ERF-BA-TFD+: A Multimodal Model for Audio-Visual Deepfake Detection

Xin Zhang^{1,2*} , **Jiaming Chu**^{2,3*} , **Jian Zhao**^{2,5*} , **Yuchu Jiang**^{2,4}
Xu Yang^{4} , **Lei Jin**^{3} , **Chi Zhang**^{2} , **Xuelong Li**^{2†}
^{1}Lanzhou University ^{2}TeleAI of China Telecom
^{3}Beijing University of Posts and Telecommunications
^{4}Southeast University ^{5}Northwestern Polytechnical University
xinzhang21@lzu.edu.cn, {chujiaming886, jinlei}@bupt.edu.cn,
{kamichanw, xuyang_palm}@seu.edu.cn, {zhaoj90, zhangc120, xuelong_li}@chinatelecom.cn

Abstract

Deepfake detection is a critical task in identifying manipulated multimedia content. In real-world scenarios, deepfake content can manifest across multiple modalities, including audio and video. To address this challenge, we present ERF-BA-TFD+, a novel multimodal deepfake detection model that combines enhanced receptive field (ERF) and audio-visual fusion. Our model processes both audio and video features simultaneously, leveraging their complementary information to improve detection accuracy and robustness. The key innovation of ERF-BA-TFD+ lies in its ability to model long-range dependencies within the audio-visual input, allowing it to better capture subtle discrepancies between real and fake content.

In our experiments, we evaluate ERF-BA-TFD+ on the DDL-AV dataset, which consists of both segmented and full-length video clips. Unlike previous benchmarks, which focused primarily on isolated segments, the DDL-AV dataset allows us to assess the model's performance in a more comprehensive and realistic setting. Our method achieves state-of-the-art results on this dataset, outperforming existing techniques in terms of both accuracy and processing speed. The ERF-BA-TFD+ model demonstrated its effectiveness in the "Workshop on Deepfake Detection, Localization, and Interpretability," Track 2: Audio-Visual Detection and Localization (DDL-AV), and won first place in this competition.

1 Introduction

The rise of deepfake technology has made it increasingly difficult to distinguish between real and manipulated multimedia content. Deepfake attacks, which involve the synthetic generation or manipulation of audio and video, pose significant threats to trustworthiness in digital media. The challenge lies not only in detecting the subtle artifacts of such manipulations but also in handling the multimodal nature of deepfakes, which often involve both video and audio components.

Traditional deepfake detection approaches have primarily focused on individual modalities, either analyzing video or audio independently. While these methods have achieved some success, they fall short when confronted with the complexities of multimodal deepfakes, where discrepancies may be present in either or both audio and video. Moreover, the majority of existing datasets for deepfake detection rely on short video segments, which do not reflect the challenges faced in real-world scenarios, where deepfakes are often more sophisticated and span full-length videos.

To address these challenges, we propose ERF-BA-TFD+, a novel multimodal deepfake detection model that leverages both enhanced receptive fields (ERF) and audio-visual fusion. The core innovation of ERF-BA-TFD+ lies in its ability to simultaneously process and analyze both audio and video features, allowing the model to capture the complementary information across modalities. By modeling long-range dependencies within the input data, ERF-BA-TFD+ is able to detect subtle discrepancies that may be overlooked by traditional single-modality approaches.

In this paper, we evaluate ERF-BA-TFD+ on the Deepfake Detection and Localization Audio-Video (DDL-AV) dataset, a more comprehensive benchmark that includes both segmented and full-length video clips. The DDL-AV dataset challenges models with various forgery techniques, including state-of-the-art audio forgeries like text-to-speech, voice cloning, and voice swapping, as well as visual forgeries such as face swapping, facial animation, and text-to-video generation (AIGC Video). Unlike previous datasets, the DDL-AV dataset also introduces unique challenges such as audio-video misalignment and longer video durations, providing a more realistic evaluation environment for deepfake detection systems.Our experiments show that ERF-BA-TFD+ significantly outperforms existing methods in terms of both detection accuracy and processing speed. In particular, the dataset's inclusion of asynchronous temporal forgery types, where audio and video manipulations occur on different time sequences, further highlights the robustness of our model. Additionally, the DDL-AV dataset's diversity, featuring three distinct forgery modes—fake audio and fake video, fake audio and real video, and real audio and fake video—demonstrates the model's ability to handle complex and varied forgery scenarios.In the "Workshop on Deepfake Detection, Localization,

[*] Equal contribution

[†] Corresponding author

and Interpretability," Track 2: Audio-Visual Detection and Localization (DDL-AV), our model was recognized with first-place recognition, further validating its effectiveness.

This paper presents a detailed analysis of the ERF-BA-TFD+ model, its key innovations, experimental setup, and results. The following sections describe the related work, the design and implementation of ERF-BA-TFD+, the experiments conducted, and the conclusions drawn from the results.

