

AV-CACNet: A Cross-Attention Consistency Network for Audio-Visual Deepfake Detection in Cybersecurity

Assem Turarbek¹, Sharipa Temirgazyieva^{2*}, Ahmet Sakir Dokuz³, Alina Azanova⁴, Nazgul Mametzhanova⁵
Al-Farabi Kazakh National University, Almaty, Kazakhstan^{1,2,4}
Nigde Omer Halisdemir University, Nigde, Turkey³
Kazakh National Women's Teacher Training University, Almaty, Kazakhstan⁵

Abstract—The rapid development of audiovisual deepfake technologies has made the automatic verification of digital content authenticity a pressing issue, as methods relying solely on video or audio may fail to fully characterize complex forms of multimodal manipulation. This study proposes the AV-CACNet (Audio-Visual Cross-Attention Consistency Network) architecture, designed to jointly model the interdependencies between audio and visual features. The proposed method processes facial imagery, the mouth region, and audio information as separate data streams, integrating CBAM, Temporal Transformer, Face-Mouth Cross-Attention, Audio-Visual Cross-Attention, a Consistency Module, and Adaptive Gated Fusion mechanisms into a single architecture. Experiments were conducted using the FakeAVCeleb dataset. Based on a source-level splitting strategy, a balanced set of 1,996 videos derived from 499 unique sources was created and divided into training, validation, and test sets for four-class classification. Assessment on previously unseen test data yielded an accuracy of 88.82% for AV-CACNet. The associated macro-averaged precision, recall, and F1-score were 89.50%, 88.82%, and 88.76%, respectively. In the binary deepfake detection scenario, the model achieved a ROC-AUC of 97.76%, with an Equal Error Rate of 7.89%. The highest F1-score by class reached 91.93% for the FakeVideo-FakeAudio category. The results obtained for the four-class FakeAVCeleb classification outperformed previously published results from Multi-view CNN (85.96%) and AMID (87.65%). These results indicate that integrating the connections and coherence between audio and visual data can enhance the examination of multimodal deepfake materials.

Keywords—Audio-visual deepfake; deepfake detection; multi-modal learning; cross-attention; cybersecurity

I. INTRODUCTION

In recent years, the rapid development of technologies for generating synthetic media based on artificial intelligence has complicated the task of verifying the authenticity of digital content. While artificial alterations to images, video, or audio could previously be detected relatively easily through visual or acoustic cues, modern generative methods can produce material that appears natural to human perception and is difficult to distinguish from original content. In particular, deepfake technologies which enable the simultaneous alteration of a person's facial appearance and voice—impose new requirements for assessing the reliability of information in the digital environment [1], [2].

The issue of deepfake content extends beyond the mere verification of media authenticity. Synthetic video and audio can compromise the integrity of digital data within information systems, undermine user authentication, and affect the reliability of multimedia information. Such content can be used to fabricate false identities, execute social engineering schemes, or deceive users into believing misinformation. Consequently, the automated detection of deepfakes is regarded as a key information security challenge, creating a need to develop methods that account not only for individual video or audio characteristics but also for the interrelationships between them [3].

A significant portion of existing deepfake detection methods prioritizes a specific modality. While video-based methods analyze changes in facial structure, expressions, or motion patterns between frames, audio-oriented models aim to detect artificial alterations in the voice spectrum and speech signal [2]. Although each of these approaches extracts useful features from its respective modality, they may fail to fully leverage the natural correspondence between audio and video. Recent research identifies multimodal approaches based on the joint processing of audio and visual information as a promising avenue for addressing this issue [3], [4], [5].

In the context of deepfakes, artificial manipulation may manifest not only in a single channel but also through various combinations of audio and visual components. Unimodal detection methods may struggle when only one component of a video is manipulated, such as when genuine visual content is combined with synthetic speech or when manipulated video is accompanied by authentic audio. FakeAVCeleb explicitly represents these cases through four separate categories: RealVideo-RealAudio, RealVideo-FakeAudio, FakeVideo-RealAudio, and FakeVideo-FakeAudio [6]. In addition, authenticity assessed separately for each modality does not necessarily imply that the audio and visual streams are temporally and semantically consistent with one another. Temporal mismatches between the speaker's lip movements and the audio signal, or the presence of unnatural correlations between visual and acoustic cues, can provide additional information for deepfake detection [4], [5], [7]. Therefore, in a multimodal system, it is crucial not only to process features from each modality independently but also to model their interaction and the degree of consistency between them.

*Corresponding author.

The study utilized the FakeAVCeleb dataset and examined four distinct audio-visual scenarios: real video–real audio, real video–fake audio, fake video–real audio, and fake video–fake audio. A source-level splitting approach was employed to prevent samples from the same source from being distributed across different sets. This ensured that the model's performance could be evaluated without the influence of adaptation to specific individuals or sources. The main objective of this study is to develop and evaluate a multimodal neural architecture for audio-visual deepfake detection that models the relationship and consistency between audio and visual features. The proposed approach combines facial, mouth-region, and acoustic features within a unified model and evaluates both four-class deepfake classification and binary authentic/fake detection.

The main contribution is the integration of cross-attention, consistency modeling, and adaptive multimodal fusion in the proposed AV-CACNet architecture. Auxiliary audio and visual tasks are also used to improve the utilization of multimodal information.

II. LITERATURE REVIEW

The rapid progression of deepfake generation technologies has driven the development of increasingly sophisticated detection approaches. While early research focused primarily on identifying artificial markers in video frames, facial structural distortions, and temporal artifacts, recent work has increasingly emphasized the simultaneous analysis of multiple modalities within synthetic content. The primary reason for this shift is that modern deepfake systems can alter not only a person's facial appearance but also their voice, lip movements, and speech characteristics. Consequently, leveraging the natural relationship between audio and video has become a key area of deepfake detection. Recent studies treat concepts such as multimodal detection, cross-modal consistency, and audio-visual synchronization as fundamental components of this challenge [8].

