Multi-modal Deepfake Detection and Localization with FPN-Transformer

Chende Zheng , **Ruiqi Suo** , **Zhoulin Ji** , **Jingyi Deng** , **Fangbin Yi** , **Chenhao Lin**∗ , **Chao Shen**

Xi'an Jiaotong University

{zhengchende, srain22, 310449, yi446320, misscc320}@stu.xjtu.edu.cn, linchenhao@xjtu.edu.cn, chaoshen@mail.xjtu.edu.cn

Abstract

The rapid advancement of generative adversarial networks (GANs) and diffusion models has enabled the creation of highly realistic deepfake content, posing significant threats to digital trust across audio-visual domains. While unimodal detection methods have shown progress in identifying synthetic media, their inability to leverage cross-modal correlations and precisely localize forged segments limits their practicality against sophisticated, fine-grained manipulations. To address this, we introduce a multi-modal deepfake detection and localization framework based on a Feature Pyramid-Transformer (FPN-Transformer), addressing critical gaps in cross-modal generalization and temporal boundary regression. The proposed approach utilizes pre-trained self-supervised models (WavLM for audio, CLIP for video) to extract hierarchical temporal features. A multi-scale feature pyramid is constructed through R-TLM blocks with localized attention mechanisms, enabling joint analysis of cross-context temporal dependencies. The dual-branch prediction head simultaneously predicts forgery probabilities and refines temporal offsets of manipulated segments, achieving frame-level localization precision. We evaluate our approach on the test set of the IJCAI'25 DDL-AV benchmark, showing a good performance with a final score of 0.7535 for cross-modal deepfake detection and localization in challenging environments. Experimental results confirm the effectiveness of our approach and provide a novel way for generalized deepfake detection. Our code is available at *https://github.com/Zig-HS/MM-DDL*.

1 Introduction

The rapid advancement of generative adversarial networks (GANs) and diffusion models has revolutionized audio-visual synthesis technologies, enabling the creation of highly realistic deepfake content that is increasingly indistinguishable from authentic media. While these breakthroughs have

driven innovations in entertainment, education, and human-computer interaction, they have also introduced severe security threats. Malicious actors exploit synthetic speech synthesis (e.g., voice conversion and text-to-speech systems) and facial manipulation techniques (e.g., face swapping and attribute editing) to generate deceptive content for identity impersonation, misinformation dissemination, and social engineering attacks. This dual-use nature of generative AI underscores the urgent need for robust deepfake detection and localization frameworks capable of safeguarding digital trust in real-world applications.

Despite significant progress in unimodal deepfake detection (e.g., audio spoofing analysis or video forgery identification), existing methods face two critical limitations. First, most approaches operate in isolated modalities, failing to leverage cross-modal correlations that could enhance detection accuracy and generalization. For instance, audio-visual consistency cues, such as lip-sync alignment or speaker identity coherence, are often ignored in single-modality pipelines. Second, while temporal localization of forged segments (e.g., identifying manipulated intervals in a video or audio clip) is crucial for forensic analysis, current solutions struggle with partial manipulations (e.g., spliced speech segments or localized facial edits) due to their reliance on simplistic binary classification or rigid post-processing heuristics. These shortcomings hinder practical deployment, particularly against evolving deepfake techniques that exploit multimodal and fine-grained editing strategies.

To address these challenges, we propose a unified multi-modal deepfake detection and localization framework based on Feature Pyramid-Transformer (FPN-Transformer). We leverage pretrained self-supervised models (WavLM [Chen *et al.*, 2022] for audio, CLIP [Radford *et al.*, 2021] for video) to extract hierarchical temporal features, followed by masked differential convolution for local context modeling. Subsequently, we construct a multi-scale feature pyramid using R-TLM [Sun *et al.*, 2021] blocks with 1D downsampling convolutions to jointly analyze audio-visual temporal dependencies through local attention mechanisms. Finally, we employ a dual-branch prediction head to simultaneously predict forgery probabilities and precisely localize the start/end offsets of manipulated segments. To validate the effectiveness of our framework, we conduct extensive experiments on the IJCAI'25 DDL-AV benchmark. Experimental results demon-

---

∗Corresponding Author

strate that our proposed method achieves superior detection and localization performance in a comprehensive evaluation. Our main contributions can be summarized as follows:

- We propose a general-purpose temporal data forgery detection model for multimodal deepfake localization, addressing the challenge that existing detection frameworks struggle to handle multimodal forged data.
- We conduct extensive experiments on the IJCAI'25 dataset to validate the effectiveness of our framework with a final score of 0.7535, providing novel solutions toward achieving generalized forgery detection.

2 Related Works

2.1 Synthetic Audio Generation

The rapid advancement of deep learning and generative models has significantly propelled the development of speech synthesis technologies. Early speech synthesis methods primarily relied on parametric synthesis (e.g., formant synthesis [Styger and Keller, 1994], linear predictive coding (LPC) [O'Shaughnessy, 1988]) and concatenative synthesis (e.g., unit selection synthesis [Hunt and Black, 1996]), yet these approaches exhibited notable limitations in naturalness and flexibility.