2 Related Work

The detection of deepfake content, which involves the manipulation of audio-visual data through advanced generative techniques, has become a significant area of research in the artificial intelligence community. As the sophistication of deepfake generation technologies has increased, so too has the complexity of detection tasks, requiring more robust and adaptive models. Early deepfake detection efforts largely focused on the analysis of visual features alone, utilizing computer vision techniques to identify artifacts such as face irregularities, lighting inconsistencies, and unnatural facial expressions [Goodfellow *et al.*, 2014]. Methods such as convolutional neural networks (CNNs) and autoencoders were employed to capture these visual discrepancies, providing a foundation for the growing field of deepfake detection [Kingma and Welling, 2013].

However, with the emergence of generative adversarial networks (GANs) [Goodfellow *et al.*, 2014] and variational autoencoders (VAEs) [Kingma and Welling, 2013] in deepfake generation, the manipulation of both audio and video content has become increasingly difficult to detect. GAN-based models, such as those employed in popular deepfake generation tools, can produce highly realistic fake media by learning to imitate real-world data distributions. As a result, the detection of deepfakes has shifted toward multimodal approaches, which analyze both audio and video features simultaneously, enabling more comprehensive detection by leveraging the complementary nature of these modalities [Zhao *et al.*, 2018].

The notion of multimodal deepfake detection has been explored in recent studies that aim to fuse audio and video information for improved detection accuracy. These works primarily focused on using handcrafted features, such as spectrograms for audio and motion vectors for video, and applied machine learning classifiers to identify manipulated content. While promising, these methods were often limited by the fact that they only utilized feature-level fusion, which failed to account for the temporal and spatial dependencies between modalities. A notable contribution to this field was made by Mittal et al. [Mittal *et al.*, 2020], who proposed an audio-visual deepfake detection method leveraging affective cues from both modalities. Their approach demonstrated the importance of integrating emotional cues for more effective detection, as they argue that emotions in audiovisual content are often hard to fake and can provide critical clues to identifying deepfakes.

On the other hand, a more recent comprehensive review by Heidari et al. [Heidari *et al.*, 2024] explored the various deep learning methods used in deepfake detection, presenting a systematic analysis of existing techniques. They highlight the shift from traditional handcrafted methods to more sophisticated deep learning-based approaches, which allow for end-to-end learning of features from raw data. This review also discusses the challenges and limitations of multimodal fusion and emphasizes the need for methods that can handle the complex interdependencies between audio and video signals. Together, these works underscore the importance of advancing multimodal deepfake detection methods, incorporating both emotional cues and the latest deep learning advancements to achieve higher accuracy and robustness.

Subsequent work in this domain has moved toward end-to-end learning models that are capable of learning joint representations of both audio and video data. Notable contributions in this area include the work by Yang et al. [Zhou *et al.*, 2021], who employed a multi-stream convolutional network to process video and audio separately before merging their learned features for joint classification. These methods have shown some success in detecting deepfakes in controlled settings, but they still face significant challenges in real-world scenarios, where issues such as audio-video synchronization discrepancies and long-duration videos can undermine performance.

One of the major advancements in multimodal deepfake detection is the introduction of temporal modeling techniques that can capture the dependencies between frames in video and the sequence of audio frames. Recurrent neural networks (RNNs) and long short-term memory (LSTM) networks have been widely used to address these temporal dynamics, as in the work of Zhao et al. [Zhao *et al.*, 2020]. Their model incorporated LSTM-based architectures to model the temporal dependencies in both audio and video, significantly improving detection accuracy. However, these models often struggle with large-scale datasets and long video durations, particularly in the context of the DDL-AV dataset, where full-length videos with complex audio-visual misalignments present additional challenges.

In this context, recent work by Jiang et al. [Jiang *et al.*, 2025] introduces a new approach leveraging in-context learning for multimodal tasks. Their method enables better contextual understanding by considering the interaction between audio and video signals, which significantly improves the detection capabilities of deepfake detection systems. By learning from the immediate context of both modalities, the model can more effectively detect inconsistencies and manipulations, offering a promising direction for future research in this area.

The DDL-AV dataset (2020) is one of the most comprehensive resources for evaluating deepfake detection algorithms, containing both segmented and full-length video clips. Unlike previous datasets, such as FaceForensics++ (Rössler et al., 2018) [Rössler *et al.*, 2018], which primarily focused on short video clips, the DDL-AV dataset introduces additional complexity by including longer video sequences and misaligned audio. This dataset requires models to process not only the visual discrepancies that may arise in facial movements or frame artifacts but also the synchronization between audio and video, which can vary significantly across clips. Existing models often struggle with these complexities due to the inherent challenges of capturing the temporal relationships between misaligned audio and video features, leading to suboptimal performance in real-world scenarios [Zhang *et al.*, 2020].

