DualStreamNet : Robust Audio-Video Deep Fake Media Detection using Complimentary Information Fusion

Shreyas Sheeranali[†] Raghavendra Ramachandra[†] Sushma Venkatesh[‡]
[†]*Norwegian University of Science and Technology (NTNU), Gjøvik, Norway.*
[‡]*MOBAI AS, Gjøvik, Norway.*
E-mail: {raghavendra.ramachandra}@ntnu.no

***Abstract*—Deepfake technology employs sophisticated machine-learning techniques to create highly convincing video and audio recordings of individuals doing or saying things that they never actually did or said. These falsified media pieces have the potential to deceive and manipulate viewers, posing significant risks to their privacy, security, and trust in digital media. In this paper, we present a novel method *DualStreamNet* for reliable Audio-Video (AV) fake media detection which exploits the complementary information. The proposed *DualStreamNet* includes independent detector for video and audio modality. We introduced a novel video fake detection framework using a SlowFast encoder as the backbone, and a novel architecture based on a 3D CNN with skip connections. We also introduced novel features to reliably detect audio fakes using a Continuous Wavelet Transform (CWT) Filter Bank that was further processed using the ResNet50 architecture. Finally, the decisions from the video and audio detectors are combined using the logical OR rule to make the final decision. Extensive experiments were performed on two publicly available audio-video fake datasets: FakeAVCeleb and SWAN-DF. The obtained results indicate the improved detection accuracy of the proposed method compared to existing methods.**

I. Introduction

The extensive use of social media platforms has led to a global proliferation of information across millions of users. However, verifying the authenticity of information on these platforms is challenging due to the absence of moderation. Additionally, the growth of generative AI has further exacerbated this situation by equipping new tools and techniques to produce fake multimedia content. Presently, generative AI tools are the driving force behind the dissemination of disinformation content, which intentionally spreads false information to mislead and harm individuals on social networks. The repercussions of multimedia content manipulation, also referred to as deep fakes, became apparent in 2017, when Reddit users started posting digitally altered pornographic videos on their webpages [7]. Since then, techniques for manipulating multimedia content have advanced significantly, enabling the creation of high-quality visual content that can easily deceive consumers [10], [29].

The advancement of generative AI has led to significant progress in the creation and detection of multimedia fakes, which refer to video and audio content in multimedia files that have been deliberately altered to deceive viewers or listeners. These fakes are also referred to as FakeAVs in the literature. Because multimodal fakes consist of both video and audio, there are three types of fake content: (a) Real Video Fake Audio (RVFA), in which only the audio content is modified by keeping the video content the same. Audio content can be modified by altering the tone [43], semantically altering the audio content [11]. (b) Fake Video Real Audio (FVRA), in which only video content is altered by keeping the audio content constant. Conventional video alternation techniques such as Morphing [41], face swapping [24], [31], Facial attribute alternations and morphing [38], lip sync [14], [34] and facial reenactment [31], [40] (c) Fake Video Fake Audio (FVFA): Here, both video and audio is manipulated using existing video and audio techniques independently with additional care for synchronization. However, emerging generative AI techniques such as diffusion models are capable of generating audio and video simultaneously [27], [39] with high

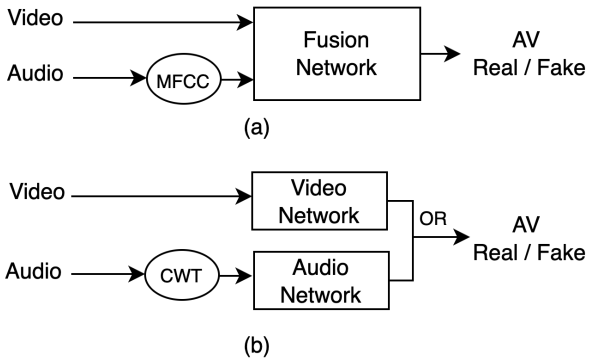

Fig. 1: In most existing works, the structure described in (a) is followed, employing the Mel Frequency Cepstrum Coefficient (MFCC) for audio feature representation and a fusion network that provides a combined score. This approach necessitates the presence of both modalities and their comparable relationships in both real and fake media. On the other hand, the proposed method in (b) utilizes Continuous Wavelet Transform (CWT) filter banks instead of MFCC and trains individual audio and video classifiers, whose decisions are fused using logical OR function to arrive at the final decision.

perpetual quality. The improvement in the generation quality of multimodal fakes makes automatic detection challenging, which has motivated researchers to pursue the development of detection techniques.