Currently, the field of speech synthesis is dominated by generative adversarial networks (GANs) and diffusion models, which can produce highly natural speech, often indistinguishable from human voices. For instance, models such as VITS [Kim et al., 2021] and YourTTS [Casanova et al., 2022] integrate variational inference and speaker adaptation techniques, enabling high-quality multi-speaker synthesis with limited data. Additionally, diffusion-based probabilistic models (e.g., DiffWave [Kong et al., 2020], WaveGrad [Chen et al., 2020a]) generate speech through iterative noise addition and denoising processes, further enhancing synthesis quality.

Recent advancements in large language models (LLMs) have spurred the development of autoregressive speech synthesis. Models like VALL-E [Wang et al., 2023] leverage text-driven audio synthesis to achieve more natural speech prosody. However, the progress in speech synthesis has also raised security concerns regarding audio deepfakes. Malicious actors may exploit voice conversion (VC) and text-to-speech (TTS) technologies for identity impersonation, posing threats to social trust and privacy security. Consequently, developing highly robust fake speech detection methods has become an urgent need for both academia and industry.

2.2 Synthetic Video Generation

Synthetic Video Generation is an emerging technological field that aims to fabricate facial expressions and movements of target individuals through AI techniques, thereby producing deceptive content. So far, Synthetic Video Generation can be categorized into three types: Face Swapping, Facial Reenactment, and Attribute Manipulation.

Face-swapping methods aim to replace a target face with a source face while preserving the original context and identity consistency. [Ding et al., 2020] introduced a method that performs reliably across various scenes and maintains identity fidelity. However, it suffers from imperfect face-region blending, often producing boundary artifacts and being sensitive to changes in lighting or pose. [Zendran and Rusiecki, 2021] combined VAE and GAN architectures to better recover facial details and improve latent space controllability, but the training process tends to be unstable, and the method struggles with facial expression alignment.

Facial reenactment methods generate new images or videos that transfer expressions or movements from one face to another. LE-GAN by [Hu et al., 2023] uses a Laplacian pyramid to enhance facial detail and produce more realistic expressions. However, it can suffer distortions under rapid motion or large pose variations. B-GAN [Liu et al., 2020] exploits frequency domain modeling to enhance texture realism and reduce GAN artifacts, yet it's sensitive to occlusions and may produce blurry edges.

Attribute manipulation methods modify facial features such as age, expression, or hairstyle. [Li et al., 2023] proposed SC-GAN, which embeds semantic directions into the latent space for fine-grained control and natural output. However, when editing multiple attributes simultaneously, the model can exhibit semantic conflicts, and undefined regions in the latent space may lead to unnatural results. GANprintR by [Neves et al., 2020] focuses on fingerprint removal from GAN-generated images to enhance their stealthiness. While effective against detection, it often sacrifices texture quality. [Manjula et al., 2022] combined StarGAN, C-GAN, and VAE for flexible multi-attribute control, but the integration is complex and prone to unstable training.

2.3 Synthetic Audio Detection

The field of fake speech detection originated with the use of various digital signal processing algorithms to analyze anomalies in audio signals for distinguishing genuine from fake audio. Early researchers conducted foundational studies through statistical modeling and acoustic feature analysis [Malik and Farid, 2010]. With the release of increasingly large-scale datasets such as ASVspoof [Wang et al., 2020], research in this domain has deepened significantly. Subsequently, deep learning-based approaches have gained prominence, achieving remarkable results across multiple dimensions.

Early approaches to forgery detection relied primarily on digital signal processing techniques without involving neural network-based methods. [Malik and Farid, 2010] proposed leveraging spectral decay characteristics to detect forged speech, achieving reasonable performance while maintaining strong interpretability. With the advent of neural networks, machine learning-based techniques were increasingly adopted by researchers as countermeasures for forged speech detection. [Alegre et al., 2012] introduced an SVM-based robust forgery detection strategy, significantly outperforming traditional digital signal processing methods. Subsequently, deep learning techniques were widely applied to forged speech detection.

Recently, the generalization capability of audio deepfake detection has garnered increasing attention from researchers, as it determines whether a detection method can perform effectively in real-world scenarios or merely achieve seemingly high metrics on limited academic datasets. [Müller et al.,

2022] discusses the impact of different factors on generalization and highlights that many existing approaches perform poorly on real-world data, indicating that some solutions in the research community have become overly tailored to specific datasets, thereby losing generalizability and practical usability. [Chen *et al.*, 2020b] improves the loss function by employing IMCL, a loss supervision mechanism that enhances feature separation, instead of relying solely on simple binary cross-entropy loss. Some researchers have also adopted single-domain generalization techniques to ensure that the feature space exhibits relatively strong generalization [Xie *et al.*, 2023]. These methods are typically trained and fine-tuned exclusively on large-scale real-world data , enabling them to achieve robust detection results against various unseen attacks.

Simple binary classification has become insufficient for handling certain complex scenarios, such as partially manipulated audio, where only specific segments of a sample are generated by speech synthesis algorithms. These manipulated segments often distort the original semantic meaning of the audio, posing significant risks. [Yi *et al.*, 2022] introduced the concept of utterance-level fake speech detection, demonstrating through their challenge dataset that baseline models employing traditional binary classification methods perform poorly when detecting samples containing partially substituted words or phrases generated by speech synthesis algorithms. Consequently, the detection and localization of partially manipulated speech has emerged as a critical research focus.