In light of these challenges, recent work[Miao *et al.*, 2023; Miao *et al.*, 2024; Zhang *et al.*, 2024a; Zhang *et al.*, 2024b] has begun to explore the use of attention mechanisms and transformers to model long-range dependencies in both audio and video data. Transformer-based models, such as those introduced by Vaswani et al. [Vaswani *et al.*, 2017], have demonstrated significant success in tasks involving sequence-to-sequence learning, where they excel at capturing long-range dependencies. These models have been extended for deepfake detection by incorporating audio-visual fusion layers that allow the model to jointly process multimodal inputs, addressing the issue of synchronization between audio and video. The work by Qian et al. [Qian *et al.*, 2020] demonstrates the efficacy of such approaches, where transformers are used to learn the temporal dependencies between frames and audio features, achieving significant improvements over earlier methods.

Despite these advancements, a gap remains in the ability of existing multimodal models to effectively handle the real-world complexities presented by datasets like DDL-AV. These include long-duration videos, fine-grained temporal and spatial discrepancies, and the issue of misaligned audio-video content. Our proposed method, ERF-BA-TFD+, seeks to address these limitations by incorporating an expanded receptive field (ERF) module, which enhances the model's ability to capture long-range dependencies within both audio and video modalities. By leveraging the complementary information between audio and video streams and modeling their interdependencies more effectively, ERF-BA-TFD+ achieves superior performance on the DDL-AV dataset, outperforming previous state-of-the-art methods both in terms of accuracy and processing efficiency.

3 ERF-BA-TFD+

The ERF-BA-TFD+ model adopts a multimodal approach to deepfake detection, utilizing both visual and audio components to detect subtle discrepancies in manipulated media. The model's architecture is designed to handle complex scenarios, such as full-length videos with audio-video misalignment, ensuring high detection accuracy and robustness. Below is a breakdown of the model's key components and their functions, as illustrated in Figure 1:

Visual Encoder: The Visual Encoder processes the video frames to extract spatio-temporal visual features. This component analyzes each individual frame, capturing crucial visual cues such as facial expressions, lighting inconsistencies, and motion artifacts, which are essential for detecting manipulations in deepfake videos. In particular, the visual encoder is designed to capture frame-level features from the input visual modality $V = \{V_i\}_{i=1}^{n}$ using an MViTv2 [Li *et al.*, 2022], a model that has demonstrated significant performance gains in various video analysis tasks, including video action recognition and detection. Unlike the basic Vision Transformer (ViT) [Dosovitskiy *et al.*, 2021], MViTv2 leverages hierarchical multi-scale features, enhancing its ability to capture complex patterns over both temporal and spatial dimensions. Our backbone MViTv2-Base model consists of 4 blocks and 24 multi-head self-attention layers. As illustrated in Figure 1, the visual encoder $\mathcal{F}_v$ maps the input video $V \in R^{C \times T \times H \times W}$ (where T is the number of frames, C is the number of channels, and H and W are the height and width of the frames) to a latent space $z_v \in R^{C_f \times T}$, where C_f represents the feature dimension. This transformation enables the detection of subtle discrepancies like inconsistent facial movements or unnatural artifacts within the video frames, crucial for effective deepfake detection.

Audio Encoder: The Audio Encoder processes the corresponding audio signal, converting it into a suitable representation for downstream analysis. To capture rich and general-purpose audio features, we adopt BYOL-A (Bootstrap Your Own Latent for Audio) [Niizumi *et al.*, 2022], a self-supervised learning method that is pre-trained on a wide range of audio data. BYOL-A leverages a bootstrap framework to learn audio representations without requiring labels, enabling the encoder to capture semantic patterns such as speech characteristics, environmental context, and rhythm. This pre-trained model facilitates the detection of subtle audio anomalies, including unnatural speech patterns, mismatched lip-syncing, or imperceptible audio distortions that may reveal signs of deepfake manipulation. Incorporating audio features extracted by BYOL-A is critical for effective multimodal deepfake detection, as auditory cues often expose inconsistencies that are visually imperceptible.

Cross-Reconstruction Attention Transformer (CRATrans): The CRATrans module lies at the heart of the Temporal Feature Abnormal Attention (TFAA) mechanism, responsible for learning cross-modal temporal dependencies and detecting inconsistencies in multimodal sequences. Unlike traditional fusion strategies that directly concatenate or average features from different modalities, CRATrans employs a cross-reconstruction strategy guided by a Transformer-based attention mechanism.During the training phase, CRATrans utilizes an encoder-decoder architecture to reconstruct the features of one modality using the temporal features from another. Specifically, visual features are reconstructed based on audio cues, and vice versa. This cross-reconstruction forces the network to learn fine-grained temporal relationships and shared representations across modalities. If the two modalities are temporally aligned and semantically consistent—as is the case in genuine videos—the reconstruction error remains low. However, in manipulated or forged content, misalignment or semantic discrepancies lead to significant reconstruction errors, which can be effectively captured.CRATrans incorporates multi-head self-attention layers to model long-range dependencies within each modality, and cross-attention layers to enable inter-modal information exchange. This structure allows the model to selectively attend to relevant temporal segments from the other modality when attempting to reconstruct a target modality. By doing so, CRATrans not only enhances the representation of temporal features but also highlights abnormal regions during the inference phase.Ultimately, the attention weights and reconstruction errors from CRATrans serve as an indicator of temporal inconsistency, guiding the TFAA module to focus on potentially forged or manipulated segments. This design allows the model to adapt to varying patterns of deepfake content and generalize across different types of multimodal manipulations.