Multimodal fake detection has gained momentum in the past few years, resulting in several detection techniques [21]. The available multimodal face detection methods are designed to detect the inconsistency between audio and video using generic deep CNN models and emotion detection models. In [25], the inconsistency in the spatial, temporal, and spectral regions of video and audio was modeled using different models, such as Xception, LipNet, and DeepSpeech2, whose features were combined using a multilayer perceptron (MLP) model to obtain the final prediction. In [6], the homogeneity between audio and video was used to detect multimodal fake data. A custom transformer-like module with InfoNCE loss was proposed to detect deep fakeAV attacks. In [44], the joint learning of audio-video data was introduced using both spatial and temporal information together with cross-attention to learn the relationships between both modalities. In [42] FTFDNet was introduced, which employs an audio-visual attention mechanism (AVAM) framework that enables the identification of crucial features for a task. The audio and video signals were processed separately using a stack of residual convolutional layers to encode the input into low-dimensional features. These features are then combined to obtain a joint representation that can be used to detect FakeAV samples. In [17], a two-stream network was introduced using a SWIN transformer to independently model video and audio. The audio samples were processed to obtain a Mel Spectrogram, and the video was processed to have only the face region that was then independently trained using SWIN transformers. The max rule is applied to combine decisions from audio and video to make the final decision. In [13], a self-supervised approach was proposed using an autoregressive model to capture the temporal synchronization between the frames and the sound of the sample. ResNet-18 [15] is used as the backbone to extract the features (2D and 3D) from video and audio that are combined through a transformer model. The final decision was made using the autoregressive model. In [8], an anomaly detection technique was proposed using audiovisual identity verification, called POIForensics. A contrastive learning paradigm is used to model the identity information of individual features of real videos and then search for inconsistencies in the embedding space.

The emotions from the audio and video signals were also used to detect FakeAV samples. In [28], a deep learning approach based on Xception and LSTMs was introduced to quantify emotional inconsistencies between video and audio. The audio signals were processed using SincNet, and the final decision was obtained by combining the information from the video and audio using an MLP. In [30], a Siamese network using triplet loss was proposed for audio video-based fake detection based on emotion modelling. In [16], a video forensic system using sentiment recognition from frames and audio signals was used to identify inconsistent and non-natural emotions. The is employed valence-arousal model was employed to evaluate the emotional responses evoked by LLD audio and faces, specifically focusing on two dimensions of emotion: positivity and arousal. This information on inconsistency in emotion between video and audio is used to make the final decision.

It appears that the inconsistency modeling of multimodal signals, such as video and audio, is more robust and generalizable than emotion modeling. Although existing emotion-based approaches have achieved reasonable detection results, they are unable to fully replicate human emotions because of their inherently subjective and complex nature. Among the inconsistency modeling techniques, backbone models and transformers are commonly used for video data, whereas audio data spectrograms and MFCC features are commonly used for detecting multimodal fake data. Although inconsistency modeling approaches have shown promising results in detecting FakeAV attacks, they still lack generalizability across different types of attack generations and variations in data quality.

The aim of this work is to develop a novel method for detecting fraudulent AV media by analyzing inconsistencies independently of both video and audio data. To achieve this, we introduced a new 3D CNN architecture that utilizes residual connections for processing video signals, and we proposed novel features based on continuous Wavelet Transform (CWT) followed by ResNet50 for modeling audio signals. Lastly, we combined the predictions from the video and audio signals using the logical OR rule to make the final decision. The primary advantage of the proposed approach lies in its ability to handle scenarios in which only one form of media (either audio or video) is accessible. The key contributions of this work are summarized below:

- Proposed a novel technique for audio-video fake media detection based on inconsistency modeling independently on video and audio and fusing the independent decisions from video and audio using logical OR rule.
- Proposed a novel 3D CNN architecture with skip connections to model the video inconsistency. The input to the proposed architecture was built over the Slow-Fast [12] network as the backbone.
- Proposed a novel Audio inconsistency modeling based on the CWT filter banks followed by the fine-tuned ResNet50.
- Extensive experiments are performed on the publicly available FakeAV datasets such as FakeAVCeleb [22] and SWAN-DF [23] and comparison with State-Of-The-Art (SOTA) video-audio fake detection techniques. The performance evaluation protocols are designed to evaluate the generalisation of the proposed method.
- The source code of the proposed method along with the evaluation protocol for SWAN-DF dataset is made public upon acceptance of the paper.

The paper is organised as follows: Section II presents the proposed audio-video based fake detection technique describing the individual modality networks, the intuition behind the

design, fusion of the scores and the implementation details. Section III describes the datasets used for evaluation, the performance evaluation protocol and performance comparison of the proposed and existing methods and Section IV draws the conclusion.

II. Proposed *DualStreamNet* for Audio-Video Fake Media Detection

Figure 2 illustrates the proposed architecture of the proposed *DualStreamNet* audio-video fake detection system. Existing studies, such as those involving joint audio-video representation learning [1], [6], are not capable of addressing scenarios where only a single modality is available. In contrast, methods based on identity verification [8] rely on the presence of an authentic video featuring the same individual. To address these limitations, we propose *DualStreamNet*: a two-stream network trained independently on audio and video data. The final predictions made by each modality were then combined using logical OR rule to determine whether the audio-video pair was genuine or fraudulent. In the following we discuss the proposed video and audio networks together with the fusion module.

