NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Applied Mathematics and Informatics"

Research Project Report on the Topic:

Multimodal Systems for Deepfake Video Detection

Submitted by the Student:

group #БПМИ231, 2nd year of study — Larin Ilya Aleksandrovich

Approved by the Project Supervisor:

Grinberg Petr Markovich
Visiting Teacher, HSE
Faculty of Computer Science, HSE University

Moscow 2025

Contents

Abstract

Generative models continue to advance, so fraudsters have access to increasingly advanced falsification methods. However, changing only the speech or appearance of a person may not be enough to convince the audience. That is why deepfakes usually alter both modalities: audio (speech) and visual (appearance). To enhance the accuracy of deepfake detection, it is important to make use of both modalities. In this study we explore multi-modal deepfake detection methods and create our own model. The solution we developed leverages advanced feature fusion, which enables it to outperform multiple other multi-modal approaches and most uni-modal solutions.

Аннотация

С развитием генеративных моделей, у мошенников появляется доступ к все более и более продвинутым методам фальсификации. Однако изменение лишь речи или внешности персоны может быть недостаточно для того, чтобы люди поверили в обман. Поэтому часто для фальсификации используются две модальности: звук (речь) и видео (лицо человека). Для улучшения качества обнаружения подобных мошеннических схем стоит также использовать обе модальности. В данной работе студенту предлагается исследовать методы мультимодального обнаружения фальсификаций и создать свою модель. Разработанное нами решение выступает лучше некоторых мультимодальных подходов и достигает результатов большинства унимодальных решений.

Keywords

Deepfake detection, Audio-visual deepfakes, Deep learning, Computer vision

1 Introduction

It is an arguable question if developing technologies is good or bad and how they affect our daily lives. But nobody can argue that spreading fake information may cause serious damage to a particular person, group of people, governments and whole countries. Nowadays fraudsters may take advantage of advanced Artificial Intelligence (AI) systems powered by Deep Learning (DL) tools to generate realistic fake content such as texts, videos, images and audio [11], [20] [15]. We have already reached the stage when its quality may convince audience and make people believe in a lie [28].

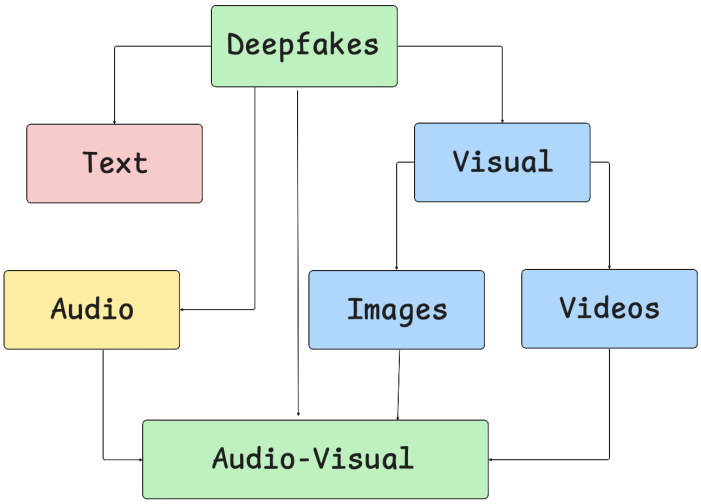

Figure 1.1: Types of deepfakes. Inspired by [9]

In this particular paper, we focus on audio-visual deepfakes, that also include audio deepfakes and video deepfakes as shown in Figure 1.1. Audio-visual deepfakes, in turn, can be divided in smaller groups by affected modalities, as shown in Figure 1.2, depending on which ones were affected or not. Every type but RARV (real audio and real video) are considered deepfakes.

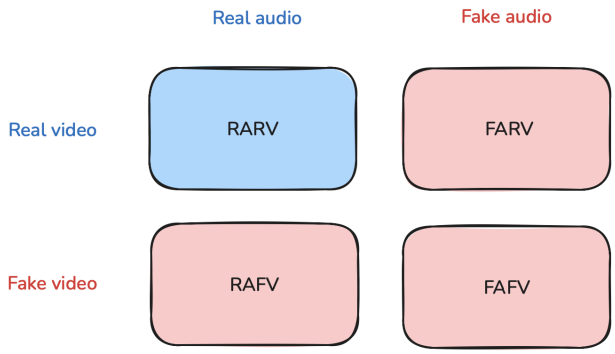

Figure 1.2: Media divided in four groups by two modalities (audio and video). Inspired by [9]

There is an analogue of this division in sphere of deepfake detection: models that only

focus on audio deepfakes and models that only focus on visual deepfakes. If model is able to somehow combine both modalities the accuracy enhances, because it gets more information to proceed: instead of one stream (only audio or only visual) it gets as an input three – both audio and visual and also alignment between two modalities. Such an approach gives model more space for detecting deepfake. This hypothesis is proven by recent studies [21], [10], [31]. High accuracy and capability to identify various forms of synthetic media is the reason why we focus on developing a multi-modal system for audio-visual deepfake detection. In this research, we use self-supervised model AV-HuBERT [24] for feature extraction. It was shown [16], [19] that utilizing models pre-trained on large data is beneficial for improving the quality and generalization capabilities of deepfake classifiers. As classifier we take graph-based AMSDF [30] classifier that uses advanced feature fusion and has achieved state-of-the-art results in audio deepfake detection. Using this approach we achieve high results in audio-visual deepfake detection and outperform multiple multi-modal solutions trained in the same environment and most uni-modal models. We present an open-source solution and all the code we used for model training is available at our official GitHub repository[1].