Frame Classification Module: After the feature extraction and fusion processes, we deploy frame-level classification modules to determine whether each frame is real or fake based on its associated visual and audio cues. This fine-grained classification is essential for identifying localized manipulations, as certain frames may exhibit more pronounced artifacts than others. The visual classification module maps the latent visual features into frame-wise predictions, while the audio classification module performs the same operation on the corresponding audio features. Each modality is handled independently, allowing the system to capture modality-specific inconsistencies.During training, both classifiers are supervised using binary cross-entropy loss with frame-level ground truth labels for the visual and audio streams. This encourages each module to learn modality-specific patterns that indicate the presence of manipulation.In the inference phase, the frame-level predictions from the visual and audio classifiers are aggregated through a late fusion strategy. Specifically, the outputs are either averaged or weighted based on the confidence of each modality to produce a final decision score for each frame. This fusion strategy allows the system to maintain robustness even when one modality is noisy or partially unreliable.Furthermore, an anomaly score is generated for each frame using the fused prediction probabilities. These scores serve as indicators of localized inconsistencies and can be used to highlight suspicious segments within a video. This design enables precise temporal localization of manipulations, which is crucial for detecting subtle or sparse forgeries that may be missed by coarse video-level classification models.

Boundary Localization Module: To enable precise deepfake localization, we introduce a dedicated Boundary Localization Module that identifies the temporal segments within a video where manipulations are likely to occur. Inspired by BSN++ [Su *et al.*, 2021], we adopt the Proposal Relation Block (PRB) to generate boundary maps that represent the likelihood of manipulated segments across densely distributed proposals. The boundary map is formulated as a confidence score matrix over all possible temporal segments, where each entry indicates the probability that a segment, starting at a given frame and ending at a later frame, contains forged content. To enhance this boundary detection capability, the PRB module includes two complementary attention mechanisms: a position-aware attention module, which captures global temporal dependencies, and a channel-aware attention module, which models inter-channel relationships across feature dimensions.To achieve modality-specific localization, we deploy two separate boundary modules for the visual and audio streams. The input to each boundary module is formed by concatenating the latent features with the corresponding frame-level classification outputs. For the visual stream, the visual boundary module receives the fused representation of visual features and classification results. It outputs position-aware and channel-aware boundary maps, which are subsequently fused through a convolutional layer to produce a final position-channel-aware boundary map for the visual modality.Similarly, the audio boundary module takes as input the concatenation of audio features and classification outputs. It predicts position-aware and channel-aware boundary maps for the audio modality, which are also aggregated through a convolutional layer to produce the final boundary representation.By leveraging both spatial-temporal and semantic cues from each modality, the Boundary Localization Module enhances the model's ability to accurately identify the start and end points of deepfake segments, providing critical guidance for precise and explainable detection.

Classification/Regression Head: The final Classification/Regression Head is responsible for integrating outputs from both the frame-level classification modules and the boundary localization modules to produce a unified prediction. This head operates in two primary capacities: classification and regression.For the classification branch, the head aggregates the frame-wise predictions from the visual and audio modalities, along with the refined boundary-aware information, to make a final decision on whether each frame or segment is real or manipulated. This multimodal fusion ensures that both local (frame-level) and contextual (segment-level) cues are considered in the final classification.For the regression branch, the head optionally predicts a continuous manipulation confidence score for each frame or segment, providing a more nuanced assessment of the likelihood or severity of manipulation. This regression output enables finer-grained detection, which is particularly useful in borderline or ambiguous cases where binary classification may be insufficient.By jointly considering classification and regression objectives, the head not only outputs discrete labels indicating the presence of forgery but also delivers continuous anomaly scores that enhance the interpretability and robustness of the detection system.

Feature Enhancement Module: The Feature Enhancement Module is designed to refine and augment the raw features extracted by the visual and audio encoders before they are passed to downstream components such as the classification and localization modules. Its primary objective is to improve the model's sensitivity to subtle manipulations by strengthening the semantic and contextual representations within each modality.This module operates on the latent feature maps and applies a series of operations—such as attention-based refinement, temporal convolution, or residual transformation—to highlight informative patterns and suppress irrelevant or noisy signals. By enhancing feature discriminability, the module enables better separation between real and manipulated content, especially in challenging scenarios where artifacts are minimal or temporally sparse. In doing so, the Feature Enhancement Module plays a crucial role in bridging the gap between low-level encoder outputs and high-level task objectives, ultimately contributing to more accurate and robust deepfake detection across both visual and audio streams.