2.4 Synthetic Video Detection

Deepfake detection methods focus primarily on three types of features: temporal features, spatial features, and frequency features. Most mainstream detection approaches are based on one of these domains or integrate multiple feature types to improve robustness.

Temporal-based deepfake detection methods focus primarily on the temporal consistency between video frames and abnormal behavioral patterns. These approaches model dynamic changes across frame sequences to identify forgeries. Since fake videos often fail to reproduce genuine human physiological behaviors, such as blinking, head movements, and gaze direction, [Li *et al.*, 2018] detect anomalies in blinking frequency to determine video authenticity. [Yang *et al.*, 2019] analyze inconsistencies in head pose by comparing estimates from all facial landmarks with those from central regions. [Peng *et al.*, 2024] focus on interframe gaze angle variation and propose a spatio-temporal feature fusion module that combines gaze dynamics, spatial attributes, and texture features for classification. Frame-to-frame discontinuities also serve as key indicators of manipulation. [Zheng *et al.*, 2024] reveal that existing detectors suffer from semantic artifacts across diverse scenes and propose a patch-shuffling strategy to break these artifacts for generalized detection. [Yin *et al.*, 2023] design a multiscale spatiotemporal aggregation module to model interframe inconsistencies, while [Choi *et al.*, 2024] observe fluctuations in latent variable styles between frames and develop a style attention module accordingly.

Spatial-based detection methods often analyzing single image frames to identify forged artifacts such as texture anomalies, blending boundaries, and lighting inconsistencies. For example, RECCE [Cao *et al.*, 2022] utilizes reconstruction consistency to detect shadows learned from real images, while LGrad [Tan *et al.*, 2023] converts images into gradient representations using a pre-trained transformation network to amplify subtle artifacts. [Ba *et al.*, 2024] go beyond isolated regions and integrate information from multiple nonoverlapping areas to detect more global inconsistencies. [Miao *et al.*, 2024] further introduce a Mixture-of-Noises Module to enhance RGB features with noise traces, improving localization accuracy in multi-face manipulation detection. Other studies emphasize differences between facial and non-facial regions or leverage fine-grained texture cues [Chai *et al.*, 2020] [Nirkin *et al.*, 2021].

Frequency-based detection methods transform visual information from the spatial or temporal domain into the frequency domain and analyze periodic signal characteristics to reveal manipulation traces.F3-Net [Qian *et al.*, 2020] proposes a dual-branch framework: one branch uses frequency-aware image decomposition (FAD) to learn fine-grained forgery patterns, while the other extracts local frequency statistics (LFS) for semantic-level analysis. HFI-Net[Miao *et al.*, 2022] incorporates global-local interaction modules to explore multi-level frequency artifacts and further enhance detection performance. FreqNet [Tan *et al.*, 2024] focuses on high-frequency components of images, combining them with a frequency-domain learning module to extract source-independent features and improve generalizability. Additionally, [Miao *et al.*, 2023] leverage multi-spectral class centers to suppress semantic information and enhance localization capabilities through frequency-aware features.

3 Methodology

Existing deepfake detection solutions are incapable of achieving generalizable cross-modal detection and localization. Based on our previous unimodal work [Ji *et al.*, 2024], our approach enables the detection and localization of multimodal temporal forgeries. Notably, rather than simultaneously processing audio-visual modalities, our approach treats cross-modal data as unified temporal feature sequences. Figure 1 illustrates the proposed dual FPN-Transformer detection architecture, which comprises three key components: temporal feature embedding and projection module, FPN-Transformer backbone module, and prediction heads. First, during the feature embedding stage, we employ pre-trained self-supervised models (WavLM/CLIP) to encode input temporal data (audio/video) into feature embeddings, generating temporal feature sequences. Subsequently, these embeddings are processed through an encoder composed of lightweight Transformer blocks to obtain the feature pyramid. And finally, classification and regression heads jointly analyze temporal features to predict forgery boundaries at each time step, specifically estimating start times and offset distances of forged segments for precise localization. The following sections will elaborate on these components in detail.
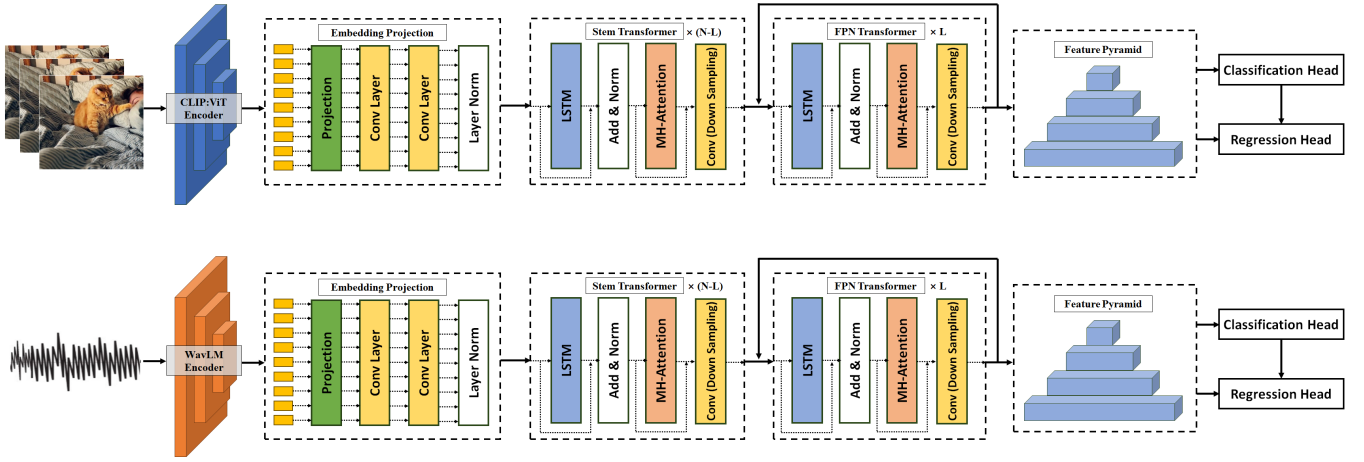
Figure 1: Framework of proposed dual FPN-Transformer detection method. First, we employ pre-trained self-supervised models (WavLM/CLIP:ViT) to encode input temporal data into feature embeddings. Then, these embeddings are processed through an encoder composed of Transformer blocks to obtain the feature pyramid. And finally, prediction heads jointly analyze temporal features to predict forgery boundaries. We train two models separately for audios and videos, and combine the output results.