2 Related work

As noted, using big SSL pre-trained models enhances performance so that we adopt the AV-HuBERT model. The same idea is used in the AV-Lip-Sync+ paper.

The AV-Lip-Sync+ [23] paper presents an approach for spoofing audio-visual deepfakes. This section focuses on the feature extraction process and models that are used for it.

The authors suggest focusing not on spectro-temporal relationships, but on multi-view synchronization. For such needs, they use the AV-HuBERT model for feature extraction. They get audio-only (F_a), visual-only (F_v) and audio-visual (F_l) features. Audio-visual features are kept as is, when audio-only and video-only are subtracted to obtain synchronization features:

$$\vec{F}_{sync} = \{|\vec{F}_{ai} - \vec{F}_{vi}|\}_{i=1}^{T}, \tag{1}$$

where T is the number of frames. After that F_l and F_{sync} are concatenated along hidden dimension and parsed with MS-TCN [6], then pooling and linear layers to get probabilities of video being real or fake.

However, the one big disadvantage of such an approach is that AV-HuBERT model works

[1]Official implementation: https://github.com/runtime57/deepfake_detection

only with the lip sequence, so it is hard to identify deepfake methods, that do not affect mouth region very much (for example, faceswap [15]). To solve this problem authors propose adding facial feature extractor and using ViViT [2] for it. ViViT features are a single-dimensional vector, they concatenate it with MS-TCN output before linear layer. With such an approach they achieved state-of-the-art results.

We can divide existing methods of multi-modal audio-visual deepfake detection into three main groups: Independent learning, Joint Learning and Matching-based learning. These types are visualized in Figure 2.1.

Independent learning. Basic idea is presented in [32]. Two or more independent streams of convolutional neural networks make predictions about source being audio or video deepfake. Predictions from each of the sources are combined to get the resulting one, that is final decision. Khalid et al. [13] ensemble predictions of uni-modal detectors to evaluate FakeAVCeleb [14]. Hashmi et al. [8] extend a similar system by adding a third stream that considers both modalities.

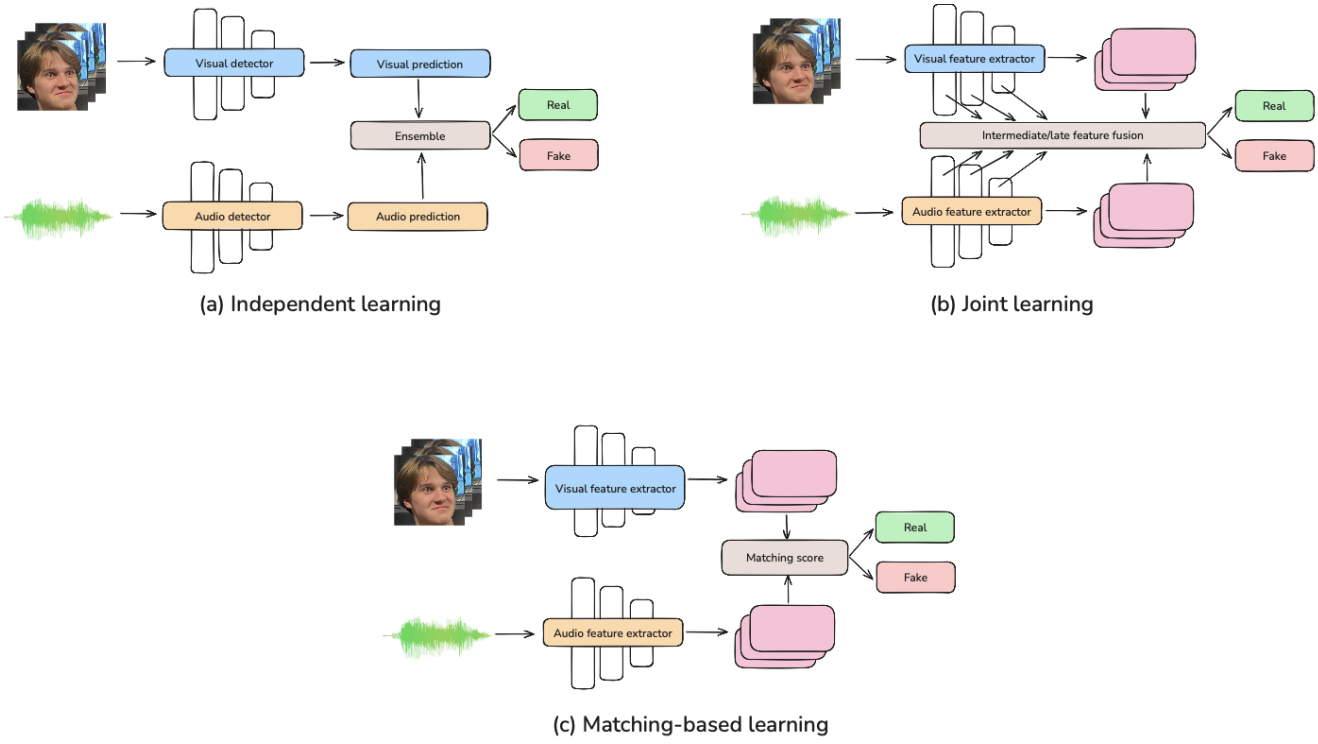

Figure 2.1: Three main paradigms for multi-modal audio-visual deepfake detection. Inspired by [17]

Matching-based learning. This method is based on natural inter-modal relationship or, in other words, synchronization between visual and audio streams that is presents real videos and usually absent in synthesized ones. Cheng et al. [4] and Tian et al. [26] use similar idea of computing function that shows how synchronized two modalities are. The resulting value is compared to a threshold and based on the comparison video is called fake or real. Marcella et

al. [3] present new loss function, which is maximized for fake data and minimized for real data. It measures the difference between first-order statistics of visual and audio feature distribution.