A. Proposed Video Network

The proposed video detector architecture comprises of two main components: the Backbone and Head. In this study, we selected the pretrained SlowFast [12] network as the backbone. The SlowFast [12] network was trained for the action recognition task using the Kinetics 400 [20] dataset. We selected the SlowFast [12] network because of its exceptional performance in action recognition and the diverse nature of the data used to train the network. We believe that the incorporation of the SlowFast [12] network will enable the extraction of reliable information on motion and semantics, which can be utilized to quantify fake video signals. The SlowFast network comprises two pathways [12]: Fast: Captures the rapidly changing motion by working at a fast refreshing speed; it has fewer channels to intentionally reduce its capability to process spatial information. It has a large temporal stride τ on the input frames; that is, it processes only one out of τ frames. A typical value of τ is 16, that is, the refreshing speed is approximately two frames sampled per second for 30-fps videos. Slow: Capture semantic information that can be provided by images or a few sparse frames with more channels to retrieve spatial information. It operates with a small temporal stride of τ/α, where $\alpha > 1$ is the frame rate ratio between the Fast and Slow pathways. α was chosen as eight in their experiments.

The backbone of our model is initialised with weights obtained from training on the Kinetics-400 [20] database, and we extract features from the penultimate block, which has 4096 channels with height, width, and time dimensions of 8. These extracted features were then passed to the proposed head using the 3D CNN layers.

In this work, we propose a novel 3D CNN-based serial architecture with skip connections inspired by U-Net [36] to effectively harness temporal, contextual, and spatial information for the detection of fake videos. The proposed video detector architecture comprises a contracting path with an encoder layer and an expansive path with a decoder layer, with skip connections that allow for the flow of contextual information. The encoder consisted of four serially connected convolutional layers, followed by ReLU and max pooling. The input convolution layer has a kernel size of $3 \times 3 \times 3$ and employs 1024 filters, while the subsequent convolutional layers maintain a constant kernel size of $3 \times 3 \times 3$ and gradually decrease the number of filters, thereby capturing both high- and low-level features as the features progress through the path. The decoder network employs transposed convolution to upscale the feature space by combining the information from the contracting path via skip connections. Subsequently, max pooling is followed by a gated linear unit [9], which is a lightweight layer instrumental in the selection of features. Finally, sigmoid activation was used to obtain a probability score for classifying the video as fake (1) or real (0).

B. Proposed Audio Model

Previous research on audio-based deepfake detection primarily relied on spectrograms or mel-frequency cepstral coefficients (MFCCs) as feature extractors. In this work, we introduce novel audio features using a Continuous Wavelet Transform (CWT) filter bank that preserves greater details of the signal and does not discard higher frequency components that could be useful for distinguishing between genuine and fake audio. Additionally, the CWT Filter bank adapts to the signal characteristics, which can enhance the robustness and generalization of the feature extraction. We computed the time-frequency features of the audio signal using the CWT and generated a scalogram by computing the absolute value of the CWT coefficients corresponding to the audio signal. To create the scalogram, we used a CWT filter bank based on the Morse wavelet, resulting in a 256×256-dimensional feature map. This feature map was then passed through the ResNet50 architecture [15], followed by a GLU and a Sigmoid layer, to compute the probability score for classifying the audio as fake (1) or real (0). We trained the audio network end-to-end by initializing the weights using the values computed on the ImageNet dataset [37]. Figure 3 shows the example of CWT filterbank features extracted on bona fide and fake samples.

C. Fusion of Video and Audio Predictions

In the next step, we integrate the outputs from the video and audio models to arrive at the final decision. For this task, we utilized the logical OR rule to combine predictions from the two modalities. We used a standard 0.5 threshold because the sigmoid function in the last layer of the classifier outputs a probability, and 0.5 is the midpoint, representing an equal likelihood of being fake or real. We used the logical OR rule because it enables computation of the final decision, even when one of the modalities (either video or audio) is unavailable.