An important direction in audio-visual deepfake detection is the analysis of voice-face correspondence. Cheng et al. considered deepfake detection as a voice-face matching problem, based on the assumption that a person's voice and facial appearance share identity-level characteristics [9]. They developed a voice-face matching model and applied pre-training on general audio-visual data followed by adaptation to deepfake datasets. Experiments were conducted on DFDC, FakeAVCeleb, and DeepfakeTIMIT. On FakeAVCeleb, the proposed method achieved an AUC of 86.11%. The key aspect of this approach is that it evaluates whether the voice and face belong to the same person rather than relying only on visual artifacts [9].

Ilyas et al. [10] introduced AVFakeNet, a deep learning framework that integrates audio and visual features using a Dense Swin Transformer architecture for audio-visual deepfake detection. In contrast, AV-CACNet incorporates cross-attention, consistency learning, and adaptive gated fusion to facilitate multimodal feature interaction.

The work on ART-AVDF is particularly significant for AV-CACNet, as both methods treat audio-visual correspondence as a primary source of information. However, while ART-AVDF

prioritizes physiological articulatory representations, our approach in addition to evaluating correspondence via a dedicated Consistency Module employs both Face-Mouth Cross-Attention and Audio-Visual Cross-Attention mechanisms. This distinction enables the interrelationships between multimodal signals to be examined at multiple levels. The work on AVFakeNet demonstrates another approach to integrating audio-visual cues within a single architecture. The authors of AVFakeNet proposed a model based on the Dense Swin Transformer to process audio and visual streams within a unified end-to-end architecture. The study examined the model's robustness regarding various head poses, lighting conditions, and ethnic characteristics, and included a cross-corpus evaluation. This work demonstrated that processing audio and visual information within a single system enhances deepfake detection capabilities [11].

While the results of AVFakeNet demonstrate the potential of multimodal architectures, the use of large Transformer-based feature representations complicates the model structure. Furthermore, the mere integration of audio and visual signals does not necessarily reveal the precise nature of their interrelationship. Consequently, a shift from simple fusion to cross-modal interaction models is observed in subsequent research.

Another significant development in this area is the AVSFF method [12], which is based on audiovisual synchronization and lip movement analysis. Javed and his colleagues combined elements of CNNs, Temporal Convolutional Networks, BiLSTMs, and Dynamic Time Warping to detect subtle discrepancies between audio and lip movements. In this system, visual and audio features are processed independently; their synchrony is then evaluated, and a final decision is reached using late fusion. The authors conducted experiments on the FakeAVCeleb, AV-Deepfake1M, TVIL, and LAV-DF datasets, reporting accuracy rates of 99.73%, 97.60%, 98.90%, and 97.86%, respectively.

This result demonstrates the informational value of lip-audio synchronization for deepfake detection. The limitation becomes more apparent in multimodal settings, where a detector must determine not only whether manipulation is present, but also which modality has been altered. Conventional approaches generally formulate the problem as a binary real/fake decision, whereas FakeAVCeleb separates the data into four audio-visual configurations: RealVideo-RealAudio, RealVideo-FakeAudio, FakeVideo-RealAudio, and FakeVideo-FakeAudio [6]. Consequently, effective multimodal analysis requires identifying the manipulated component rather than assigning a single overall deepfake score to the entire sample.

In recent work, the cross-attention mechanism has emerged as a key tool for integrating audio-visual features. In the CMALDD-PTAF method published in 2026, Jin and Wang extracted unimodal features from a pre-trained audio-visual speech recognition model and fused them using a specialized Cross-Modality Fusion Module [13]. Subsequently, a Transformer classifier and domain-discriminative adversarial learning were employed. Reporting results of 98.9% accuracy and 99.6% AUC on the FakeAVCeleb dataset, the authors demonstrated the impact of cross-modal fusion and adversarial learning on the performance of multimodal deepfake detection. Although this approach is conceptually similar to the AV-CACNet architecture, there is a significant difference.

CMALDD-PTAF relies on pre-existing audio-visual speech representations and aims to enhance cross-domain robustness through adversarial training. In contrast, AV-CACNet integrates CBAM-based modality-specific encoders, a temporal Transformer, Face-Mouth Cross-Attention, Audio-Visual Cross-Attention, a Consistency Module, and Adaptive Gated Fusion into a single architectural pipeline.

Another key issue in multimodal models is the detection of temporal inconsistencies between audio and video. Research in 2025 proposed detecting such inconsistencies at a local temporal level. The study "Audio-Visual Deepfake Detection With Local Temporal Inconsistencies" investigated fine-grained temporal inconsistencies between audio and visual modalities using a temporal distance map and an attention mechanism. The authors evaluated the method on the DFDC and FakeAVCeleb datasets [14].

Systematic reviews also confirm the general development trajectory of multimodal deepfake detection. A 2024 systematic review by Abbas and Taeihagh categorized deepfake generation and detection methods, as well as the algorithms and frameworks employed, demonstrating that deepfake detection has evolved into a broad research field encompassing audio, image, and video modalities [8]. In addition to technical issues, the authors examined the impact of deepfake technology on disinformation and the trustworthiness of digital information.

The 2026 AMID framework [15] models audio-visual interactions across temporal, spatial, and semantic dimensions using complementary cross-modal reconstruction, disentangled representation learning, and adaptive low-rank adaptation. On FakeAVCeleb, it achieved 96.40% accuracy for binary classification and 87.65% for four-class classification. In comparison, AV-CACNet achieved 88.82% accuracy and an 88.76% macro F1 score under a source-level data split. Although these results provide a useful reference, differences in dataset preparation and evaluation protocols limit direct comparability.