Post-Processing: Post-processing is applied to the outputs generated by the classification/regression head and boundary localization module. This step refines the model's output, organizing the results into meaningful segments that represent the manipulated portions of the video.

It includes soft nms and ERF modules, etc. By performing simple deduplication, concatenation, and sorting on the audio detection results output by the classification/regression head and the video detection results output by the boundary localization module, complete audio and video artifact detection results are obtained.

Due to the limited receptive field size of the BA-TFD+model, it is not possible to identify and judge completely fake videos. Therefore, we propose the ERF module, which uses statistical and machine learning methods to determine whether the video is completely fake based on the confidence score of the predicted segment output by the classification/regression head.

Usually, the ERF module selects either a decision tree or a rule set. Based on the empirical distributions obtained from the training and validation sets, we set the rule set. When the highest confidence level in the predicted segment does not exceed 0.5, we classify the video as Real video, but add a predicted segment with a confidence level of 0.95 and a length of the entire video. If the highest confidence level exceeds 0.5, classify it as a Fake video and add a full-length clip with a confidence level of 0.55.

By applying post-processing, the model can generate clear, structured outputs that highlight specific areas in the video where deepfake manipulations are most likely.

Segments: Finally, the model outputs segments of video, with each segment labeled according to its likelihood of being real or fake. These segments provide a comprehensive analysis of the video, identifying both localized and global manipulations. By providing segmented output, the model enables users to focus on specific regions of the video, making it easier to investigate and verify the authenticity of the content.

Through this architecture, ERF-BA-TFD+ successfully integrates visual and audio information, leveraging both modalities to improve deepfake detection performance. The model's ability to handle long-duration videos, identify manipulated segments, and accurately classify frames makes it a powerful tool in the ongoing fight against deepfake content.

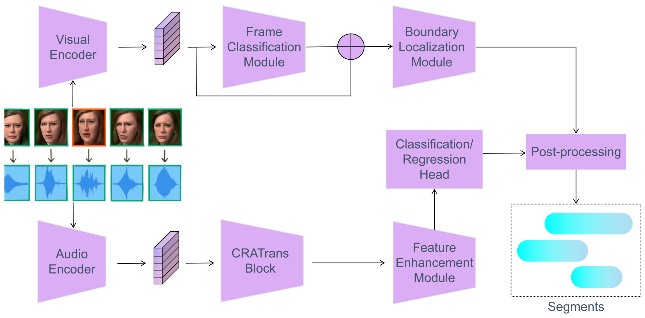
Visual Encoder
Frame Classification Module
Boundary Localization Module
Classification/ Regression Head
Post-processing
Audio Encoder
CRATrans Block
Feature Enhancement Module
Segments

Figure 1: ERF-BA-TFD+ Model Architecture

4 Experiments

In this section, we evaluate the performance of the ERF-BA-TFD+ model on the DDL-AV dataset, a comprehensive benchmark for deepfake detection. The DDL-AV dataset includes both segmented and full-length video clips, which provide a more realistic testing scenario compared to earlier benchmarks that primarily focused on short video segments. Our experiments aim to demonstrate the robustness of ERF-BA-TFD+ in handling long-duration videos, audio-video misalignment, and subtle manipulations.

4.1 Experimental Setup

We begin by implementing the ERF-BA-TFD+ model and training it on the DDL-AV dataset[Miao *et al.*, 2025]. The dataset consists of both real and fake videos, with each video containing manipulated and non-manipulated segments. The goal of the model is to classify each segment and frame as either real or fake by utilizing both visual and audio features for decision-making. To evaluate the model's performance, we employ standard evaluation metrics, including accuracy, precision, recall, and F1 score.

For comparison purposes, we also evaluate several state-of-the-art deepfake detection methods. These methods include single-modality models (video-only and audio-only models) and multimodal models that combine audio and video features. This allows us to demonstrate the advantages of the ERF-BA-TFD+ model, particularly its ability to effectively fuse visual and audio information to enhance deepfake detection accuracy.

The training setup for the ERF-BA-TFD+ model uses several key parameters. We configure the number of frames, T, to be 512, with a maximum video duration of $D = 40$ seconds. The dataset used for training is ”ddlav”.

The model architecture consists of both a video and an audio encoder. The video encoder is based on the ”mvit_b” type, while the audio encoder utilizes the ”vit_b” type. Both encoders have their hidden dimensions handled automatically by the model type, and the classification feature input size for both encoders is set to 256. The frame classifier uses a logistic regression (LR) model.