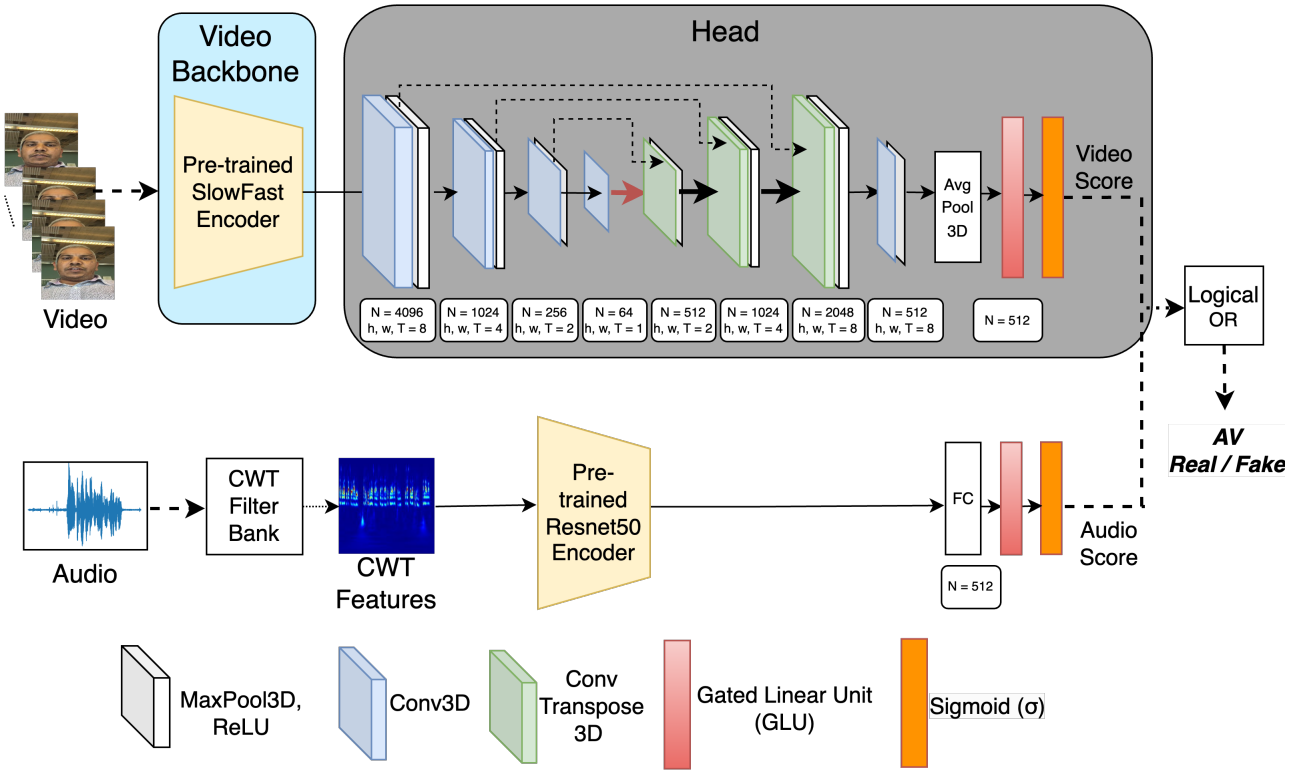

Fig. 2: Block Diagram of the Proposed Audio-Video fake detection network, **DualStreamNet**: Videos are trained using the SlowFast Network based module and the audio files are pre-processed with CWT Filter Bank before feeding into Resnet50-backbone. Both the streams are trained individually with BCE Loss, and the logical OR of the two predictions are used for making the final decision.

D. Implementation Details

For our experiments, we utilized an NVIDIA A100 GPU with a 20 GB VRAM partition. The proposed model was optimized using the binary cross-entropy loss for audio and video modalities. AdamW was selected as the optimizer, with a learning rate of 1e-6, weight decay of 1e-7, and a Cosine Annealing Scheduler. Training was conducted for 30 epochs for FakeAV training and fine-tuning on the SWAN dataset.

III. Experiments and Results

In this section, we present the quantitative results of the proposed multimodal fake detection on two datasets: FakeAVCeleb [22] and SWAN-DF [23]. The quantitative performance of the proposed method was compared with ten different state-of-the-art techniques on the FakeAVCeleb dataset and one state-of-the-art technique on the SWAN-DF. The results are presented in terms of accuracy, with a higher value indicating superior detection performance.

A. Multimodal Audio-Video Fake Databases

In this section, we briefly discuss the details of the two different multimodal audio video datasets employed in this work to benchmark the performance of the proposed AV fake detection technique.

Dataset	Split	Audio		Video	
		Real	Fake	Real	Fake
FakeAV	Train	8177	9062	800	16439
	Test	2032	2273	200	4105
SWAN (Subset)	Train	1216	1152	640	576
	Test	320	320	160	160
SWAN: Session-1	Train	640	7216	640	4608
	Test	160	1856	160	1152
SWAN: Session-2 to 6	Train	320	-	320	-
	Test	80	-	80	-

TABLE I: Dataset-wise train-test and real-fake split for FakeAVCeleb and SWAN. Sessions 2-6 do not have fake samples since SWAN-DF [23] contains fake media created only from session 1 real video and audio.

1) FakeAVCeleb: FakeAVCeleb [22] comprises videos of celebrities with different ethnic backgrounds–Caucasian, Black, South Asian, and East Asian–belonging to diverse age groups with equal proportions for each gender. Fake videos were generated using Faceswap [24] and Faceswap GAN (FSGAN) [31] with Wav2Lip [6]. Face++ [1] was used to measure similarity of faces to select the closest ones for fusion,