Joint learning. This paradigm uses as a basis idea of integrating audio and visual modalities. Two separated feature extractors proceed different streams, then fusion module integrates these two features for audio-visual deepfake detection. As shown in Figure 2.1, there are intermediate fusion (combining high-level abstract features) and late fusion (combining in the final stages). Raza et al. [22] use late fusion idea applying MLP mixer [27] layer to enhance effective fusion.

In AMSDF paper Sahibzada Adil Shahzad et al. present a new state-of-the-art approach for audio spoofing detection. It can be split into three parts: feature extraction, Multi-view Correlation Measurement Network (MCMN), classification head. MCMN is based on Audio-Text-Emotions correlations and uses advances of GATs [29] to better find dependencies between modalities and highlight forgery cues.

Firstly, they apply intra-view graph attention mechanism (IGAM) to extracted features, then three graphs are connected in pairs and all together. Next, HGFM is applied to every group of graphs (A-T, A-E, T-E, A-T-E where A stands for Audio, T stands for text and E - for emotions). It gives us M_m (master node for modality m) and H_m (heterogeneous graph). The classifier is essentially a GRS operation applied to groups $\{M_m| \forall m\}$, $\{G_m| \forall m\}$, $\{H_m| \forall m\}$. GRS operation derives D-dimensional features, that are concatenated and pushed through MLP.

3 Methodology

Inspired by AV-Lip-sync+ and AMSDF, we propose a new solution with enhanced fusion of modalities.

3.1 Feature extraction

We extract features using three different models: AV-HuBERT pre-trained on LRS3 [1] and VoxCeleb2 [5] datasets for audio-visual features (F_l), ViViT pre-trained on Kinetics [33] dataset for facial features (F_v), and AASIST encoder (which we train from scratch) for audio features (F_a). As in [30] we use ViViT to capture not lip sequence only, but full video and give ourselves more space for spoofing deepfake. For modality m we obtain feature F_m of shape $T_m \times D_m$. Preprocessing and feature extraction are illustrated in Figure 3.1 and closely described in Section 4.2.

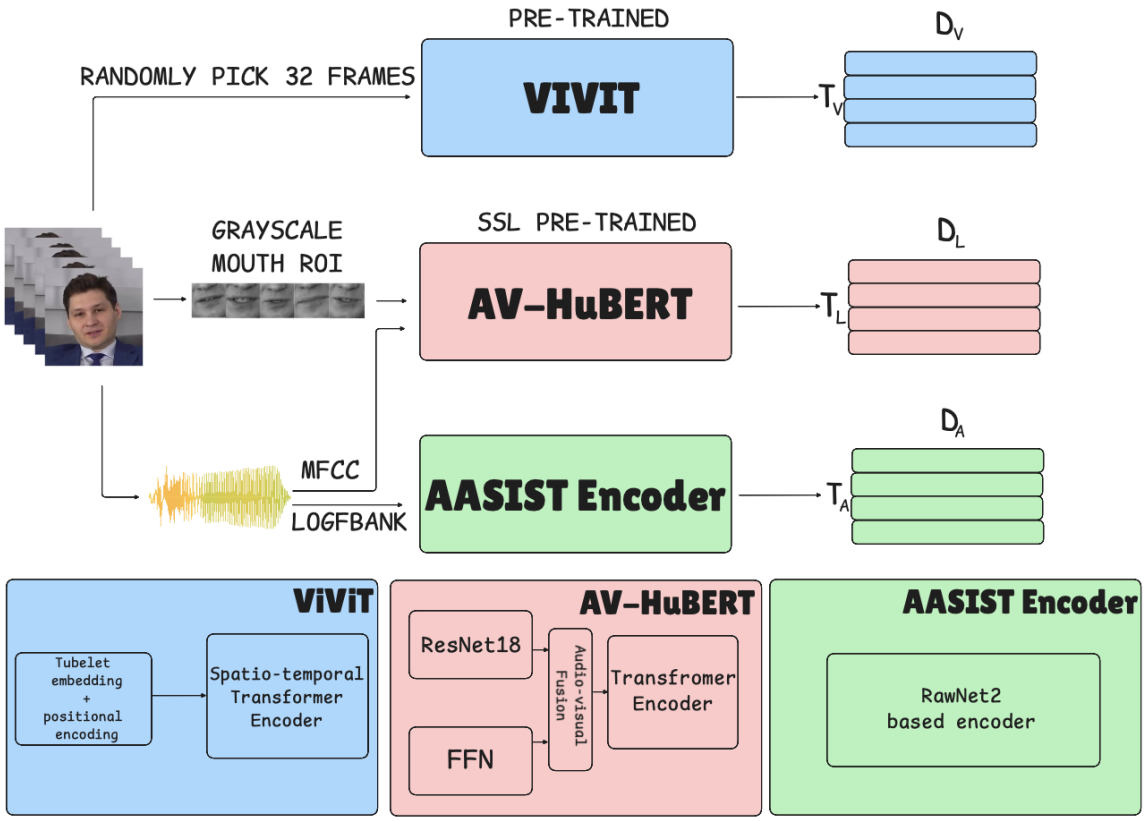

Figure 3.1: Feature extraction structure illustrated.