3.1 Feature Embedding and Projection

Problem Definition. Traditional audio or video forgery detection and localization methods exclusively focus on single modal inputs, inherently limiting their cross-modal generalizability. To address this limitation, we propose a unified problem formulation for multimodal temporal data (audio/video). Specifically, given a temporal input sequence $X = \{x_1, \ldots, x_T\}$, our objective is to generate the corresponding output sequence $Y = \{y_1, \ldots, y_N\}$ through a mapping function $f : X \to Y$. Each element $y_n = (p_n, d_n^s, d_n^e)$ of the output Y denotes one potential forged segment, where p_n denotes the forgery probability of the n-th segment, d_n^s denotes the start time offset relative to the input sequence, and d_n^e denotes the end time offset relative to the input sequence.

Notably, for genuine (non-forged) sequences, the expected output Y should be an empty set. This formulation enables precise temporal localization of forged segments while maintaining cross-modal consistency through unified temporal feature representations.

Considering the distinct temporal characteristics across modalities (e.g., audio data exhibits dense temporal resolution with sampling rates reaching kilohertz levels, while video data presents sparse temporal structure with frame rates limited to hertz levels), we implement a unified feature embedding framework through temporal encoding. Specifically, the input sequence $X = \{x_1, \ldots, x_T\}$ is encoded by encoder e into feature representations $Z = \{z_1, \ldots, z_M\}$. For each feature representation z_i, f_L denotes the temporal sequence length corresponding to z_i and f_S denotes the temporal offset between consecutive feature representations. Consequently, the length M of feature sequence Z is given by the following equation:

$$M = \left\lfloor \frac{T - f_L}{f_S} \right\rfloor + 1 \qquad (1)$$

where $\lfloor \cdot \rfloor$ denotes integer flooring.

To facilitate batch processing, variable-length sequences are standardized through padding/truncation to a maximum length M_{max}, with masking mechanisms ensuring valid temporal context propagation.

Inspired by previous research [Ojha *et al.*, 2023], we employ pre-trained self-supervised models with frozen weights as encoder e. Specifically, WavLM-LARGE [Chen *et al.*, 2022] is adopted for audio data while CLIP:ViT-B/16 [Radford *et al.*, 2021] is adopted for video data. Compared to conventional feature calculating approaches (e.g., LFCC, MFCC, image DFT), these self-supervised models have been exposed to massive training data, enabling superior capability in capturing low-level features critical for differentiating genuine and forged content [Yang *et al.*, 2021]. The architectural design leverages this property to detect subtle discrepancies inherent in generative artifacts.

Subsequently, we employ a set of masked differential convolutional networks to implement feature projection, which facilitates positional embedding integration and effectively captures local temporal context. Specifically, for a given feature embedding $z \in Z$, the output of the masked 1D differential convolution at timestamp t_0 is formulated as:

$$\text{MDC}(t_0) = \theta \cdot \left(-z(t_0) \cdot \sum_{t_n \in D} w(t_n) \right) \qquad (2)$$

$$+ \sum_{t_n \in D} w(t_n) \cdot z(t_0 + t_n) \qquad (3)$$

where t_0 denotes the current timestamp, t_n denotes the enumerated timestamps in offset set D, $w(t_n)$ denotes the learnable convolutional weights, and $\theta \in [0, 1]$ is a hyperparameter balancing the contribution between intensity-level and gradient-level information

3.2 FPN-Transformer Architecture

We employ N layers of R-TLM blocks [Sun *et al.*, 2021] to perform deep feature encoding. Compared to standard Trans-

former architectures, R-TLM incorporates additional LSTM and Fusion layers to explicitly model cross-context representation interactions. The multi-head self-attention (MSA) layer in R-TLM integrates temporal context across the sequence. Notably, we apply localized attention masking to constrain computations within sliding windows, motivated by two factors: (1) forged segments exhibit localized temporal characteristics, and (2) this design significantly reduces computational complexity.

To capture hierarchical temporal features at multiple scales for constructing a feature pyramid, we integrate R-TLM with strided 1D depthwise convolutions. Specifically, we introduce a strided depthwise 1D convolution after each MSA layer. By aggregating outputs from multi-level R-TLM structures, we obtain a hierarchical feature pyramid $\mathcal{F} = \{F^{(1)}, ..., F^{(L)}\}$ with L levels.