[1]https://www.faceplusplus.com/face-comparing/

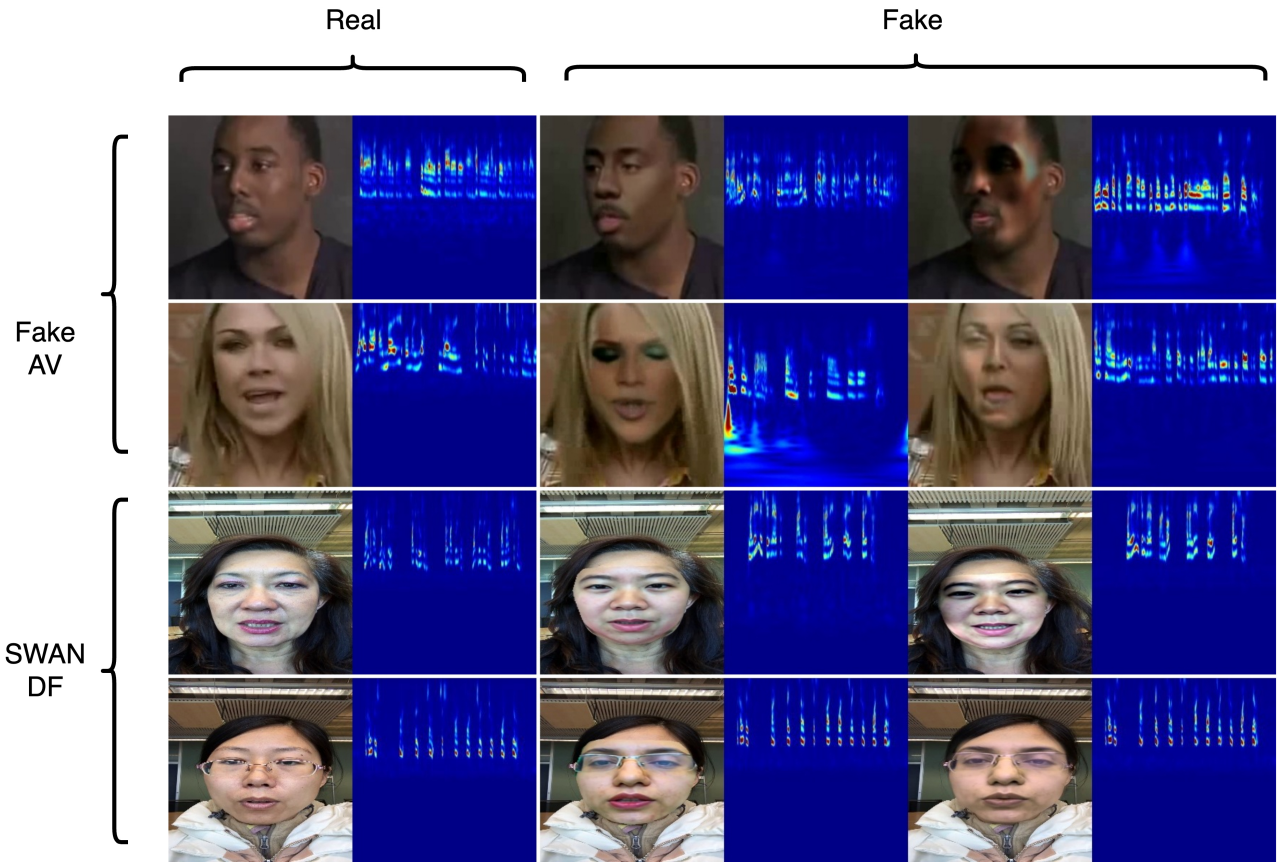

Fig. 3: Representative images from the datasets - a sample from the video followed by the corresponding CWT Filter Bank image

individually for each ethnicity. Meanwhile, fake audio was synthesized using a Real-Time Voice Cloning tool (RTVC), such as SV2TTS [18]. In this study, we followed the evaluation protocol defined in [6] to evaluate the performance of the proposed and existing methods. Table I lists the database splits used in this work to benchmark quantitative results.

2) SWAN-DF: SWAN-DF [23] is the first photorealistic quality database for audio-visual deepfakes created from the SWAN [35] biometric database. Video deepfakes are created using the DeepFaceLab [2] [32] open-source repository using the 256x256 resolution variants whose images generated using LIAE architecture. More than 20 types of blending techniques were deployed to ensure that deepfake was not detected based on the residues left during blending. Other methods used have not been open sourced for ethical reasons. Audio deepfakes were created using four voice conversion methods, YourTTS [4], HiFiVC [19], DiffVC [33], and FreeVC [26] and two text-to-speech methods, Adaspeech [5] and TorToiSe TTS [2].

SWAN-DF's media are created entirely from session-1 of the IDIAP subset of the SWAN database, which has six sessions in total. Since this is a newly released dataset in which there is no evaluation protocol for benchmarking the detection accuracy. Therefore, we propose an evaluation protocol to have an 80:20 train and test, where the train and test sets have a disjoint set of identities. Since the SWAN dataset has six sessions, we used bona fide (or real) samples from all sessions to compute the detection performance. However, fake samples were generated using session 1 data. Table I lists the statistics of the SWAN-DF dataset used to benchmark the performance of the proposed and existing face detection methods.