3.2 Classification head

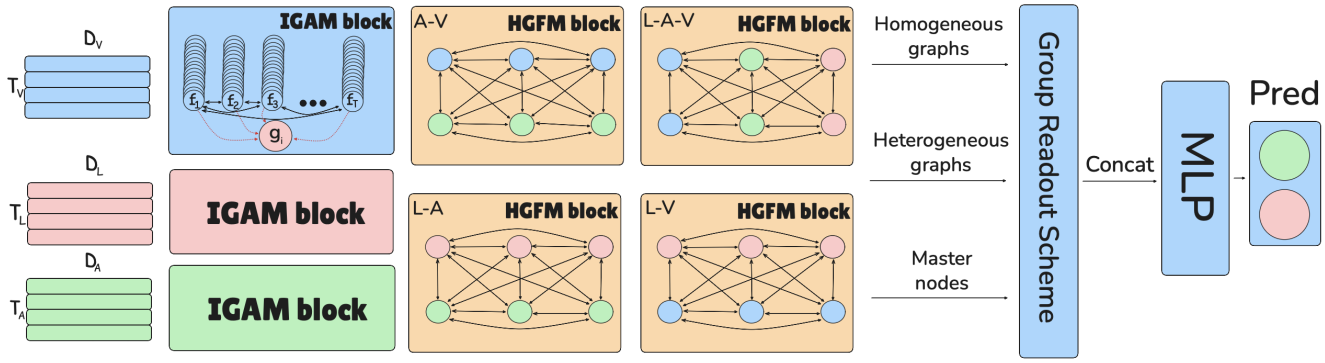

Figure 3.2: Pipeline illustrated. Inspired by [12] and [30].

This part of the pipeline is illustrated in Figure 3.2. We use HGFM-based classification head designed in [12] and modified in [30]. The main advantage of this approach is that we first apply intra-modality attention for each $m \in \{L, A, V\}$ and so that we can spot forgeries inside one modality. Then we merge modalities in every possible way (L+A+V, L+A, L+V, A+V) and apply inter-modality attention to create space for spotting strange dependencies between different modalities.

Firstly, we build separate graphs for all modalities. Assuming we work with modality m

and have features $F_m \in \mathbb{R}^{T_m \times D_m}$, we create T_m nodes and i-th node is associated with a vector $f_m^i \in \mathbb{R}^{D_m}$. We apply GATs [29] to this graph to update nodes using their neighbours. In this way, weights are assigned for relationships between nodes.

$$\alpha_m^{i,j} = softmax(W_m \cdot (f_m^i \odot f_m^j)), \tag{2}$$

where $W_m \in \mathbb{R}^{1 \times D_m}$ is a learnable map and α_m is the attention weight. Let the output of the GAT operation be homogeneous graph $G_m = \{g_m^1, ..., g_m^{T_m}\}$. Then g_m^i is computed as follows.

$$g_m^i = \text{SELU}(W_{\text{neigh}} \cdot (\sum_{j=1}^{T_m} \alpha_m^{i,j} f_m^j) + W_{\text{proj}} \cdot f_m^i), \tag{3}$$

where W_{neigh} and W_{proj} are learnable maps used for, respectively, transforming aggregated information from neighbouring nodes and projecting the nodes from original graph, and SELU$(\cdot)$ is the activation function. This step provides an updated graph that contains information about local relationships between nodes inside one modality.

Next, we apply the HGFM operation to every group of graphs. This operation uses an attention mechanism to find correlations between different modalities. For example, we can take L+A group. We create a heterogeneous graph $H_{\text{L+A}}$ that contains $T_L + T_A$ nodes and every node has dimension of d. We numerate nodes of graph $H_{\text{L+A}}$ in the given order: first nodes of graph G_{L} with indices from 1 to T_{L}, then nodes of graph G_A with indices from $T_{\text{L}} + 1$ to $T_{\text{L}} + T_{\text{A}}$. Node $h_{\text{L+A}}^i \in H_{\text{L+A}}$ is defined as:

$$h_{\text{L+A}}^i = \begin{cases} W_{\text{L}} \cdot g_{\text{L}}^i & \text{for } i \leq T_{\text{L}} \\ W_{\text{A}} \cdot g_{\text{A}}^{i-T_{\text{L}}} & \text{else} \end{cases} \tag{4}$$

We apply GAT to this graph as we do in an IGAM, but in a more complex form to handle multiple modalities. We update nodes the same way as described in eq. (3). To calculate α_i, weights are adjusted with learnable vectors $\lambda_{\text{L+A}}, \lambda_{\text{L+L}}, \lambda_{\text{A+A}}$ for different types of connected nodes (corresponding to cross-modality connections, audio-visual connections and audio-only connections).

We create a master node that aggregates all available information from each group. Let $m_{\text{L+A}}$ be the mean of all the nodes from our graph $H_{\text{L+A}}$:

$$m_{\text{L+A}} = \text{mean}_i\left(h_{\text{L+A}}^i\right). \tag{5}$$

For $m_{\text{L+A}}$ we apply the transformation from eq. (3) (we calculate new α_{L+A}^m using eq. (2), but replacing f_m^i with $m_{\text{L+A}}$).