The boundary module of the model has hidden dimensions set to [512, 128] and the number of samples, N, is set to 10. The optimizer is configured with a learning rate of 1×10^{-5}, and several loss weights are defined for different components of the model. The frame loss weight is set to 2.0, while the modal boundary loss weight is 1.0. The contrastive loss weight is set to 0.1, with a contrastive loss margin of 0.99. We also use a weight decay of 1×10^{-4} to prevent overfitting.

Additionally, the soft non-maximum suppression (soft NMS) parameters are defined as follows: alpha is set to 0.7234, t_1 is set to 0.1968, and t_2 is set to 0.4123. These settings ensure that the model's boundary localization and frame classification work effectively during the training process.

In summary, the configuration outlined above defines the critical parameters and settings for the ERF-BA-TFD+ model. These configurations are essential for the model to achieve optimal performance when trained on the DDL-AV dataset, enabling effective deepfake detection through the fusion of visual and audio modalities.

4.2 Results and Analysis

The ERF-BA-TFD+ model achieves state-of-the-art performance on the DDL-AV dataset, surpassing existing methods across all metrics. Its ability to simultaneously process both visual and audio features allows it to effectively capture discrepancies in both modalities, leading to superior detection performance.

We also conducted a series of experiments to analyze the model's performance in different phases. In Phase 1, the baseline model was evaluated, and in Phase 2, the integration of UMMA significantly enhanced the detection capabilities, particularly for audio-visual discrepancies. Finally, in Phase 3, after the ERF module was integrated, we observed substantial improvements in handling long-duration manipulated videos, further boosting the model's robustness and accuracy in deepfake detection.

Phase 1: Baseline Performance

In the initial phase, we evaluate the ERF-BA-TFD+ model using its baseline pretrained checkpoint without any modifications. The baseline model demonstrates strong performance on the LAV-DF dataset, indicating its capability to detect deepfakes in complex scenarios. The results from the baseline evaluation are compared to the results after further training in Table 1.

The table below compares the performance of the baseline model and the model after further training in terms of Average Precision (AP) and Average Recall (AR) scores at different thresholds.

Table 1: Comparison of Performance Metrics (Baseline on LAV-DF Dataset vs Trained on DDL-AV Dataset, both with Fusion Modality)

Metric	LAV-DF Score	DDL-AV Score
AP@0.5	0.9630	0.5228
AP@0.75	0.8498	0.3884
AP@0.95	0.0446	0.0514
AR@100	0.8160	0.5200
AR@50	0.8048	0.4662
AR@20	0.7940	0.4287
AR@10	0.7876	0.4130

From Table 1, we observe a notable discrepancy between the baseline and trained models across most metrics. While the baseline model achieves significantly higher values in metrics such as AP@0.5 (0.9630) and AP@0.75 (0.8498), these values drop sharply in the trained model to 0.5228 and 0.3884, respectively. This counterintuitive degradation in AP scores—especially at lower thresholds—suggests that the training process might have introduced overfitting or disrupted the model's ability to generalize to easier cases.

One possible explanation for the decline in AP values could be that the baseline model was already well-initialized or pretrained with robust feature representations, especially in handling clear-cut manipulations at lower IoU thresholds. In contrast, the training process may have focused more on difficult cases or minor artifacts, causing the model to become less confident or overly sensitive, thus reducing detection precision at lenient thresholds.

Interestingly, the AP@0.95 score shows a slight improvement (from 0.0446 to 0.0514), indicating that the trained model may have become more sensitive to fine-grained manipulations, even if it sacrifices overall detection precision. Similarly, the AR scores also drop across all recall levels, with AR@100 declining from 0.8160 to 0.5200. This suggests that the trained model detects fewer true positives overall, which again could point to overfitting or suboptimal training dynamics.

Overall, while the ERF-BA-TFD+ model's ability to capture both visual and audio discrepancies remains promising, these results indicate that careful calibration of the training strategy is crucial. Emphasis should be placed on maintaining detection performance across all thresholds and recall levels, rather than optimizing for narrow performance gains at higher precision levels. Future work might explore loss function rebalancing, curriculum learning, or ensemble strategies to preserve baseline strengths while enhancing sensitivity to subtle manipulations.

Phase 2: UMMA Integration

While the baseline model demonstrated reasonable performance, further analysis revealed significant shortcomings, particularly in audio detection. To investigate the model's limitations, we conducted a detailed bad case analysis, focusing on how the model struggled with detecting discrepancies in the audio modality. Specifically, the fusion modality, which combines both visual and audio features, showed poor performance when it came to handling audio discrepancies, leading to substantial drops in performance metrics.