A critical component involves temporal alignment between feature sequence timestamps τ and original input timestamps t. Given a virtual timestamp τ_i at the i-th pyramid level and its cumulative stride factor s_i, we map τ_i to the corresponding physical timestamp t in the raw input domain through:

$$t = \left\lfloor \frac{s_i}{2} \right\rfloor + \tau_i \cdot s_i \qquad (4)$$

3.3 Prediction Head

We employ a dual-branch prediction head to decode the feature pyramid $\mathcal{F}$ into the desired output Y. The decoder consists of:

Classification Head. Given feature pyramid $\mathcal{F}$, the classification head evaluates all L pyramid levels at each timestamp t to predict the forgery probability $p(t)$. This is implemented through lightweight 1D convolutional networks attached to each pyramid level, with parameters shared across levels. Specifically, the classification network comprises 3 convolutional layers (kernel size=3), layer normalization (applied to first two layers), and ReLU activation. A final sigmoid function is applied to output dimensions to produce probabilistic forgery predictions.

Regression Head. Distinct from the classification head, the regression head predicts temporal boundaries only when timestamp t lies within forged segments (During inference, we utilize the classification head's output to determine whether a timestamp lies within forged segments). For each pyramid level, we predefine an output regression range to model the start offset d_t^s and end offset d_t^e. The regression head employs 1D convolutional networks with ReLU activation to ensure precise distance estimation. Specifically, the most probable forgery span $[s_t, e_t]$ corresponding to timestamp t is determined by:

$$c_t = \arg\max p(c_t), \quad s_t = t - d_t^s, \quad e_t = t - d_t^e \quad (5)$$

3.4 Loss Function

Our prediction task involves dual objectives: (1) binary classification of forgery probability at each timestamp t, and (2) temporal boundary regression for forged segments (start/end offsets). We design a composite loss function combining two components:

Classification Loss. We employ focal loss [Lin *et al.*, 2017] to address class imbalance between forged and genuine segments. For timestamp t, the classification loss ($\mathcal{L}_{\text{cls}}$) is formulated as:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{T^+} \sum_{t=1}^{T} 1(t \in \Omega^+) \cdot [\log(p_t) + \gamma \cdot (1 - p_t)^\alpha] \quad (6)$$

where p_t denotes predicted forgery probability, Ω^+ represents forged regions, $T^+ = |\Omega^+|$ is the total number of positive samples, and α and γ are hyperparameters balancing hard mining effects.

Regression Loss. For timestamps $t \in \Omega^+$, we minimize DIoU loss ($\mathcal{L}_{\text{reg}}$) [Zheng *et al.*, 2020] between predicted boundaries $\hat{s}_t, \hat{e}_t$ and ground-truth s_t^*, e_t^*:

$$\mathcal{L}_{\text{reg}} = \frac{1}{T^+} \sum_{t \in \Omega^+} (1 - \text{DIoU}(\hat{s}_t, \hat{e}_t; s_t^*, e_t^*)) \qquad (7)$$

Final Loss. The overall objective combines both components:

$$\mathcal{L}_{\text{total}} = \frac{1}{T^+} \sum_{t=1}^{T} \left[\lambda \mathcal{L}_{\text{cls}} + \mathcal{I}(t \in \Omega^+) \mathcal{L}_{\text{reg}}(t) \right] \qquad (8)$$

where $\lambda \in [0, 1]$ is the balancing ratio between classification and localization tasks, and $\mathcal{I}(t \in \Omega^+)$ is an indicator function (1 if timestamp t lies in forged regions Ω^+, 0 otherwise).

4 Experiments

4.1 Implementation Details

Preprocessing. Audio data are resampled to 16 kHz and processed using WavLM-LARGE to extract 1024-dimensional feature vectors at 20 ms intervals (50 FPS). Video data undergoes frame extraction at 25 FPS, followed by resizing to 224×224 pixels and normalization. Per-frame feature extraction employs CLIP:ViT-B/16, generating 768-dimensional embeddings. All implementations utilize PyTorch on NVIDIA L40 GPUs.

Training. We adopt the AdamW optimizer with mini-batch processing, incorporating a 5-epoch warmup phase and cosine decay for learning rate scheduling. The initial learning rate is 1×10^{-3} and the weight decay is 1×10^{-2}. Variable-length sequences are standardized through padding/truncation to a maximum length of 1024, with masking mechanisms ensuring valid temporal context propagation. The number of R-TLM blocks N is set to 6, and the number of FPN levels L is set to 5. θ is set to 0.6 and the balancing ratio λ of the loss is set to 0.01. The model is trained for fixed epochs (up to 95) with a batch size of 64, and separate models are trained for audio and video modalities.