We obtain the representation of our information in different forms: homogeneous graphs G_{L}, G_{L}, G_{L}, heterogeneous graphs $H_{\mathrm{L+A}}$, $H_{\mathrm{L+V}}$, $H_{\mathrm{A+V}}$, $H_{\mathrm{L+A+V}}$ and master nodes $m_{\mathrm{L+A}}$, $m_{\mathrm{L+V}}$, $m_{\mathrm{A+V}}$ and $m_{\mathrm{L+A+V}}$.

Homogeneous graphs contain information about local intra-modality relationships. Heterogeneous graphs encode information about local inter-modality relationships. We have four groups of modalities which provide our approach greater capacity for detecting forgeries. Moreover,for each heterogeneous graph there is a master node that captures global inter-modality information. This allows the model not to focus on relationships between individual nodes, but to consider each group of modalities as a whole.

We combine homogeneous graphs, heterogeneous graphs and master nodes into three groups accordingly. Then every group is passed to the GRS block. For example, we have a graph $Q \in \mathbb{R}^{T_Q \times D_Q}$ that contains T_Q nodes and each node has a dimension of D_Q. Then GRS operation, firstly, pools T most important nodes. The importance score $I_Q \in \mathbb{R}^{T_Q \times 1}$ is calculated using a learnable weight matrix W_Q:

$$I_Q = \text{sigmoid}\left(W_Q^T \cdot Q\right) \tag{6}$$

This procedure standardizes all graphs to the same shape. We apply a dot product operation with the corresponding importance score for each of the T pooled nodes to assign their weights.

In every group after pooling operation all graphs are merged together. Then the results of node-wise maximum and node-wise averaging operations are concatenated. We concatenate obtained representations into a single vector with shape of $3 \cdot T$, a dropout layer with rate 0.5 is applied, and the vector is passed to an MLP classifier, that produces a probability vector indicating if input data is a deepfake or not.

3.3 Evaluation metrics

We use four metrics to evaluate our model: Precision, Recall F1 score and Accuracy.

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

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$\text{F1 score} = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{10}$$

where TP, TN, FP, FN are respectively True Positive, True Negative, False Positive and False Negative predictions.

4 Experimental setup

4.1 Dataset

We use FakeAVCeleb [14] dataset for both: training and evaluation. It is a standard benchmark and it contains both audio and visual modalities unlike most other datasets (for example, VoxCeleb1). The dataset information is provided in Table 4.1.

Table 4.1: Dataset information

	Split	real	faceswap	fsgan	rtvc	wav2lip	faceswap -wav2lip	fsgan -wav2lip
FakeAVCeleb	Full dataset	500	2317	3964	500	9602	2717	3553
	Training set	430	614	3087	430	7212	2223	2774

This dataset is based on 500 videos sourced from YouTube of celebrities (both male and female) from different regions (Africa, South Asia, East Asia, Europe and America) speaking. That gives us the ability to create a model that will perform well under real-life conditions.

It contains 500 real videos and more than 20000 fake videos which are created using different methods: wav2lip [15], Faceswap [15], FSGAN [20], SV2TTS [11] and some of their combinations (six methods in total). It contains all types of manipulations: visual, audio and audio-visual. 430 subjects are a basis for a train set and others 70 are used for evaluation. However, it gives us more than 16000 train samples and only 430 of them are real. To address the class imbalance problem in the training split we add 3570 real videos sourced from the VoxCeleb2 dataset [5].

To evaluate our approach we use 8 test splits. 6 sets correspond to the used methods: FSGAN, Wav2Lip, FaceSwap, FaceSwap-Wav2Lip, FSGAN-Wav2Lip, RTVC. The more representative sets are Test-set-1 and Test-set-2. Test-set-1 contains an equal number of samples created using each method. Test-set-2 contains an equal number of samples in terms of types of deepfakes (FVRA, RVFA, FVFA). Every test set consists of 70 real and 70 fake samples except for Test-set-1 (66 real and fake samples, because there are 6 methods) and Test-set-2 (69 real and fake samples, because there are 3 groups).

These splits are inspired by [23]. So that we can compare our approach to the state-of-the-art solution in detail.

4.2 Preprocessing

We rely on the official implementation of AV-HuBERT[2]. Preprocessing for this model includes separating audio and video. For encoding input data AV-HuBERT uses ResNet18 for visual modality and Feed Forward Network for audio modality. Audio is extracted in mono format and a sampling rate of 16 kHz and then converted to log filterbank energies. As a video input this model needs mouth region, so we use Dlib CNN detector with a pre-trained checkpoint and then crop every frame to 96×96 format. For audio, every 4 frames are stacked using code from AV-HuBERT repository to get the same number of frames as a video (video is 25 fps and audio is extracted at 100 fps). Every sample (both visual and audio modalities) is cropped or circularly padded to a length of 75 frames (3s). Output dimensions are $T \times D$ which is 75×768 in our case.

We use HuggingFace implementation of the ViViT model with original Google checkpoint (pre-trained on the Kinetics dataset [33]). ViViT uses tubelet embedding and positional encoding to split input into spatio-temporal blocks. Then they are passed to spatio-temporal encoder.We randomly pick 32 frames and process them with an image processor. As an output, we obtain tensor shaped $T \times H \times \times W \times C$. Number of channels C is 3, T is 32, H and W both equal 224. ViViT consists of a spatial transformer and a temporal transformer. ViViT splits input in patches sized $2 \times 16 \times 16 \times 3$. We obtain features with shape $S \times D$ which is 3136×768. Then we reshape them to get $T' \times (H' \cdot W') \times D$ representation and take the average over the spatial dimension. Next, we apply linear interpolation along time to double the temporal length. The output has shape $T' \times D$ which is 32×768 in our case.