[2]https://github.com/iperov/DeepFaceLab

B. Results and Discussion

In this section, we present the quantitative results of the proposed AV fake detection, DualStreamNet and existing methods on two different datasets: FakeAVCeleb [22] and SWAN-DF [23]. The quantitative performance of the proposed and existing methods is presented using the detection accuracy; thus, a higher value of accuracy indicates better detection performance. Table II shows the performance of DualStreamNet and existing methods on FakeAVCeleb [22] which were evaluated using the same experimental protocols as those defined in [6]. The performance of the proposed method was compared with of ten different state-of-the-art AV

AV Detection Algorithm	Detection Accuracy (%)		
	Aud & Vid	Video	Audio
XceptionNet [21]	43.94	73.06	76.26
Meso-4 [21]	45.93	43.15	50.36
EfficientNet-B0 [21]	63.18	59.64	50
MesoInception-4 [21]	72.87	77.88	53.96
VGG-16 [21]	78.04	81.03	67.14
VFD [6]	81.52	–	–
POI-Forensics [8]	86.6	–	–
DST-Net [17]	92.59	90.94	98.73
BA-TFD [3]	79.81	–	–
Multimodaltrace [1]	92.9	–	–
DualStreamNet (Proposed)	**98.7**	**98.58**	**99.88**

TABLE II: Comparison of the proposed method detection accuracy (%) with previous works on the FakeAVCeleb Dataset. Modality-wise results were not reported in [1], [3], [6], [8] as they have not been reported in the literature.

AV Detection Algorithm	Train	Test	Detection Accuracy (%)	
			English	French
BA-TFD [3]	S1 Eng	S1	93.01	92.49
		S2	93.26	92.66
		S3	93.08	92.46
		S4	93.1	92.53
		S5	93.1	92.52
		S6	93.27	92.69
DualStreamNet (Proposed)	S1 Eng	S1	98.77	99.18
		S2	99.86	100
		S3	98.6	99.85
		S4	99.86	100
		S5	99.72	100
		S6	97.91	98.32

TABLE III: **Experiment-1**: Detection Accuracy(%) of the proposed and existing method on the SWAN database when trained with **English** media.

techniques. Based on the obtained results, the following can be observed.

- The proposed method shows the best performance compared with the existing techniques with the detection accuracy of 98.70% with Audio-video data. Furthermore, the proposed method indicated the best detection accuracy of 98.58% for video and 99.88% for audio.
- The outstanding performance of the proposed method can be attributed the use of slow-fast encoder backend and the novel CNN architecture for head in the video stream. Furthermore, we introduced a CWT filter bank for audio signals that can capture details related to fake audio. Furthermore, the combination of both at the decision level further improves detection accuracy.

Tables III, IV and V show the quantitative results of the proposed and existing methods for the SWAN-DF dataset. Because there is no detection evaluation of SWAN-DF, in this work, we consider BA-TFD [3] as the state-of-the-art

AV Detection Algorithm	Train	Test	Detection Accuracy (%)	
			English	French
BA-TFD [3]	S1 French	S1	89.65	88.11
		S2	89.75	88.17
		S3	89.46	87.88
		S4	89.48	87.95
		S5	89.55	87.92
		S6	89.59	88.01
DualStreamNet (Proposed)	S1 French	S1	96	95.86
		S2	95.99	95.84
		S3	95.97	95.84
		S4	95.99	95.84
		S5	95.99	95.84
		S6	95.99	95.84

TABLE IV: **Experiment-2**: Detection Accuracy(%) of the proposed and existing methods on the SWAN-DF database when trained with **French** media.

AV Detection Algorithm	Train	Test	Detection Accuracy (%)	
			English	French
BA-TFD	S1 Eng + French	S1	82.85	82.67
		S2	83.11	82.96
		S3	83.05	83.12
		S4	83.08	83.24
		S5	83.09	83.15
		S6	83.15	83.02
DualStreamNet (Proposed)	S1 Eng + French	S1	97.14	96.45
		S2	97.13	96.43
		S3	97.13	96.43
		S4	97.13	96.43
		S5	97.13	96.43
		S6	97.13	96.43

TABLE V: **Experiment-3**: Detection Accuracy (%) of the proposed and existing methods on the SWAN database **fullset** where trained using both **English and French** media.