Inference. During inference, the complete sequence is input to the model with a batch size of 1. Non-Maximum Suppression (NMS) [Neubeck and Van Gool, 2006] is applied to refine predictions by eliminating highly overlapping and inefficient instances, yielding the final forged segment outputs. For unimodal temporal data, the maximum forgery confidence among all predicted segments is treated as the overall
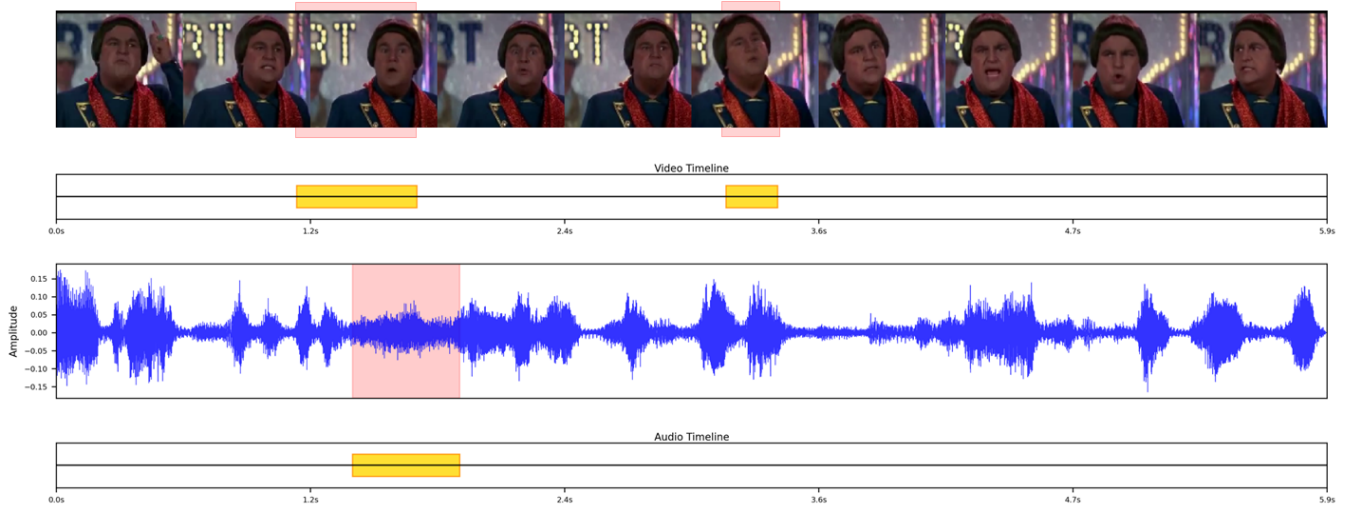
Figure 2: Visualization results of our proposed method. Red represents forged segments, and yellow represents our predicted results. Our method can accurately predict the presence of forged video and audio segments in the samples for both video and audio modalities.

Training strategy		Feature Embedding		Final Score
Initial Learning Rate	Epochs	Audio	Video	
1×10^{-3}	3	wavLM	CLIP	0.7535
1×10^{-3}	6	wavLM	CLIP	0.7501
1×10^{-3}	15	wavLM	CLIP	0.6590
1×10^{-3}	36	wavLM	CLIP	0.6174
1×10^{-3}	60	wavLM	CLIP	0.6144
1×10^{-3}	95	wavLM	CLIP	0.6000
1×10^{-3}	6	wav2vec	XCLIP	0.7361
1×10^{-3}	60	wav2vec	XCLIP	0.5644
1×10^{-3}	95	wavLM	XCLIP	0.5873
1×10^{-3}	95	wav2vec	XCLIP	0.5798
3×10^{-4}	2	wavLM	CLIP	0.5821
3×10^{-4}	5	wavLM	CLIP	0.7164
3×10^{-4}	6	wavLM	CLIP	0.7218
3×10^{-4}	7	wavLM	CLIP	0.7218
3×10^{-4}	8	wavLM	CLIP	0.7340
3×10^{-4}	9	wavLM	CLIP	0.7252
3×10^{-4}	10	wavLM	CLIP	0.7201
3×10^{-4}	11	wavLM	CLIP	0.7256
3×10^{-4}	12	wavLM	CLIP	0.7227
3×10^{-4}	13	wavLM	CLIP	0.7182
3×10^{-4}	14	wavLM	CLIP	0.7126
3×10^{-4}	15	wavLM	CLIP	0.7084

Table 1: Comparison experiments of our proposed method under varying experimental conditions, including self-supervised feature extractors, training epochs, and initial learning rates.

confidence score for the entire sequence. The final outputs are obtained by containing both audio and video modalities, where the higher confidence score between the two modalities is selected as the final sample-level forgery confidence, while the union of predicted audio and video forged segments forms the complete output set of forged regions.

Dataset. We evaluate our framework on the IJCAI'25 DDL-AV dataset [Miao *et al.*, 2025; Zhang *et al.*, 2024a; Zhang *et al.*, 2024b] for model training and testing. The dataset comprises 200k video samples in the training set, 20k in the validation set, and 111k in the test set. These samples cover 9 deepfake audio forgery methods and 18 deepfake video forgery techniques, ensuring comprehensive evaluation

across diverse generation paradigms.