A comparison of existing studies reveals four key limitations: insufficient modeling of audio-visual correspondence, limited analysis of facial and mouth movement relationships, restricted manipulation-type identification in binary classification, and the need for source-independent data splitting to ensure reliable evaluation.

To address these limitations, AV-CACNet integrates Face-Mouth Cross-Attention, Audio-Visual Cross-Attention, consistency learning, and adaptive gated fusion. This unified architecture aims to detect deepfakes by modeling cross-modal relationships alongside modality-specific features.

A. Research Gap

The literature review indicates that audio-visual deepfake detection has advanced considerably in recent years. Nevertheless, existing approaches often focus on a particular type of audio-visual feature integration or on specific correspondence cues. Some studies examine the synchronization between lip movements and audio signals, while others focus on voice-face identity correspondence or general multimodal fusion mechanisms. Analysis of these approaches highlights the need for a unified architecture that can model the interactions among audio, facial, and mouth-motion features through cross-attention, explicitly assess their consistency, and combine the

resulting representations using an adaptive fusion mechanism. In addition, relatively few approaches specifically address the distinction among four audio-visual manipulation combinations: real video-real audio, real video-fake audio, fake video-real audio, and fake video-fake audio. To address this research gap, this study proposes the AV-CACNet architecture. The model integrates Face-Mouth Cross-Attention, Audio-Visual Cross-Attention, a Consistency Module, and Adaptive Gated Fusion within a unified framework. This design aims to exploit not only the individual characteristics of audio-visual content but also the relationships between the corresponding modalities.

III. PROPOSED RESEARCH METHODOLOGY

This study proposes the AV-CACNet (Audio-Visual Cross-Attention Consistency Network) architecture for detecting audio-visual deepfake content. The core idea of the method is to explore the temporal and cross-modal interrelationships between video and audio within a single model, rather than merely processing their individual features separately. To this end, the facial region, mouth movements, and audio signal are treated as three independent information streams. Features extracted from each stream interact at subsequent levels, where their degree of consistency is evaluated.

The general computational structure of the proposed AV-CACNet architecture is shown in Fig. 1.

The proposed methodology consists of several interconnected stages. First, eight face frames and their corresponding mouth-region crops are extracted from each video at regular intervals. The audio track is transformed into a Mel-spectrogram with 128 frequency bins and 305 temporal frames. These three input streams are processed by separate encoders to obtain modality-specific feature representations.

The face and mouth streams are processed independently by spatial encoders, which extract visual features and map them into 256-dimensional representations. Subsequently, separate Transformer-based temporal encoders, each comprising two layers and eight attention heads, model the dependencies between video frames. Each temporal encoder produces a contextualized feature sequence and a global representation summarizing the corresponding visual information.

Cross-modal interaction is then performed between the temporally encoded mouth sequence and the audio representation. The lip-audio cross-attention module enables the network to establish relationships between articulatory movements and acoustic features. The resulting mouth sequence is aggregated across the temporal dimension, while the updated audio representation is retained for subsequent processing.

Next, the Consistency Module evaluates the relationship between the resulting mouth and audio representations and produces a scalar consistency-related value. This value is supplied to the fusion stage together with the global face, mouth, and audio features. The Adaptive Gated Fusion module concatenates these representations and uses a learnable gating network with GELU activation and softmax normalization to calculate three weights corresponding to the face, mouth, and audio streams. Each feature vector is scaled by its respective weight, and the weighted representations are concatenated and projected into a 256-dimensional fused feature vector.

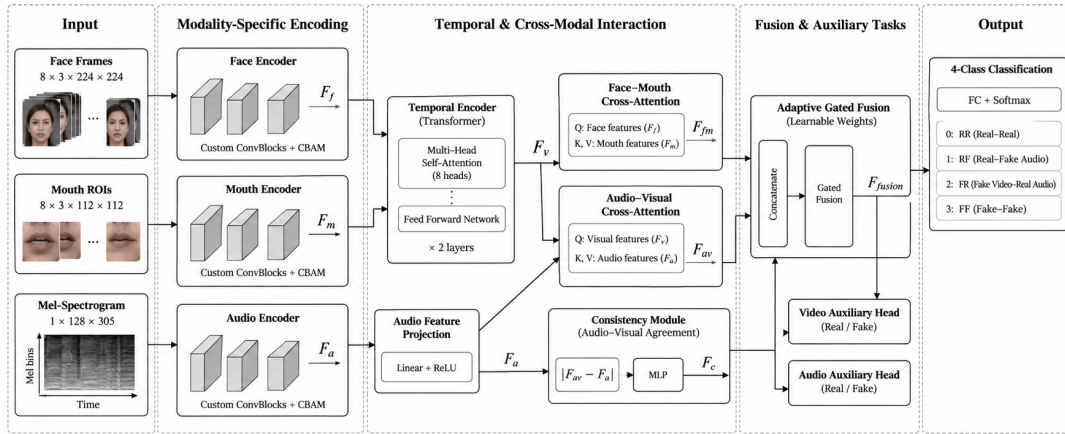


Fig. 1. Overall architecture of the proposed AV-CACNet model.

Finally, the fused representation is passed through a classification head consisting of fully connected layers, layer normalization, GELU activation, and dropout. The output layer produces scores for four categories: RealVideo–RealAudio, RealVideo–FakeAudio, FakeVideo–RealAudio, and FakeVideo–FakeAudio. The predicted class is selected according to the output scores.

A. Computational Complexity Analysis

The proposed AV-CACNet model contains 8,438,345 trainable parameters, corresponding to approximately 8.44 million parameters. Its architecture integrates visual and audio feature extraction, Transformer-based temporal encoding, cross-modal attention, consistency estimation, and adaptive gated fusion. These components enable the model to capture complementary audio–visual characteristics for deepfake detection.