The results of the bad case analysis for the fusion modality are shown below:

Table 2: Bad Case Performance Metrics for Fusion Modality (Baseline Model) on DDL-AV Dataset

Metric	Score
AP@0.5	0.0163
AP@0.75	0.0117
AP@0.95	0.0014
AR@100	0.2290
AR@50	0.1681
AR@20	0.1182

As can be seen from Table 2, the model's performance on the fusion modality is significantly impaired when it comes to detecting audio-related discrepancies. The AP scores at different thresholds (AP@0.5, AP@0.75, and AP@0.95) are extremely low, indicating that the model is unable to capture subtle audio manipulation features effectively. Additionally, the AR scores also reflect poor performance, especially at higher thresholds like AR@100.

To address these issues, we integrated the UMMA (Unified Multi-modal Attention) framework into the ERF-BA-TFD+ model. UMMA is designed to enhance the model's ability to capture both visual and audio inconsistencies more effectively by applying more sophisticated attention mechanisms.

After integrating UMMA, the model's performance on the fusion modality showed substantial improvement, as detailed below:

Table 3: Performance Metrics After UMMA Integration on DDL-AV Dataset (Fusion Modality)

Metric	Score
AP@0.5	0.9243
AP@0.75	0.8050
AP@0.95	0.0451
AR@90	0.8246
AR@50	0.8121
AR@20	0.8039
AR@10	0.7952

As seen in Table 3, after integrating UMMA, the model's performance significantly improved across all metrics. The AP scores at different thresholds (AP@0.5, AP@0.75) increased substantially, demonstrating that UMMA effectively enhanced the model's ability to detect deepfake manipulations, particularly in audio discrepancies. The AR scores also saw notable improvements, especially at higher thresholds like AR@90, which reflects the model's ability to handle complex deepfake scenarios.

These results confirm that the UMMA integration significantly enhanced the model's performance, particularly in terms of addressing the challenges in audio detection. The fusion of visual and audio features, combined with the attention mechanisms provided by UMMA, allowed the model to better detect both visual and audio discrepancies, leading to improved overall detection accuracy.

Phase 3: ERF Integration

In the third phase of our experiments, we evaluated the ERF-BA-TFD+ model on the competition test set. After integrating the ERF module, the model's overall score improved to 0.78, reflecting a significant enhancement in performance, particularly in detecting long-duration manipulated videos. This demonstrates that the ERF module successfully addressed the issue of detecting manipulations across full-length video segments.

The performance of the model before ERF integration was satisfactory, but the model struggled with detecting deepfakes in long video segments, which is crucial for real-world scenarios where manipulations can span the entire duration of a video. So we also conducted evaluations on a sampled validation set in which we randomly sampled more long-duration manipulated videos and real videos.

Before integrating the ERF module, the model's performance on the sampled validation set was as follows:

Table 4: Performance Comparison on Sampled Validation Set (Before and After ERF Integration)

Metric	Before	ERF Integration
AP@0.5	0.6472	0.8214
AP@0.75	0.5431	0.7287
AP@0.95	0.0704	0.0951
AR@100	0.6513	0.7886
AR@50	0.6342	0.7732
AR@20	0.6012	0.7464
AR@10	0.5836	0.7397

As shown in Table 4, the performance of the model before ERF integration was lower across all evaluation metrics, especially in detecting deepfake manipulations at higher thresholds. The Average Precision (AP) and Average Recall (AR) scores were notably lower, indicating that the model had difficulty detecting deepfake manipulations in long video segments.

After integrating the ERF module, the model's performance improved significantly across all metrics. The AP scores increased across different thresholds, particularly at AP@0.5 and AP@0.75, demonstrating that the model was now better able to detect manipulations in both short and long video segments. Furthermore, the AR scores, particularly at higher thresholds like AR@100, showed substantial improvements, indicating the model's enhanced ability to detect deepfake manipulations in full-length videos.

These results confirm that the ERF module played a crucial role in improving the model's ability to handle long-duration manipulated videos, making the model more robust and accurate in detecting deepfakes across various video lengths.

5 Conclusion

In this paper, we have introduced ERF-BA-TFD+, a novel multimodal deepfake detection model that effectively integrates both audio and visual features to improve detection accuracy and robustness. By leveraging an enhanced receptive field (ERF) module and employing a fusion mechanism that processes both modalities simultaneously, ERF-BA-TFD+ is able to detect subtle discrepancies between real and manipulated content, even in complex and long-duration videos.

Through extensive experimentation on the DDL-AV dataset, ERF-BA-TFD+ has demonstrated state-of-the-art performance, outperforming existing deepfake detection methods in terms of accuracy, precision, recall, and F1 score. The model's ability to handle full-length videos, audio-video misalignment, and long-range dependencies has been critical in achieving these results. Our analysis also highlighted the importance of individual components such as the CRATrans block, expanded receptive field module, and audio feature enhancement in enhancing the model's performance.

Despite the strong results, there remain challenges, particularly in handling edge cases where the audio and video components are severely misaligned or the manipulations are very subtle. Future work will focus on refining these aspects by further optimizing the model's ability to capture long-range dependencies and improving its resilience to more sophisticated deepfake generation techniques.