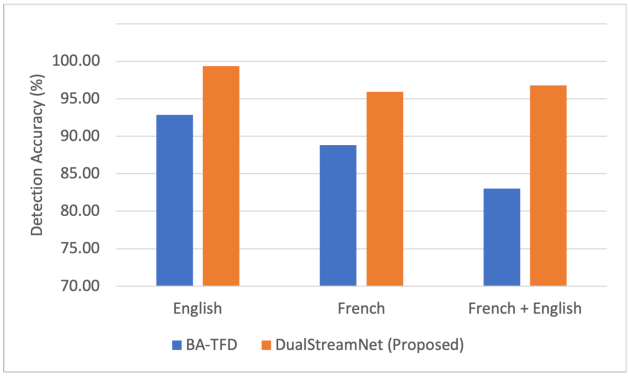

Fig. 4: Comparison of BA-TFD [3] with the average accuracy(%) across all sessions for all three variants of training - only English media, only French; French and English.

method to compare the detection accuracy with the proposed method. In this work, we present three different experiments that are targeted to evaluate the influence of language on fake audio detection, thereby influencing the overall performance of the proposed fake AV detection. **Experiment-1:** In this experiment, English language audio was used to train the audio detector, and was tested using both English and French language audio. **Experiment-2:** In this experiment, French language audio was used to train the audio detector and was tested with both English and French language audio. **Experiment 3:** Both English and French language were used for the training and testing. Based on the obtained results, the following conclusions can be drawn:

- The proposed method indicates the improved detection accuracy compared to the existing method BA-TFD [3] on all three experiments.
- Among three experiments, the training the model with English and testing with English yields the best performance. Training on the detection models in both languages will result in degraded performance in detecting AV fake samples when the French language is used. Therefore, the use of language has an impact on training data.
- When training and testing is performed on the same language, both the proposed and SOTA methods indicates the best results compared to the cross language.
- Use of bona fide data from different sessions that has reflected the different data collection environment has indicated the less effect on the detection accuracy of both the the proposed and SOTA methods. The constant detection accuracy across different sessions are due to the perfect detection of bona fide (or real) videos. It is worth nothing that, only bona fide samples are changed across different sessions while fake samples remains constant.

Thus, based on the extensive experiments reported on two publicly available deep fake AV datasets, the proposed method indicated the best performance in detecting fake AV samples compared to different SOTA. This validates the efficacy of the proposed method in detecting fake AV samples across datasets and languages.

C. Ablation Study

In this section, we present an ablation study on the different components of the proposed method, including the individual performance for video and audio fake detection. To this end, we used the FakeAVCeleb dataset.

Backbone		Head		Detection Accuracy(%)
Frozen	End-to-end	Skip Conn.	GLU	
x	✓	✓	x	82.59
✓	x	✓	✓	74.09
x	✓	x	✓	52.83
x	✓	✓	✓	**98.58**

TABLE VI: Video architecture ablations on FakeAV database

Table VI indicates the detection accuracy corresponding to the ablation study on the proposed video based fake detection model. We performed four different ablation studies considering the Backbone and Head. Within the backbone, we checked whether the use of end-to-end training would benefit fine-tuning with initial weights. For the head, we checked the importance of the skip connection and GLU block. Based on the ablation study results, we note that: (1) the skip connection in the proposed head plays an important role in improving the detection accuracy; (2) GLU plays an important role in improving the detection accuracy, as without GLU, the detection accuracy drops to 82.59%. (3) The use of end-to-end training of the backbone improves the detection accuracy compared with fine-tuning using frozen weights. (4) Thus, the use of end-to-end training with the proposed head yielded the best detection accuracy of 98.58%.

Backbone		Head	Detection Accuracy(%)
Frozen	End-to-end	GLU	
x	✓	x	99.87
✓	x	✓	94.96
x	✓	✓	**99.88**

TABLE VII: Audio architecture ablations on FakeAV database

Table VII shows the ablation study on using the backbone (based on ResNet50) and GLU in the fake audio framework. Based on the ablation study, we noted that end-to-end training had a significant impact on achieving an outstanding detection accuracy of 99.88% in detecting fake audio.

IV. Conclusion

The use of deep fake media poses a significant challenge to contemporary society, and several approaches have been explored to distinguish between them. In this work, we present DualStreamNet, a novel framework to detect audio-video-based fake media that process audio and video signals independently and combine individual decisions using the logical OR rule. We proposed a novel architecture for video-based fake detection using serial 3D CNN with skip connection and GLU, and introduced novel features using CWT filter banks with ResNet50 architecture to detect fake audio. Extensive experiments were conducted using two publicly available audio video fake datasets, FakeAVCeleb and SWAN-DF. Results obtained on FakeAVCeleb indicate the best attack detection performance, with a detection accuracy of 99.88%. With SWAN-DF, the proposed method consistently outperformed the temporal forgery- and boundary-matching-based BA-TFD networks by more than 5% in terms of the accuracy for English, French, and combined training. In future work, the concept of Low-Rank Adaptation of Large Language Models (LoRAs) can be deployed to easily fine-tune a classifier when new fake media generation algorithms are developed.

References

[1] M. Anas Raza and K. Mahmood Malik. Multimodaltrace: Deepfake detection using audiovisual representation learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 993–1000, 2023.

[2] J. Betker. TorToiSe text-to-speech, Apr. 2022.

[3] Z. Cai, K. Stefanov, A. Dhall, and M. Hayat. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization, 2023.

[4] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR, 2022.

[5] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, S. Zhao, and T.-Y. Liu. Adaspeech: Adaptive text to speech for custom voice. *arXiv preprint arXiv:2103.00993*, 2021.