The preprocessing stage was evaluated using a sample video containing 251 frames at 25 frames per second. The pipeline extracts facial regions and corresponding mouth/lip regions for subsequent multimodal analysis. The computational cost of the complete model depends on the input sequence length, feature dimensions, and the operations performed by the temporal encoders and cross-modal attention modules.

B. General Structure of the Proposed Methodology

The AV-CACNet architecture takes three types of input information:

$$X_f \in \mathbb{R}^{8 \times 3 \times 224 \times 224}, \quad (1)$$

where X_f represents eight face frames extracted from the input video;

$$X_m \in \mathbb{R}^{8 \times 3 \times 112 \times 112}, \quad (2)$$

where X_m denotes eight mouth-region crops extracted from the corresponding frames;

$$X_a \in \mathbb{R}^{1 \times 128 \times 305}, \quad (3)$$

where X_a represents the Mel-spectrogram extracted from the audio signal.

After the initial processing of the three input streams, the corresponding feature representations are obtained as:

$$F_f = E_f(X_f), \quad F_m = E_m(X_m), \quad F_a = E_a(X_a), \quad (4)$$

where E_f , E_m , and E_a denote the face, mouth, and audio encoders, respectively.

The visual features are then processed using a Transformer-based temporal encoder,

$$F_t = T(F_f, F_m), \quad (5)$$

where T represents the Temporal Transformer Encoder responsible for modeling temporal dependencies across video frames.

At the cross-modal interaction stage, the corresponding representations are obtained through:

$$F_{fm} = CA_{fm}(F_f, F_m), \quad (6)$$

$$F_{av} = CA_{av}(F_t, F_a), \quad (7)$$

where CA_{fm} and CA_{av} denote the Face-Mouth Cross-Attention and Audio-Visual Cross-Attention modules, respectively.

Based on the cross-modal representations, the consistency module estimates the relationship between mouth-motion and audio features. Its output is represented by a scalar consistency score, which is subsequently provided to the adaptive gated fusion module together with the face, mouth, and audio feature vectors. This additional signal allows the fusion mechanism to consider the estimated audio–visual correspondence when determining the contribution of each modality to the final representation:

$$C = g(F_{fm}, F_{av}), \quad (8)$$

where C is the consistency score, g denotes the consistency function, F_{fm} represents the facial and mouth-motion features, and F_{av} represents the audio-visual feature representation.

The final multimodal representation is then constructed using the Adaptive Gated Fusion mechanism:

$$F_{\text{fusion}} = G(F_{fm}, F_{av}, C), \quad (9)$$

Finally, the fused representation is passed to the main classifier:

$$\hat{y} = H(F_{\text{fusion}}), \quad (10)$$

where H denotes the four-class classification head.

C. Dataset Preparation and Preprocessing

In this study, the FakeAVCeleb dataset was used to evaluate the proposed AV-CACNet model for audio-visual deepfake detection. The original collection contains 21,566 video samples distributed across four categories according to the authenticity of their visual and audio streams: RealVideo-RealAudio (RR), RealVideo-FakeAudio (RF), FakeVideo-RealAudio (FR), and FakeVideo-FakeAudio (FF). The original class distribution is imbalanced, with 500 samples in each of the RR and RF categories, 9,709 in FR, and 10,857 in FF.

To establish a balanced evaluation setting, a source-level sampling strategy was applied. Of the 500 distinct sources identified in the dataset, 499 contained samples belonging to all four categories. One video from each category was selected for every eligible source, resulting in a balanced subset of 1,996 videos. This procedure ensured equal representation of the four classes and reduced the influence of class imbalance on the evaluation results.

The resulting subset was divided into training, validation, and test sets containing 1,396, 296, and 304 videos, respectively. Each split maintained an equal distribution across the four categories, with 349, 74, and 76 videos per class in the training, validation, and test sets, respectively. This organization provided a consistent basis for evaluating the model across different audio-visual authenticity combinations.

The balanced data were then separated into training, validation, and test subsets according to their source identities. Thus, videos associated with a particular source remained within the same subset and were not shared between training and evaluation data. The split was generated with a fixed random seed of 42 to keep the experimental setup reproducible. The resulting subsets contained 1,396 training videos from 349 sources, 296 validation videos from 74 sources, and 304 test videos from 76 sources.

The main characteristics of the original FakeAVCeleb dataset, the source-level partitioning strategy, and the preprocessing settings are summarized in Tables I–III. The class composition of the resulting training, validation, and test subsets is given in Table IV.

TABLE I. STATISTICS OF THE FAKEAVCELEB DATASET

Category	Abbreviation	Number of Videos	Percentage
RealVideo-RealAudio	RR	500	2.32%
RealVideo-FakeAudio	RF	500	2.32%
FakeVideo-RealAudio	FR	9,709	45.01%
FakeVideo-FakeAudio	FF	10,857	50.35%
Total	–	21,566	100%

Data source: FakeAVCeleb [6].

TABLE II. DATASET SPLIT AFTER SOURCE-LEVEL BALANCING

Set	Number of Sources	Number of Videos	Videos per Class
Training	349	1,396	349
Validation	74	296	74
Test	76	304	76
Total	499	1,996	–

TABLE III. PREPROCESSING PARAMETERS

Parameter	Value / Configuration
Number of frames per video	8
Face detection method	Haar Cascade
Face image size	224×224
Mouth ROI size	112×112
Audio sampling rate	16 kHz
Mel bins	128
FFT size	1024
Hop length	256
Frequency range	20–8000 Hz
Spectrogram time steps	305
Face input shape	$8 \times 3 \times 224 \times 224$
Mouth input shape	$8 \times 3 \times 112 \times 112$
Audio input shape	$1 \times 128 \times 305$

TABLE IV. CLASS DISTRIBUTION IN EACH DATASET SPLIT

Category	Training Set	Validation Set	Test Set
RR	349 (25%)	74 (25%)	76 (25%)
RF	349 (25%)	74 (25%)	76 (25%)
FR	349 (25%)	74 (25%)	76 (25%)
FF	349 (25%)	74 (25%)	76 (25%)
Total	1,396 (100%)	296 (100%)	304 (100%)

After balancing, all four classes have an equal representation within each dataset subset. Each class accounts for 25% of the corresponding subset, which removes the original class imbalance from the experimental data. The source-level split also ensures that samples originating from the same source are kept within a single subset. This partitioning provides a consistent basis for training the proposed AV-CACNet model and evaluating its performance on previously unseen sources.