Metrics. To comprehensively assess our method's performance across diverse scenarios, we adopt multi-dimensional metrics following the IJCAI'25 Workshop on Deepfake Detection and Localization dataset (DDL-AV dataset). Detection performance is quantified using Area Under Curve (AUC), while localization quality is measured via Average Precision (AP) and Average Recall (AR). The localization score is computed as:

$$\text{Score} = \frac{1}{16}\text{AP@0.5} + \frac{2}{16}\text{AP@0.75} + \frac{2}{16}\text{AP@0.9} \quad (9)$$

$$+ \frac{3}{16}\text{AP@0.95} + \frac{1}{16}\text{AR@30} + \frac{2}{16}\text{AR@20} \quad (10)$$

$$+ \frac{2}{16}\text{AR@10} + \frac{3}{16}\text{AR@5} \quad (11)$$

The final comprehensive performance metric is defined as:

$$\text{Final Score} = \frac{\text{AUC} + \text{Score}}{2} \quad (12)$$

This formulation balances detection accuracy and localization precision through weighted aggregation of threshold-specific metrics, ensuring robust evaluation of both global and fine-grained forgery identification capabilities.

4.2 Experimental Results

Visualization results. Figure 2 illustrates the visualization of our method's prediction results on multimodal input data. The visualizations demonstrate that our framework effectively detects and localizes forged segments in both video and audio modalities, with clear temporal alignment between predicted boundaries and ground-truth forgery regions. This empirical evidence validates the cross-modal consistency and precision of our detection mechanism.

Quantitative results. Table 1 presents a comprehensive performance comparison of our proposed framework under varying experimental conditions, including self-supervised feature extractors, training epochs, and initial learning rates. Our optimal configuration achieves a final score of **0.7535**, significantly outperforming baseline variants.

- **Self-supervised feature extractors.** To evaluate modality-specific feature representation capabilities, we contrast two audio feature extractors (Wav2Vec [Baevski *et al.*, 2020] vs. WavLM) and two video feature extractors (CLIP vs. XCLIP [Ni *et al.*, 2022]). While WavLM and CLIP consistently yield superior performance (Table 1), the marginal drops observed in Wav2Vec/XCLIP configurations can be attributed to: 1) Wav2Vec has lower model capacity (768-dimensional features vs. WavLM's 1024D) and reduced contextual modeling capability due to its shallow transformer architecture; 2) Temporal misalignment arising from XCLIP's original pretraining on 8 FPS videos, which conflicts with DDL-AV's 25 FPS format. Despite incorporating temporal encoding mechanisms, XCLIP struggles to adapt to higher frame rate dynamics, leading to suboptimal boundary localization; Notably, the performance gaps across extractors remain statistically insignificant, demonstrating the

robustness of our framework against variations in low-level feature representations. This suggests that our hierarchical FPN-Transformer architecture successfully abstracts modality-specific discrepancies into unified temporal feature spaces.

- **Training Depth and Learning Rate Analysis.** When training with 1×10^{-3} initial learning rate, deeper training (higher epochs) correlates with reduced generalization performance on DDL-AV's test set, which contains forgery methods absent in training. This manifests as declining scores beyond 3 epochs (optimal point). Reducing the learning rate to 3×10^{-4} reveals a non-monotonic generalization trend: performance improves initially but degrades at excessive depths. This reveals a critical trade-off between training depth and generalization robustness. Our findings suggest that balancing model capacity with controlled training duration may mitigate performance drops on out-of-distribution forgery samples, offering a novel perspective for improving detection systems' adaptability to evolving deepfake techniques.

5 Conclusion and Discussion

This paper presents a novel multi-modal deepfake detection and localization framework based on FPN-Transformer, addressing critical limitations in cross-modal generalization and temporal boundary regression. By leveraging pre-trained self-supervised models (WavLM for audio, CLIP for video) to extract hierarchical temporal features, combined with a multi-scale feature pyramid constructed via R-TLM blocks and localized attention mechanisms, our approach achieves robust analysis of cross-context temporal dependencies. The dual-branch prediction head enables simultaneous forgery probability estimation and precise temporal offset refinement, achieving frame-level localization accuracy. Evaluated on the IJCAI'25 DDL-AV benchmark, our method attains a final score of 0.7535, demonstrating superior performance in detecting and localizing sophisticated, fine-grained manipulations in challenging environments.

Our work advances the field by providing a unified solution for generalized deepfake detection, bridging the gap between unimodal approaches and complex real-world scenarios. Future research may explore dynamic adaptation to evolving generative techniques and further optimize computational efficiency for real-time applications.

Acknowledgments

References

[Alegre *et al.*, 2012] Federico Alegre, Ravichander Vipperla, and Nicholas Evans. Spoofing countermeasures for the

protection of automatic speaker recognition from attacks with artificial signals. *Proc. 13th Interspeech*, 29:54–58, 2012.

[Ba *et al.*, 2024] Zhongjie Ba, Qingyu Liu, Zhenguang Liu, Shuang Wu, Feng Lin, Li Lu, and Kui Ren. Exposing the deception: Uncovering more forgery clues for deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 719–728, 2024.

[Baevski *et al.*, 2020] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[Cao *et al.*, 2022] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4113–4122, 2022.

[Casanova *et al.*, 2022] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir 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.

[Chai *et al.*, 2020] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVI 16*, pages 103–120. Springer, 2020.

[Chen *et al.*, 2020a] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

[Chen *et al.*, 2020b] Tianxiang Chen, Avrosh Kumar, Parav Nagarsheth, Ganesh Sivaraman, and Elie Khoury. Generalization of audio deepfake detection. In *Odyssey*, pages 132–137, 2020.

[Chen *et al.*, 2022] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

[Choi *et al.*, 2024] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake detection video detection. *arXiv e-prints*, pages arXiv–2403, 2024.