AASIST encoder takes 4 seconds of audio in mono format and a sampling rate of 16 kHz. AASIST encoder is based on RawNet2 [25], except that output of the first layer (sinc-convolution layer) is a 2-dimensional image Time $\times$ Frequency and not a 1-dimensional sequence with F channels. We use circular padding for those samples, which are shorter than needed. Output dimensions are $S \times D$ which is 29×64 in our case.

Finally, for classification head we use pool graphs to size $T = 25$ as in original paper [30].

4.3 Baselines

We use two more different approaches for comparison. One of them is a basic variant and the other is used in a number of deepfake detection models, some of them achieve state-of-the-art results, such as AV-Lip-Sync+ [23].

[2]Official AV-HuBERT implementation: https://github.com/facebookresearch/av_hubert (visited on September 16, 2025)

4.3.1 MLP

This is the simplest classifier. We provide its results, to show how much better our main idea performs. In this approach we take the mean over the temporal dimension and obtain a 1-D vector for each modality.

$$P_m = \text{mean}(F_m) \in \mathbb{R}^{D_m} \quad \forall m \in \{\text{l}, \text{v}, \text{a}\} \tag{11}$$

$$F_c = P_l \oplus P_v \oplus P_a \in \mathbb{R}^{D_l + D_v + D_a} \tag{12}$$

We apply an MLP to F_c which outputs a 2-element probability vector of video being real or fake. The prediction is made by choosing the option with greater probability.

$$\hat{y} = \arg\max(\text{MLP}(F_c)) \tag{13}$$

4.3.2 MS-TCN

As another baseline, we use MS-TCN utilized in AV-Lip-Sync+ paper. MS-TCN focuses on spectro-temporal correlation and that is the opposite of how we apply the AMSDF classifier. So comparing these classifiers by passing them the same features on the same training and evaluation data gives us ability to show if GAT-based classifier performs worse or better.

For MS-TCN classifier, which takes $D \times T$ shaped features, we do not pool F_m, but instead transpose it.

$$P_m = F_m^T \in \mathbb{R}^{D_m \times T} \quad \forall m \in \{\text{l}, \text{v}, \text{a}\} \tag{14}$$

And fused features are transposed F_l, F_v, F_a concatenated along the hidden dimension.

$$F_c = P_l \oplus P_v \oplus P_a \in \mathbb{R}^{(D_l + D_v + D_a) \times T} \tag{15}$$

Then we pass F_c to the MS-TCN adapted from [18], that is followed by a temporal pooling layer and a linear layer. In the end, we still have a 2-element probability vector and the prediction is made with the argmax function.

4.4 Training

As mentioned earlier, we do not fine-tune pre-trained models (AV-HuBERT and ViViT). Since we extract features when creating dataset and save them before the training process.

After we add 3570 videos from the VoxCeleb2 dataset to the training set, there is still a problem: we have nearly four times more fake videos than real videos. Although most other papers suggest using oversampling for the same reason, we do not do so, because of limited computational resources. To deal with the data imbalance problem, we use weighted cross-entropy loss. Weights are chosen according to following formulas:

$$\omega_{\text{real}} = \frac{R+F}{2 \cdot R}, \quad \omega_{\text{fake}} = \frac{R+F}{2 \cdot F}, \tag{16}$$

where R and F are relatively number of real samples and fake samples.

Table 4.2: Training configuration for different classification heads

	MLP	MS-TCN	AMSDF
Epochs	10	30	15
Learning rate	1×10^{-4}	1×10^{-5}	1×10^{-3}
Scheduler	StepLR ($\gamma = 0.9$)		
Optimizer	Adam		

Table 4.2 shows the parameters we used for training each baseline. We choose the learning rate that yielded the best results after several experiments, selected independently for each classification head.

For every approach we trained 4 different models with each subset of features: L+A+V, L+A, L+V, A+V. We expect the best results from the model that uses all three types of features. We do not fine-tune the pre-trained models due to limited computational resources and train AASIST encoder and the classification heads from scratch.

We name each model in the format {classifier type}-{features}, for example, AMSDF-LAV or MSTCN-AV.

5 Results

To guarantee a fair comparison we train all baselines under identical conditions: the same input features and the same training set, which allows us to measure how much better our approach performs.

In Table 5.1 we provide the detailed performance of the AMSDF-LAV model on the FakeAVCeleb dataset. We can see that it performs significantly worse on test cases where only the visual modality is affected by the deepfake method. This may be explained by the very small number of samples in the corresponding groups as we can see in Table 4.1.

Table 5.1: Results of AMSDF-LAV model on FakeAVCeleb dataset.

		Prec	Rec	F1	Acc	visual changes	audio changes
FaceSwap	Real	0.53	0.97	0.69	0.57	✓	
	Fake	0.83	0.14	0.24			
FSGAN	Real	0.59	0.96	0.72	0.65	✓	
	Fake	0.88	0.31	0.46			
Wav2Lip	Real	0.76	0.97	0.85	0.84		✓
	Fake	0.96	0.70	0.81			
RTVC	Real	1	0.96	0.98	0.98		✓
	Fake	0.96	1	0.98			
FaceSwap+Wav2Lip	Real	0.94	0.97	0.96	0.96	✓	✓
	Fake	0.97	0.94	0.96			
FSGAN+Wav2Lip	Real	0.76	0.97	0.86	0.84	✓	✓
	Fake	0.96	0.70	0.81			
Test-set-1	Real	0.76	0.96	0.85	0.82	✓	✓
	Fake	0.94	0.68	0.79			
Test-set-2	Real	0.83	0.97	0.89	0.89	✓	✓
	Fake	0.96	0.80	0.87			

The model performs extremely well on RTVC and FaceSwap+Wav2Lip test sets. It shows high results on Test-set-2, Wav2Lip, FSGAN+Wav2Lip, Test-set-1. Note that, in all these cases, the F1 score is high, which indicates that our model not only effective at finding deepfake or real videos, but also balanced. However, F1 score for real videos is greater than or equal to F1 score for fake ones.