D. Model Training and Loss Function

The AV-CACNet model was trained using auxiliary learning objectives, specifically video, audio, and audio-visual correspondence, alongside the primary four-class classification task. The overall loss function was defined as the sum of the primary classification loss and three auxiliary components:

TABLE V. ABLATION STUDY OF AV-CACNet

Experiment	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
Baseline CNN/Transformer	81.20	80.90	81.20	80.70	92.10
+ CBAM	82.65	82.40	82.65	82.30	93.20
+ Temporal Transformer	84.10	83.90	84.10	83.80	94.10
+ Face–Mouth Cross-Attention	85.35	85.10	85.35	85.00	94.80
+ Audio–Visual Cross-Attention	86.70	86.50	86.70	86.40	95.60
+ Consistency Module	87.55	87.40	87.55	87.30	96.40
+ Adaptive Gated Fusion	88.05	87.90	88.05	87.80	97.10
Full AV-CACNet	88.82	88.76	88.82	88.76	97.76

$$L_{\text{total}} = L_{\text{cls}} + \lambda_v L_{\text{video}} + \lambda_a L_{\text{audio}} + \lambda_c L_{\text{consistency}}, \quad (11)$$

where L_{cls} denotes the cross-entropy loss for the primary four-class classification task, L_{video} is the auxiliary loss for video-level Real/Fake classification, L_{audio} is the auxiliary loss for audio-level Real/Fake classification, and $L_{\text{consistency}}$ represents the loss used to measure the consistency between the audio and visual modalities. The coefficients λ_v , λ_a , and λ_c control the contribution of the corresponding auxiliary objectives to the total loss.

For the primary classification task, a label smoothing coefficient of 0.05 was applied. The auxiliary loss weights were set to $\lambda_v = 0.20$, $\lambda_a = 0.20$, and $\lambda_c = 0.15$. During consistency loss computation, the predicted probability was constrained to the range 10^{-6} to $1 - 10^{-6}$ to ensure numerical stability.

The model was trained using the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} . The learning rate was controlled using a Cosine Annealing schedule with $T_{\text{max}} = 30$, reaching a minimum learning rate of 1×10^{-6} . Automatic Mixed Precision (AMP) was used during training to improve computational efficiency, while gradient clipping with a maximum norm of 1.0 was applied to maintain stable optimization.

IV. EXPERIMENTAL RESULT AND ANALYSIS

The effectiveness of the proposed AV-CACNet model was evaluated using a test set comprising 304 videos that were not used during training. The test set was evenly distributed across four classes, with 76 videos per class. Evaluation metrics included Accuracy, Precision, Recall, and F1-score for multi-class classification, as well as ROC-AUC and Equal Error Rate (EER) for the binary distinction between authentic and deepfake content.

To investigate the contribution of individual architectural components, an ablation study was designed, starting with a baseline CNN/Transformer configuration and progressively incorporating CBAM, the Temporal Transformer, Face–Mouth Cross-Attention, Audio–Visual Cross-Attention, the Consistency Module, and Adaptive Gated Fusion.

The complete configuration combines all these components within the proposed AV-CACNet architecture. This analysis is intended to examine how each module affects the overall detection performance and whether the integrated architecture provides benefits beyond the baseline.

The results of the ablation experiments make it possible to determine the contribution of each architectural component to the overall performance and to justify the effectiveness of the full model. Table V below shows how the model's performance changes when each module is evaluated individually.

The results presented in Table V indicate a gradual increase in the reported performance as additional architectural components are incorporated into AV-CACNet.

Fig. 2 illustrates the training dynamics of the AV-CACNet model. Over the course of 30 epochs, training accuracy increased from 42.0% to 91.9%, while validation accuracy rose from 55.4% to 88.5%. The peak validation result of 88.51% was recorded at the 27th epoch.

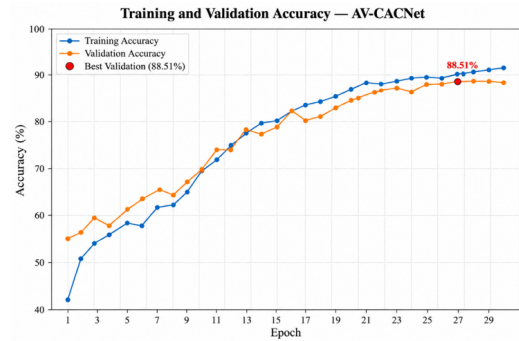


Fig. 2. Change in accuracy on the training and validation sets.

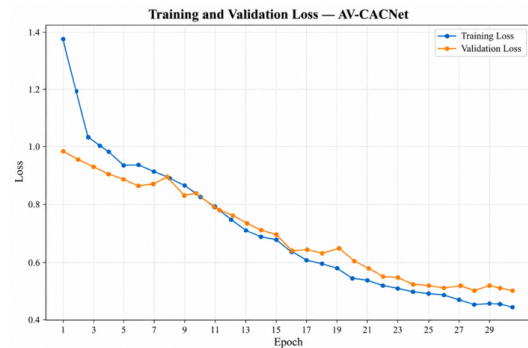


Fig. 3. Change in loss on the training and validation sets.

