Research Proposal

Title:

Multilingual and Explainable Audio-Visual Deepfake Detection Using GAN-Aware Cross-Modal Consistency Networks

1. Abstract

Deepfake technologies, driven by generative adversarial networks (GANs), pose escalating threats to information integrity across media platforms. Although significant progress has been made in deepfake detection using multimodal techniques, current models lack generalization across languages, fail to offer explainability, and often overlook full-length temporal inconsistencies. This proposal aims to design a multilingual, explainable, and GAN-aware deepfake detection model that leverages a dual-stream architecture integrating audio and visual modalities. The model will highlight modality-specific inconsistencies using attention-based saliency mechanisms and test generalization across English, French, Chinese, and Urdu. By addressing key gaps in current state-of-the-art systems, the research aims to significantly improve the robustness, transparency, and cross-cultural applicability of deepfake detection technologies.

2. Introduction / Background

Deepfakes are synthetic media where a person in an existing image or video is replaced with someone else's likeness using deep learning techniques. GANs have accelerated the realism and spread of deepfakes, making detection increasingly difficult. Most existing detectors focus solely on facial manipulations, ignoring cross-modal inconsistencies between voice and facial movements. Additionally, nearly all research and benchmark datasets are English-centric, limiting model generalization.

As multilingual and global media become more prevalent, the need for a robust, explainable, and culturally adaptable detection mechanism is critical. There is also a growing demand for models that offer explainability and forensic traceability to support legal, journalistic, and security-related verification.

3. Problem Statement

Current deepfake detection models face three critical limitations:

1. Lack of **cross-lingual robustness**, making them ineffective in multilingual settings.
2. Absence of **modality-specific explainability**, leaving decision-making processes opaque.
3. Inadequate **temporal modelling**, as they often rely on short video segments.

4. Objectives

- Develop a **dual-stream Transformer model** that learns both audio and visual patterns and their cross-modal relationships.

- Implement a **modality-specific explainability mechanism** to localize the source of forgery.
- Evaluate **multilingual generalization** using synthetic and real datasets across several languages.
- Integrate **GAN-aware modules** to capture subtle generative artifacts in both modalities.
- Benchmark against state-of-the-art methods using both quantitative and qualitative metrics.

5. Literature Review (Summary)

Recent models like AVFF, AVTENet, and ART-AVDF leverage joint audio-visual fusion but are limited to English datasets and lack explainability. Transformer-based models like AVT2-DWF offer promising fusion mechanisms but do not address multilingualism or attribution transparency. The literature also lacks studies that evaluate cross-lingual performance or explain decisions at a modality level, leaving a clear research gap.

6. Methodology

6.1 Model Architecture:

- Dual-stream architecture using Transformer encoders for both audio and video.
- Cross-modal consistency module with attention-based fusion.
- GAN-aware feature extraction for texture and frequency anomalies.

6.2 Datasets:

- Real datasets: DFDC, FakeAVCeleb, VOXCeleb.
- Synthetic multilingual data: Generated using TTS (Text-to-Speech) + Wav2Lip in English, Chinese, French, and Urdu.

6.3 Training Strategy:

- Stage 1: Self-supervised learning on real AV pairs.
- Stage 2: Supervised fine-tuning on multilingual fake-real pairs.

6.4 Evaluation Metrics:

- Quantitative: Accuracy, AUC, F1-score, EER.
- Qualitative: Saliency maps, modality attribution scores, visualizations.
- Baselines: AVFF, AVTENet, AVT2-DWF.

7. Novelty and Significance

This work introduces:

- The first **multilingual AV deepfake detection benchmark**.
- A novel **modality-aware explainability mechanism**.
- A GAN-aware, temporal consistency-driven detection pipeline.

The proposed work directly addresses gaps in cross-lingual generalization, transparency, and real-world applicability.

8. Expected Outcomes

- A high-performing, explainable deepfake detection system with multilingual generalization.
- Synthetic multilingual datasets for AV forgery detection.
- Visual and numerical insights into detection performance and modality saliency.
- Reproducible code and open-source benchmarks for the community.

9. Timeline (3-Month Plan)

Month	**Activity**
1	- Comprehensive literature review and gap analysis - Collection of benchmark datasets and generation of multilingual synthetic data - Initial experiments with baseline models (AVFF, AVTENet)
2	- Design and development of the dual-stream Transformer-based detection model - Implement GAN-aware and cross-modal attention modules - Begin self-supervised pretraining on real audio-visual data
3	- Supervised fine-tuning on multilingual fake-real datasets - Integration of explainability (saliency/attention) mechanisms - Model evaluation, ablation studies, and baseline comparisons - Prepare draft of research paper for submission to SCI Q1 journal

10. References

1. Li, H., et al. "AVFF: Audio-Visual Forgery Detection Framework." *CVPR*, 2024.
2. Zhou, P., et al. "AVTENet: Audio-Visual Transformer Ensemble for Deepfake Detection." *IEEE TIFS*, 2023.
3. Nanduri, V., et al. "ART-AVDF: Articulatory Representation for Audio-Visual Deepfake Detection." *Computer Vision and Image Understanding*, 2024.
4. Wang, Y., et al. "AVT2-DWF: Dual-Weight Fusion for Audio-Visual Deepfake Detection." *Pattern Recognition*, 2024.
5. Korshunov, P., & Marcel, S. "Deepfakes: A New Threat to Multimedia Forensics?" *IEEE Signal Processing Magazine*, 2020.