In Table 5.2, we can see the performance of all models on every test set. We provide only the Accuracy metric to make the table easier to understand. It is clear that our model outperforms the others on almost all test sets, except for Wav2Lip and FSGAN-Wav2Lip, where it shows the second-best result, close to the best. We highlight how graph-based approach is by far the best for our balanced Test-set-1 and Test-set-2.

Among models with an MLP classifier the highest results are shown by the one that uses all three modalities for deepfake detection. Surprisingly, MSTCN-LV outperforms MSTCN-LAV, but in two other groups this situation does not occur, so we can assume it's a statistical artifact.

If we split models into groups by which input features they use, we obtain the expected picture: the AMSDF model shows the best results and the MLP the worst. An exception is once again the LV group, where MSTCN shows slightly better results than the AMSDF classifier. The reason for such results may be that MS-TCN is originally developed for visual-temporal tasks. It is created for working with time sequences so high results on the input where visual features dominate, are not surprising.

We expect to see even higher results in case of fine-tuning pre-trained AV-HuBERT and

Table 5.2: Results of all models on the FakeAVCeleb dataset. In this Table we provide only the Accuracy metric.

	faceswap	fsgan	wav2lip	rtvc	faceswap -wav2lip	fsgan -wav2lip	Test-set-1	Test-set-2
Xception (audio-only) [13]	-	-	-	-	-	-	-	0.76
VGG16 (video-only) [13]	-	-	-	-	-	-	-	0.81
LipForensics (video-only) [7]	-	-	-	-	-	-	-	0.76
MLP-LA	0.5	0.54	0.63	0.83	0.77	0.63	0.73	0.77
MLP-LV	0.53	0.54	0.73	0.88	0.86	0.73	0.75	0.79
MLP-AV	0.52	0.53	0.71	0.86	0.86	0.71	0.72	0.78
MLP-LAV	0.55	0.59	**0.88**	0.88	0.88	**0.88**	0.78	0.82
MSTCN-LA	0.55	0.55	0.76	0.95	0.95	0.76	0.77	0.82
MSTCN-LV	0.53	0.64	0.76	0.93	0.81	0.76	0.80	0.85
MSTCN-AV	0.53	0.54	0.79	0.86	0.88	0.79	0.77	0.80
MSTCN-LAV	0.54	0.54	0.77	0.88	0.88	0.77	0.78	0.81
AMSDF-LA	0.52	0.51	0.74	0.97	**0.96**	0.74	0.81	0.82
AMSDF-LV	0.56	0.61	0.72	0.90	0.81	0.72	0.79	0.82
AMSDF-AV	0.53	0.54	0.77	0.94	0.95	0.77	0.80	0.83
AMSDF-LAV	**0.57**	**0.65**	0.84	**0.98**	**0.96**	0.84	**0.82**	**0.89**

ViViT feature extractors. Nevertheless, we have achieved parity with most uni-modal approaches as shown in Table 5.2.

6 Conclusion

In this work, we compared different classification heads' performance on the same input features extracted with AV-HuBERT, ViViT and AASIST encoder in the task of audio-visual deepfake detection. AMSDF classification head based on IGAM and HGFM outperformed other approaches and achieved good results, considering that, due to limited computational resources, we did not fine-tune pre-trained models.

In general, we confirmed our hypothesis: feature fusion is a key to improving audio-visual deepfake detection and graph-based networks provide an efficient way to achieve this. In the future, we hope to improve performance even more by fine-tuning the pre-trained models and enhancing the classification head.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. *LRS3-TED: a large-scale dataset for visual speech recognition*. 2018. arXiv: 1809.00496 [cs.CV]. URL: https://arxiv.org/abs/1809.00496.

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. *ViViT: A Video Vision Transformer*. 2021. arXiv: 2103.15691 [cs.CV]. URL: https://arxiv.org/abs/2103.15691.

[3] Marcella Astrid, Enjie Ghorbel, and Djamila Aouada. *Statistics-aware Audio-visual Deepfake Detector*. 2024. arXiv: 2407.11650 [cs.CV]. URL: https://arxiv.org/abs/2407.11650.

[4] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. *Voice-Face Homogeneity Tells Deepfake*. 2022. arXiv: 2203.02195 [cs.CV]. URL: https://arxiv.org/abs/2203.02195.

[5] J. S. Chung, A. Nagrani, and A. Zisserman. "VoxCeleb2: Deep Speaker Recognition". In: *INTERSPEECH*. 2018.

[6] Yazan Abu Farha and Juergen Gall. *MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation*. 2019. arXiv: 1903.01945 [cs.CV]. URL: https://arxiv.org/abs/1903.01945.

[7] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. *Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection*. 2021. arXiv: 2012.07657 [cs.CV]. URL: https://arxiv.org/abs/2012.07657.

[8] 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)". In: *IEEE* 2.5 (2022), p. 6.