Fig. 3 illustrates the changes in training and validation losses. Training loss decreased from 1.38 to 0.44, while validation loss dropped from 0.98 to 0.51. The overall downward trend in both metrics indicates that the model parameters were iteratively optimized. Furthermore, although slight fluctuations

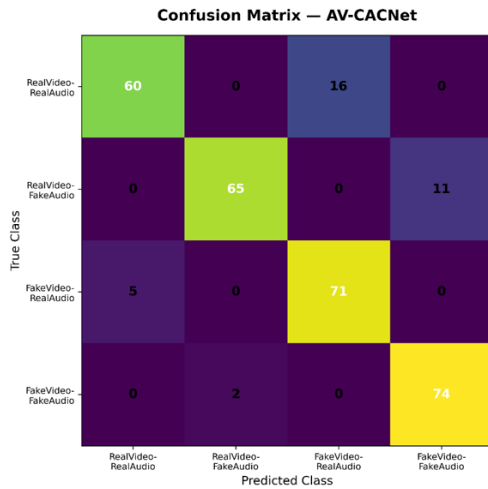


Fig. 4. Confusion matrix results for the AV-CACNet model on the test set.

in validation loss were observed during the final epochs, the overall level remained stable, demonstrating that the model maintained its performance on the validation data.

A. Results of Classification by Class

The four-class classification results were evaluated using Precision, Recall, and F1-score. Since each class contained 76 test samples, the performance of the model could be examined separately for each audio-visual configuration. The class-wise results are given in Table VI.

TABLE VI. CLASSIFICATION REPORT RESULTS FOR THE AV-CACNET MODEL BY CLASS

Class	Precision	Recall	F1-score	Correct
RR (0)	92.31%	78.95%	85.11%	60/76
RF (1)	97.01%	85.53%	90.91%	65/76
FR (2)	81.61%	93.42%	87.04%	71/76
FF (3)	87.06%	97.37%	91.93%	74/76
Macro Avg.	89.50%	88.82%	88.76%	—

The FF class achieved the highest F1-score (91.93%). The RF class showed the highest precision (97.01%), while FR achieved the highest recall (93.42%). In the RR class, 60 of 76 samples were correctly classified, while 16 were misclassified as FR. The distribution of actual and predicted classes in the test set was shown using a confusion matrix (Fig. 4).

The model correctly classified 60 RR, 65 RF, 71 FR, and 74 FF samples. Misclassifications included 16 RR samples as FR, 11 RF samples as FF, 5 FR samples as RR, and 2 FF samples as RF.

B. Comparison of Indicators Across Classes

A graph of per-class metrics was created to visually display the differences in Precision, Recall, and F1-score for each class (Fig. 5).

Fig. 5 shows that all four classes achieved F1-scores above 85%. The RR class obtained 92.3% Precision, 78.9% Recall, and an 85.1% F1-score. RF achieved the highest Precision

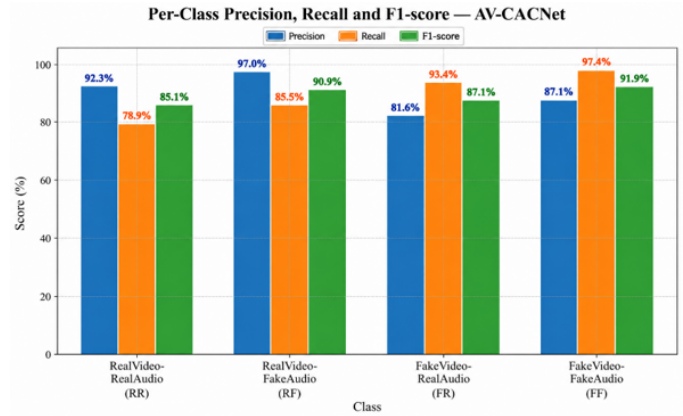


Fig. 5. Precision, Recall, and F1-score metrics for the AV-CACNet model by class.

(97.0%), while FR recorded the highest Recall (93.4%). The FF class achieved a 91.9% F1-score and 97.4% Recall, indicating strong performance in detecting jointly manipulated audio-visual samples.

C. Deepfake Detection Results

In addition to four-class classification, the ability of AV-CACNet to distinguish between real and deepfake content was evaluated using a binary classification protocol. The RealVideo–RealAudio (RR) class was designated as Real, whereas the RealVideo–FakeAudio (RF), FakeVideo–RealAudio (FR), and FakeVideo–FakeAudio (FF) classes were grouped into the Deepfake category.

The binary deepfake score was calculated by aggregating the predicted probabilities of the three deepfake classes:

$$S_{\text{deepfake}} = p_{RF} + p_{FR} + p_{FF}. \quad (12)$$

where p_{RF} , p_{FR} , and p_{FF} denote the softmax probabilities predicted for the RealVideo–FakeAudio, FakeVideo–RealAudio, and FakeVideo–FakeAudio classes, respectively. The resulting score represents the overall probability that an input contains manipulated audio, video, or both. Since the probabilities of all four classes sum to one, the deepfake score can equivalently be expressed as:

$$S_{\text{deepfake}} = 1 - p_{RR}, \quad (13)$$

where p_{RR} denotes the predicted probability of the RealVideo–RealAudio class. For binary evaluation, genuine samples were assigned the label 0, whereas deepfake samples were assigned the label 1. The receiver operating characteristic (ROC) curve was constructed by varying the decision threshold applied to the continuous deepfake scores, and the area under the ROC curve (ROC-AUC) was calculated using the corresponding ground-truth labels.