[6] H. Cheng, Y. Guo, T. Wang, Q. Li, X. Chang, and L. Nie. Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–22, 2023.

[7] S. Cole. AI-Assisted Fake Porn Is Here — vice.com. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn, 2017. [Accessed 31-01-2024].

[8] D. Cozzolino, A. Pianese, M. Nießner, and L. Verdoliva. Audio-visual person-of-interest deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 943–952, 2023.

[9] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 933–941. JMLR.org, 2017.

[10] A. De Ruiter. The distinct wrong of deepfakes. *Philosophy & Technology*, 34(4):1311–1332, 2021.

[11] A. Dixit, N. Kaur, and S. Kingra. Review of audio deepfake detection techniques: Issues and prospects. *Expert Systems*, 40(8):e13322.

[12] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[13] C. Feng, Z. Chen, and A. Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023.

[14] J. Guan, Z. Zhang, H. Zhou, T. Hu, K. Wang, D. He, H. Feng, J. Liu, E. Ding, Z. Liu, et al. Stylesync: High-fidelity generalized and personalized lip sync in style-based generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1515, 2023.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] B. Hosler, D. Salvi, A. Murray, F. Antonacci, P. Bestagini, S. Tubaro, and M. C. Stamm. Do deepfakes feel emotions? a semantic approach to detecting deepfakes via emotional inconsistencies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1013–1022, 2021.

[17] 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*, 136:110124, 2023.

[18] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[19] A. Kashkin, I. Karpukhin, and S. Shishkin. Hifi-vc: High quality asr-based voice conversion. *arXiv preprint arXiv:2203.16937*, 2022.

[20] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[21] H. Khalid, M. Kim, S. Tariq, and S. S. Woo. Evaluation of an audio-video multimodal deepfake dataset using unimodal and multimodal detectors. In *Proceedings of the 1st Workshop on Synthetic Multimedia - Audiovisual Deepfake Generation and Detection*, MM '21. ACM, Oct. 2021.

[22] H. Khalid, S. Tariq, M. Kim, and S. Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

[23] P. Korshunov, H. Chen, P. N. Garner, and S. Marcel. Vulnerability of automatic identity recognition to audio-visual deepfakes. In *IEEE International Joint Conference on Biometrics*, Sept. 2023.

[24] I. Korshunova, W. Shi, J. Dambre, and L. Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3677–3685, 2017.

[25] J. K. Lewis, I. Toubal, H. Chen, V. Sandesera, M. Lomnitz, Z. Hampel-Arias, C. Prasad, and K. Palaniappan. Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9, Los Alamitos, CA, USA, oct 2020. IEEE Computer Society.

[26] J. Li, W. Tu, and L. Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[27] Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[28] M. Lomnitz, Z. Hampel-Arias, V. Sandesara, and S. Hu. Multimodal approach for deepfake detection. In *2020 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–9. IEEE, 2020.

[29] Y. Mirsky and W. Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.

[30] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832, 2020.

[31] Y. Nirkin, Y. Keller, and T. Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7184–7193, 2019.

[32] I. Perov, D. Gao, N. Chervoniy, K. Liu, S. Marangonda, C. Umé, M. Dpfks, C. S. Facenheim, L. RP, J. Jiang, et al. Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*, 2020.

[33] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, M. Kudinov, and J. Wei. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. *arXiv preprint arXiv:2109.13821*, 2021.

[34] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 484–492, New York, NY, USA, 2020. Association for Computing Machinery.

[35] R. Ramachandra, M. Stokkenes, A. Mohammadi, S. Venkatesh, K. Raja, P. Wasnik, E. Poiret, S. Marcel, and C. Busch. Smartphone multi-modal biometric authentication: Database and evaluation. *arXiv preprint arXiv:1912.02487*, 2019.

[36] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[37] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[38] J. M. Singh and R. Ramachandra. Deep composite face image attacks: Generation, vulnerability and detection. *IEEE Access*, 11:76468–76485, 2023.

[39] M. Stypułkowski, K. Vougioukas, S. He, M. Zieba, S. Petridis, and M. Pantic. Diffused heads: Diffusion models beat gans on talking-face generation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5091–5100, 2024.

[40] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[41] S. Venkatesh, R. Raghavendra, K. Raja, and C. Busch. Face morphing attack generation and detection: A comprehensive survey. *IEEE Transactions on Technology and Society*, 2(3):128–145, March 2021.

[42] G. Wang, P. Zhang, L. Xie, W. Huang, Y. Zha, and Y. Zhang. An audio-visual attention based multimodal network for fake talking face videos detection. *arXiv preprint arXiv:2203.05178*, 2022.

[43] T.-Y. Wang, I. Kawaguchi, H. Kuzuoka, and M. Otsuki. Effect of manipulated amplitude and frequency of human voice on dominance and persuasiveness in audio conferences. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), nov 2018.

[44] W. Yang, X. Zhou, Z. Chen, B. Guo, Z. Ba, Z. Xia, X. Cao, and K. Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023.