[9] Ammarah Hashmi, Sahibzada Adil Shahzad, Chia-Wen Lin, Yu Tsao, and Hsin-Min Wang. "Understanding Audiovisual Deepfake Detection: Techniques, Challenges, Human Factors and Perceptual Insights". In: *arXiv:2411.07650* (2024).

[10] 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". In: *Applied Soft Computing* 136 (2023), p. 110124. ISSN: 1568-4946. DOI: https://doi.org/10.

1016/j.asoc.2023.110124. URL: https://www.sciencedirect.com/science/article/pii/S1568494623001424.

[11] Ye Jia, Yu Zhang, Ron J. Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. *Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis*. 2019. arXiv: 1806.04558 [cs.CL]. URL: https://arxiv.org/abs/1806.04558.

[12] Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. *AASIST: Audio Anti-Spoofing using Integrated Spectro-Temporal Graph Attention Networks*. 2021. arXiv: 2110.01200 [eess.AS]. URL: https://arxiv.org/abs/2110.01200.

[13] 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*. MM '21. ACM, Oct. 2021, pp. 7–15. DOI: 10.1145/3476099.3484315. URL: http://dx.doi.org/10.1145/3476099.3484315.

[14] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. *FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset*. 2022. arXiv: 2108.05080 [cs.CV]. URL: https://arxiv.org/abs/2108.05080.

[15] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. *Fast Face-swap Using Convolutional Neural Networks*. 2017. arXiv: 1611.09577 [cs.CV]. URL: https://arxiv.org/abs/1611.09577.

[16] Menglu Li, Yasaman Ahmadiadli, and Xiao-Ping Zhang. *Audio Anti-Spoofing Detection: A Survey*. 2024. arXiv: 2404.13914 [cs.SD]. URL: https://arxiv.org/abs/2404.13914.

[17] Ping Liu, Qiqi Tao, and Joey Tianyi Zhou. *Evolving from Single-modal to Multi-modal Facial Deepfake Detection: A Survey*. 2024. arXiv: 2406.06965 [cs.CV]. URL: https://arxiv.org/abs/2406.06965.

[18] Brais Martinez, Pingchuan Ma, Stavros Petridis, and Maja Pantic. "Lipreading Using Temporal Convolutional Networks". In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 6319–6323. DOI: 10.1109/ICASSP40776.2020.9053841.

[19] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. *Exploring Self-Supervised Vision Transformers for Deepfake Detection: A Comparative Analysis*. 2024. arXiv: 2405.00355 [cs.CV]. URL: https://arxiv.org/abs/2405.00355.

[20] Yuval Nirkin, Yosi Keller, and Tal Hassner. "FSGAN: Subject agnostic face swapping and reenactment". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2019, pp. 7184–7193.

[21] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. *AVFF: Audio-Visual Feature Fusion for Video Deepfake Detection*. 2024. arXiv: 2406.02951 [cs.CV]. URL: https://arxiv.org/abs/2406.02951.

[22] Muhammad Anas Raza and Khalid Mahmood Malik. "Multimodaltrace: Deepfake Detection Using Audiovisual Representation Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. June 2023, pp. 993–1000.

[23] Sahibzada Adil Shahzad, Ammarah Hashmi, Yan-Tsung Peng, Yu Tsao, and Hsin-Min Wang. *AV-Lip-Sync+: Leveraging AV-HuBERT to Exploit Multimodal Inconsistency for Video Deepfake Detection*. 2023. arXiv: 2311.02733 [cs.CV]. URL: https://arxiv.org/abs/2311.02733.

[24] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. "Learning Audio-Visual Speech Representation by Masked Multimodal Cluster Prediction". In: *arXiv preprint arXiv:2201.02184* (2022).

[25] Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. *End-to-end anti-spoofing with RawNet2*. 2021. arXiv: 2011.01108 [eess.AS]. URL: https://arxiv.org/abs/2011.01108.

[26] Mulin Tian, Mahyar Khayatkhoei, Joe Mathai, and Wael AbdAlmageed. *Unsupervised Multimodal Deepfake Detection Using Intra- and Cross-Modal Inconsistencies*. 2024. arXiv: 2311.17088 [cs.CV]. URL: https://arxiv.org/abs/2311.17088.

[27] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. *MLP-Mixer: An all-MLP Architecture for Vision*. 2021. arXiv: 2105.01601 [cs.CV]. URL: https://arxiv.org/abs/2105.01601.

[28] Cristian Vaccari and Andrew Chadwick. "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News". In: *Social Media + Society* 6.1 (2020), p. 2056305120903408. DOI: 10.1177/2056305120903408. URL: https://doi.org/10.1177/2056305120903408.

[29] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. *Graph Attention Networks*. 2018. arXiv: 1710.10903 [stat.ML]. URL: https://arxiv.org/abs/1710.10903.

[30] Junyan Wu, Qilin Yin, Ziqi Sheng, Wei Lu, Jiwu Huang, and Bin Li. "Audio Multi-View Spoofing Detection Framework Based on Audio-Text-Emotion Correlations". In: *IEEE Transactions on Information Forensics and Security* 19 (2024), pp. 7133–7146. DOI: 10.1109/TIFS.2024.3431888.

[31] Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. "AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake". In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 2015–2029. DOI: 10.1109/TIFS.2023.3262148.

[32] Yipin Zhou and Ser-Nam Lim. "Joint Audio-Visual Deepfake Detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, pp. 14800–14809.

[33] Andrew Zisserman, Joao Carreira, Karen Simonyan, Will Kay, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, and Mustafa Suleyman. "The Kinetics Human Action Video Dataset". In: 2017.