[Ding *et al.*, 2020] Xinyi Ding, Zohreh Raziei, Eric C Larson, Eli V Olinick, Paul Krueger, and Michael Hahsler. Swapped face detection using deep learning and subjective assessment. *EURASIP Journal on Information Security*, 2020:1–12, 2020.

[Hu *et al.*, 2023] Chen Hu, Xianghua Xie, and Lin Wu. Face reenactment via generative landmark guidance. *Image and Vision Computing*, 130:104611, 2023.

[Hunt and Black, 1996] Andrew J Hunt and Alan W Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, volume 1, pages 373–376. IEEE, 1996.

[Ji *et al.*, 2024] Zhoulin Ji, Chenhao Lin, Hang Wang, and Chao Shen. Speech-forensics: towards comprehensive synthetic speech dataset establishment and analysis. *arXiv preprint arXiv:2412.09032*, 2024.

[Kim *et al.*, 2021] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

[Kong *et al.*, 2020] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[Li *et al.*, 2018] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking. *arXiv preprint arXiv:1806.02877*, 2018.

[Li *et al.*, 2023] Shuai Li, Liang Liu, Ji Liu, Wenfeng Song, Aimin Hao, and Hong Qin. Sc-gan: Subspace clustering based gan for automatic expression manipulation. *Pattern Recognition*, 134:109072, 2023.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Liu *et al.*, 2020] Yaoyao Liu, Qianru Sun, Xiangnan He, An-An Liu, Yuting Su, and Tat-Seng Chua. Generating face images with attributes for free. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2733–2743, 2020.

[Malik and Farid, 2010] Hafiz Malik and Hany Farid. Audio forensics from acoustic reverberation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1710–1713. IEEE, 2010.

[Manjula *et al.*, 2022] AK Manjula, R Thirukkumaran, K Hrithik Raj, Ashwin Athappan, and R Paramesha Reddy. Deep fakes image animation using generative adversarial networks. In *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, pages 1–6. IEEE, 2022.

[Miao *et al.*, 2022] Changtao Miao, Zichang Tan, Qi Chu, Nenghai Yu, and Guodong Guo. Hierarchical frequency-assisted interactive networks for face manipulation detection. *IEEE Transactions on Information Forensics and Security*, 17:3008–3021, 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.

[Müller *et al.*, 2022] Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

[Neubeck and Van Gool, 2006] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, volume 3, pages 850–855. IEEE, 2006.

[Neves *et al.*, 2020] Joao C Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez. Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1038–1048, 2020.

[Ni *et al.*, 2022] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European conference on computer vision*, pages 1–18. Springer, 2022.

[Nirkin *et al.*, 2021] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6111–6121, 2021.

[Ojha *et al.*, 2023] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023.

[O'Shaughnessy, 1988] Douglas O'Shaughnessy. Linear predictive coding. *IEEE potentials*, 7(1):29–32, 1988.

[Peng *et al.*, 2024] Chunlei Peng, Zimin Miao, Decheng Liu, Nannan Wang, Ruimin Hu, and Xinbo Gao. Where deepfakes gaze at? spatial-temporal gaze inconsistency analysis for video face forgery detection. *IEEE Transactions on Information Forensics and Security*, 2024.

[Qian *et al.*, 2020] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Styger and Keller, 1994] Thomas Styger and Eric Keller. Formant synthesis. *Fundamentals of speech synthesis and speech recognition: Basic concepts, state of the art, and future challenges*, pages 109–128, 1994.

[Sun *et al.*, 2021] Guangzhi Sun, Chao Zhang, and Philip C Woodland. Transformer language models with lstm-based cross-utterance information representation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7363–7367. IEEE, 2021.

[Tan *et al.*, 2023] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023.

[Tan *et al.*, 2024] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5052–5060, 2024.

[Wang *et al.*, 2020] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020.

[Wang *et al.*, 2023] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

[Xie *et al.*, 2023] Yuankun Xie, Haonan Cheng, Yutian Wang, and Long Ye. Single domain generalization for audio deepfake detection. In *DADA@ IJCAI*, pages 58–63, 2023.

[Yang *et al.*, 2019] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[Yang *et al.*, 2021] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin,

Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

[Yi *et al.*, 2022] Jiangyan Yi, Ruibo Fu, Jianhua Tao, Shuai Nie, Haoxin Ma, Chenglong Wang, Tao Wang, Zhengkun Tian, Ye Bai, Cunhang Fan, et al. Add 2022: the first audio deep synthesis detection challenge. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9216–9220. IEEE, 2022.

[Yin *et al.*, 2023] Qilin Yin, Wei Lu, Bin Li, and Jiwu Huang. Dynamic difference learning with spatio–temporal correlation for deepfake video detection. *IEEE Transactions on Information Forensics and Security*, 18:4046–4058, 2023.

[Zendran and Rusiecki, 2021] Michał Zendran and Andrzej Rusiecki. Swapping face images with generative neural networks for deepfake technology–experimental study. *Procedia computer science*, 192:834–843, 2021.

[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.

[Zheng *et al.*, 2020] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.

[Zheng *et al.*, 2024] Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, and Chao Shen. Breaking semantic artifacts for generalized ai-generated image detection. *Advances in Neural Information Processing Systems*, 37:59570–59596, 2024.