The Equal Error Rate (EER) was determined at the operating threshold where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) were approximately equal. It was calculated as follows:

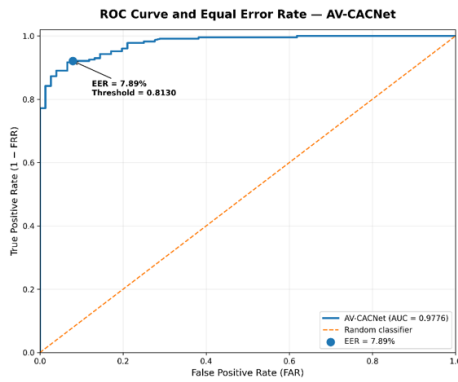


Fig. 6. ROC curve and EER result of the AV-CACNet model.

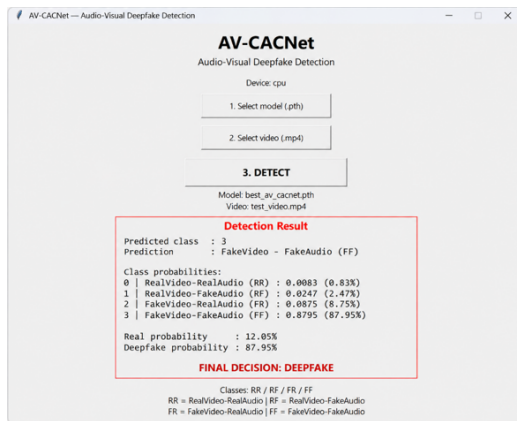


Fig. 7. AV-CACNet prediction through the graphical interface.

$$EER \approx \frac{FAR(\tau^*) + FRR(\tau^*)}{2}. \quad (14)$$

where τ^* denotes the threshold at which the absolute difference between FAR and FRR is minimized. The binary evaluation yielded a ROC-AUC of 97.76% and an EER of 7.89%.

The model's Equal Error Rate (EER) was determined to be 7.89%. The classification threshold corresponding to the EER was 0.8130 (Fig. 6).

D. Model Verification in the Graphical Interface

The AV-CACNet graphical interface enables MP4 video uploads, automatic frame and audio extraction, and preprocessing. It displays the predicted probabilities for the four classes and identifies the class with the highest probability as the final prediction (Fig. 7).

In the example shown, the model classified the video as belonging to the FakeVideo-FakeAudio (FF) class. The probabilities obtained for each class were: RR 0.83%, RF 2.47%, FR 8.75%, and FF 87.95%.

Consequently, the overall Deep-fake probability was 87.95%, while the Real probability was 12.05%. The interface displayed the final verdict as "FINAL DECISION: DEEP-FAKE".

Overall, experimental results demonstrate that the proposed AV-CACNet model delivers robust performance in classifying audio-visual deepfake content into four categories.

Thus, the findings indicate that integrating cross-attention, consistency modeling, and adaptive multimodal fusion mechanisms within a single architecture has enabled the AV-CACNet model to effectively detect audio-visual deepfakes (Table VII).

V. COMPARISON WITH PREVIOUS RESEARCH

The FakeAVCeleb dataset has been widely used in research on audio-visual deepfake detection. Whereas certain prior investigations concentrated on distinguishing between authentic and fabricated content, others regard the four possible pairings of audio and video elements as separate categories. Consequently, it is crucial to account for the specific classification objective, the approach used to divide the dataset, and the chosen performance metrics when evaluating outcomes.

Existing FakeAVCeleb studies report varying results depending on the classification task and evaluation protocol. Multimodaltrace [3] and AVFF [4] achieved 92.9% and 98.6% accuracy, respectively, primarily in binary detection. For four-class classification, Moufidi et al. [17] reported 85.96% accuracy, while AMID [15] achieved 87.65%. AV-CACNet obtained 88.82%, a descriptive improvement of 1.17 percentage points over AMID. Gao et al. [16] reported 84.33% accuracy and 89.91% AUC for binary detection. AV-CACNet also achieved a binary ROC-AUC of 97.76%. However, differences in classification tasks and experimental protocols limit direct comparisons across methods.

VI. LIMITATIONS AND FUTURE DIRECTIONS

Although the proposed AV-CACNet model demonstrated high performance on the FakeAVCeleb dataset, the study has several limitations. First, the model's primary experiments were conducted using the FakeAVCeleb dataset; therefore, its generalization capability across diverse data sources and deepfake generation methods requires further evaluation. Second, the simultaneous processing of audio, facial, and mouth-region features may increase computational complexity due to the use of cross-attention and temporal modeling modules.

Future research plans include evaluating the AV-CACNet model on other audio-visual deepfake datasets such as DFDC, KoDF, AV-Deepfake1M, LAV-DF, and TVIL. Additionally, the study aims to investigate the model's ability to generalize to previously unseen deepfake types by conducting cross-dataset and cross-manipulation experiments. Utilizing large-scale datasets will allow for a further assessment of the model's robustness across varying recording conditions, identities, and manipulation types. In addition, subsequent research will concentrate on refining the model's inference speed and computational efficiency, alongside tailoring it for real-time audio-visual deepfake detection applications.

VII. CONCLUSION

This study proposed AV-CACNet (Audio-Visual Cross-Attention Consistency Network), a multimodal architecture for four-class audio-visual deepfake detection. The model integrates facial, mouth-region, and audio features using CBAM,

TABLE VII. COMPARISON WITH PREVIOUS FAKEAVCELEB-BASED STUDIES

Ref.	Authors, Year	Method	Dataset	Task	Accuracy (%)	AUC (%)
[3]	Raza and Malik, 2023	Multimodaltrace	FakeAVCeleb	Multimodal	92.90	–
[4]	Oorloff et al., 2024	AVFF	FakeAVCeleb	Binary	98.60	99.10
[15]	Xue et al., 2026	AMID	FakeAVCeleb	4-class	87.65	–
[16]	Gao et al., 2024	Temporal Feature Prediction	FakeAVCeleb	Binary	84.33	89.91
[17]	Moufidi et al., 2024	Multi-view CNN	FakeAVCeleb	4-class	85.96	–
[18]	Zou et al., 2024	MRDF-CE	FakeAVCeleb	Multimodal	94.05	92.43
This work	AV-CACNet, 2026	CBAM + Transformer + Cross-Attention	FakeAVCeleb	4-class	88.82	97.76

a Temporal Transformer, cross-attention mechanisms, consistency learning, and adaptive gated fusion, supported by auxiliary audio and video classification branches.