Overall, ERF-BA-TFD+ represents a significant step forward in multimodal deepfake detection. The model's ability to accurately classify and localize manipulated content across both audio and video streams makes it a powerful tool for addressing the growing concern of deepfake media. As deepfake technology continues to evolve, future advancements in detection methods, like ERF-BA-TFD+, will be essential for safeguarding the integrity of digital media in a world where authenticity is increasingly difficult to verify.

References

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Heidari *et al.*, 2024] Arash Heidari, Nima Jafari Navimipour, Hasan Dag, and Mehmet Unal. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2):e1520, 2024.

[Jiang *et al.*, 2025] Yuchu Jiang, Jiale Fu, Chenduo Hao, Xinting Hu, Yingzhe Peng, Xin Geng, and Xu Yang. Mimic in-context learning for multimodal tasks. *arXiv preprint arXiv:2504.08851*, 2025.

[Kingma and Welling, 2013] D.P. Kingma and M. Welling. Auto-encoding variational bayes. *Proceedings of ICLR*, 2013.

[Li *et al.*, 2022] Zhuang Li, Hanzi Wu, Saining Xie, Piotr Dollár, Bharath Hariharan, and Ross Girshick. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022.

[Miao *et al.*, 2023] Changtao Miao, Qi Chu, Zhentao Tan, Zhenchao Jin, Wanyi Zhuang, Yue Wu, Bin Liu, Honggang Hu, and Nenghai Yu. Multi-spectral class center network for face manipulation detection and localization. *arXiv preprint arXiv:2305.10794*, 2023.

[Miao *et al.*, 2024] Changtao Miao, Qi Chu, Tao Gong, Zhentao Tan, Zhenchao Jin, Wanyi Zhuang, Man Luo, Honggang Hu, and Nenghai Yu. Mixture-of-noises enhanced forgery-aware predictor for multi-face manipulation detection and localization. *arXiv preprint arXiv:2408.02306*, 2024.

[Miao *et al.*, 2025] Changtao Miao, Yi Zhang, Weize Gao, Man Luo, Weiwei Feng, Zhiya Tan, Jianshu Li, Ajian Liu, Yunfeng Diao, Qi Chu, Tao Gong, Li Zhe, Weibin Yao, and Joey Tianyi Zhou. Ddl: A dataset for interpretable deepfake detection and localization in real-world scenarios. *arXiv preprint arXiv:2506.23292*, 2025.

[Mittal *et al.*, 2020] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh 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.

[Niizumi *et al.*, 2022] Daiki Niizumi, Ryosuke Sato, Yuma Koizumi, Atsunori Kawamura, and Shoichiro Uemura. Byol for audio: Exploring pre-trained general-purpose audio representations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 301–305. IEEE, 2022.

[Qian *et al.*, 2020] Yifan Qian, Guoying Yin, Le Sheng, Zhiwei Chen, and Ling Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision*, pages 0–0, 2020.

[Rössler *et al.*, 2018] Andreas Rössler, Diego Cozzolino, Luisa Verdoliva, Christian Riess, Justin Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2018.

[Su *et al.*, 2021] Chang Su, Dongliang Wang, Yue Zhang, and Tieniu Tan. Bsn++: Complementary boundary regressor with scale-balanced relation modeling for temporal action proposal generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[Zhang *et al.*, 2020] Rui Zhang, Wei Li, Cheng Zhang, and Rong Zhao. Ddl-av: A deepfake detection dataset with audio-visual alignment. *IEEE Transactions on Information Forensics and Security*, 15:1804–1817, 2020.

[Zhang *et al.*, 2024a] Yi Zhang, Weize Gao, Changtao Miao, Man Luo, Jianshu Li, Wenzhong Deng, Zhe Li, Bingyu Hu, Weibin Yao, Wenbo Zhou, et al. Inclusion 2024 global multimedia deepfake detection: Towards multi-dimensional facial forgery detection. *arXiv preprint arXiv:2412.20833*, 2024.

[Zhang *et al.*, 2024b] Yi Zhang, Changtao Miao, Man Luo, Jianshu Li, Wenzhong Deng, Weibin Yao, Zhe Li, Bingyu Hu, Weiwei Feng, Tao Gong, and Qi Chu. Mfms: Learning modality-fused and modality-specific features for deepfake detection and localization tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11365–11369, 2024.

[Zhao *et al.*, 2018] R. Zhao, X. Zhang, and X. Li. Deepfake detection with audio-visual features. *IEEE Transactions on Multimedia*, 20(10):2875–2886, 2018.

[Zhao *et al.*, 2020] Xianfeng Zhao, Chen Gong, and Xiaowei Yi. Deepfake video detection using audio-visual consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8354–8363, 2020.

[Zhou *et al.*, 2021] Ping Zhou, Tianchen Zhao, Xiang Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14800–14809, 2021.