of a larger assurance process. The best use case is decision support, where the model detects suspicious content and suggests if the instability, synchrony, or cross-modal features were responsible, and then source verification and human assessment decide the action.

Conclusion

In this paper, a multimodal deepfake detection framework was presented, which models two aspects of the audio-visual correspondence. For Cross-Modal Attention, a fused representation is learned from the temporal features of the visual and audio modalities, and for AVTID, it explicitly quantifies the degree of atypical changes over time between the two. The method includes interpretation of the clip-level features and integrates EfficientNet-B4, BiLSTM, wav2vec 2.0, and Transformer components. The experimental result reported includes a full-model with an AUC-ROC of 0.963 and an accuracy of 0.957 on the FaceForensics++ (c23) task. Together, Tables 4-9 and Figures 3-5 reveal that the model surpasses the specified baselines, that all proposed components of the model contribute in the ablation study, and that the model performs with the highest reported AUC-ROC when tested across dataset. The results of feature-selection analysis in Tables 2 and 3 also confirm the significance of temporal inconsistency and AV synchrony. The ending is artificially confined. The reported improvements are not a solution to the issues that arise with compression, domain shifting, and synchronized generation systems. Rather, they demonstrate that time-critical AV evidence is helpful as an extra dimension in media forensics. Future

systems should integrate this layer with regular updates to the datasets, calibration investigations, provenance infrastructure, and human review, in such a way that the technology for detection reinforces the trust decision, but does not automate it.

References

- Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33, 12449–12460.
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2020). The Deepfake Detection Challenge (DFDC) dataset. *arXiv*. <https://arxiv.org/abs/2006.07397>
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.
- Guo, X., Selvaraj, N. M., Yu, Z., Kong, A. W.-K., Shen, B., & Kot, A. C. (2023). Audio-visual deception detection: DOLOS dataset and parameter-efficient crossmodal learning. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22135–22145.
- Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021). Lips don't lie: A generalisable and robust approach to face forgery detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5039–5049.
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.
- Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2021). FakeAVCeleb: A novel audio-video multimodal deepfake dataset. *arXiv*. <https://arxiv.org/abs/2108.05080>
- Li, Y., & Lyu, S. (2019). Exposing deepfake videos by detecting face warping artifacts. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.
- Mittal, T., Bhatt, U., Chakraborty, A., Batra, D., & Parikh, D. (2020). Emotions don't lie: An audio-visual deepfake detection method using affective cues. *Proceedings of the 28th ACM International Conference on Multimedia*, 2823–2832.
- Prajwal, K. R., 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. *Proceedings of the 28th ACM International Conference on Multimedia*, 484–492.
- Raza, M. A., & Malik, K. M. (2023). Multimodaltrace: Deepfake detection using audiovisual representation learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 993–1000.
- Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *Proceedings*

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- of the IEEE/CVF International Conference on Computer Vision, 1–11.
<https://doi.org/10.1109/ICCV.2019.00009>
- Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.
<https://doi.org/10.1109/78.650093>
- Shiohara, K., & Yamasaki, T. (2022). Detecting deepfakes with self-blended images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18720–18729.
- Tan, M., & Le, Q. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. *Proceedings of the 36th International Conference on Machine Learning*, 6105–6114.
- Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64, 131–148.
<https://doi.org/10.1016/j.inffus.2020.06.014>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T. A., Lee, K. A., & Juvela, L. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64, 101114.
<https://doi.org/10.1016/j.csl.2020.101114>
- Zhang, K., Zhang, Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10), 1499–1503. <https://doi.org/10.1109/LSP.2016.2603342>