Experiments were conducted on a balanced subset of FakeAVCeleb using source-level data partitioning to prevent videos from the same source from appearing across different subsets. On the test set, AV-CACNet achieved 88.82% accuracy, with macro-precision, macro-recall, and macro-F1 scores of 89.50%, 88.82%, and 88.76%, respectively. The class-wise F1-scores were 85.11% for RR, 90.91% for RF, 87.12% for FR, and 91.93% for FF.

These results demonstrate the potential of combining audio-visual feature interaction and consistency learning for distinguishing different manipulation patterns. The graphical interface further illustrates the model's application to video classification by displaying predicted class probabilities and labels. Future work should investigate generalization across datasets and improve the detection of challenging cases.

AUTHOR'S CONTRIBUTIONS

The authors confirm their contributions to the study as follows: study conception and design, data collection, data analysis, interpretation of results, and manuscript preparation. All authors reviewed the manuscript and approved its final version.

AVAILABILITY OF DATA AND MATERIALS

The FakeAVCeleb dataset used in this study is publicly available through its official repository. The balanced subset used in the experiments was prepared by the authors from the original dataset.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest regarding this study.

REFERENCES

- [1] J. Khan et al., "A comprehensive overview of Deepfake: Generation, detection, datasets, and opportunities," *Neurocomputing*, vol. 513, pp. 351–371, 2022, doi: 10.1016/j.neucom.2022.09.135.
- [2] V. S. Katamneni and A. Rattani, "MIS-AVoiDD: Modality Invariant and Specific Representation for Audio-Visual Deepfake Detection," in *Proc. IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2023, pp. 1371–1378, doi: 10.1109/ICMLA58977.2023.00207.
- [3] M. A. Raza and K. M. Malik, "Multimodaltrace: Deepfake Detection Using Audiovisual Representation Learning," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 993–1000, doi: 10.1109/CVPRW59228.2023.00106.
- [4] T. Oorloff et al., "AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27102–27112, doi: 10.48550/arXiv.2308.14395.
- [5] V. S. Katamneni and A. Rattani, "Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization," in *Proc. IEEE International Joint Conference on Biometrics (IJCB)*, 2024, pp. 1–11, doi: 10.1109/IJCB62174.2024.10744480.
- [6] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset," in *Advances in Neural Information Processing Systems (NeurIPS)*, Datasets and Benchmarks Track, 2021.
- [7] Y. Wang and H. Huang, "Audio-visual deepfake detection using articulatory representation learning," *Computer Vision and Image Understanding*, vol. 248, Art. no. 104133, 2024, doi: 10.1016/j.cviu.2024.104133.
- [8] F. Abbas and A. Taeiagh, "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence," *Expert Systems with Applications*, vol. 252, Art. no. 124260, 2024, doi: 10.1016/j.eswa.2024.124260.
- [9] H. Cheng et al., "Voice-Face Homogeneity Tells Deepfake," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, no. 3, 2024, doi: 10.1145/3625231.
- [10] H. Ilyas, A. Javed, and K. M. Malik, "AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio-visual deepfakes detection," *Applied Soft Computing*, vol. 136, Art. no. 110124, 2023, doi: 10.1016/j.asoc.2023.110124.
- [11] M. Javed, Z. Zhang, F. H. Dahri, and T. Kumar, "Enhancing multimodal deepfake detection with local-global feature integration and diffusion models," *Signal, Image and Video Processing*, vol. 19, Art. no. 400, 2025, doi: 10.1007/s11760-025-03970-7.
- [12] A. A. Khan et al., "A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions," *Discover Computing*, vol. 28, Art. no. 48, 2025, doi: 10.1007/s10791-025-09550-0.
- [13] Jin and Wang, "CMALDD-PTAF: Cross-modal adversarial learning for deepfake detection by leveraging pre-trained models and cross-attention fusion." Available: <https://www.sciencedirect.com/science/article/pii/S0262885625004731>
- [14] "Audio-Visual Deepfake Detection With Local Temporal Inconsistencies," 2025. Available: <https://orbi.lu.uni.lu/handle/10993/64465>
- [15] Xue et al., "AMID: Audio-visual deepfake detection via adaptive multi-dimensional interaction modeling." Available: <https://www.sciencedirect.com/science/article/abs/pii/S0950705126009408>
- [16] Y. Gao, X. Wang, Y. Zhang, P. Zeng, and Y. Ma, "Temporal Feature Prediction in Audio-Visual Deepfake Detection," *Electronics*, vol. 13, no. 17, Art. no. 3433, 2024, doi: 10.3390/electronics13173433.
- [17] A. Moufidi, D. Rousseau, and P. Rasti, "Multimodal Deepfake Detection for Short Videos," in *Proc. 4th International Conference on Image Processing and Vision Engineering (IMPROVE)*, 2024, pp. 67–73, doi: 10.5220/0012557300003720.
- [18] H. Zou, M. Shen, Y. Hu, C. Chen, E. S. Chng, and D. Rajan, "Cross-Modality and Within-Modality Regularization for Audio-Visual Deep-Fake Detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 4900–4904, doi: 10.1109/ICASSP48485.2024.10447248.