Deepfake Detection: A Multimodal Survey

Meng Wang

School of Intelligent Science and Technology, Geely University of China, Chengdu, China

Abstract. The rapid advancement of generative artificial intelligence has catalyzed the emergence of deepfake technologies capable of cross-modal data fusion, posing systemic threats to digital security. To address these challenges, the academic community has developed multidimensional detection frameworks that integrate three core components: spatiotemporal consistency verification, cross-modal feature alignment, and semantic correlation inference. By synergistically processing multimodal data streams—including video, audio, and text—these frameworks leverage the complementarity and contradictions inherent in cross-modal features to identify forgery artifacts, substantially enhancing detection efficacy for sophisticated synthetic content. This study systematically examines the algorithmic architectures underpinning multimodal detection technologies, with focused analysis on optimized feature fusion strategies, innovative dynamic temporal modeling approaches, and cutting-edge adversarial training mechanisms. It further explores their application potential in critical scenarios such as political communication authentication and judicial digital forensics. The research confirms the paradigm's unique advantages in countering complex forgery attacks, establishing scalable technical pathways for developing intelligent defense systems against advanced deepfake threats.

1 Introduction

Generative AI has emerged as a key innovation driver, propelled by advances in deep learning architectures and computational capacity. Modern synthesis technologies now integrate multimodal heterogeneous data streams, evolving from early single-modality deepfake systems to architecturally complex frameworks with multidimensional generative capabilities.

The exponential advancement of generative AI through cross-modal integration has systematically amplified the stealth and operational efficiency of synthetic disinformation production, introducing novel vulnerabilities across three critical domains: information authentication protocols, personal data sovereignty preservation, and sociopolitical discourse equilibrium.

Deepfake, a lexical blend of "deep learning" and "fake," denotes artificial intelligence systems that utilize deep neural networks to simulate and forge audiovisual content. This algorithmic technology synthesizes hyper-realistic synthetic media, including images, videos, and audio recordings, through advanced computational pattern replication.

morgana3366@outlook.com

Deepfake's defining characteristics center on three technical pillars: perceptual realism achieving visual indistinguishability from authentic recordings, operational accessibility through open-source toolkits enabling non-experts to produce synthetic media with consumer-grade hardware, and inherent dual-use potential. While facilitating legitimate applications in visual effects and historical preservation, the technology's pattern replication capabilities are systematically exploited for disinformation campaigns, identity fraud architectures, and political discourse manipulation through algorithmically generated media.

The substantial societal risks posed by deepfake technologies have prompted comprehensive countermeasures across institutional domains. To address threats to political security stemming from synthetic media targeting public figures, national governments are actively developing regulatory frameworks and technical standards to govern synthetic content production and dissemination.

Notable audiovisual synthetic media cases demonstrate advanced manipulation techniques. The forged Obama speech leveraged Wav2Lip's audiovisual synchronization to align authentic vocal recordings with synthetically generated lip movements, producing hyper-realistic political discourse fabrication. Concurrently, the Zuckerberg case employed algorithmic text synthesis to fabricate declarative statements attributed to the public figure, exemplifying multimodal disinformation architectures that integrate linguistic and visual forgery modalities.

Multimodal Deepfake detection denotes methodologies for coordinated identification of synthetic media through simultaneous analysis of video, audio, and textual data streams. Unlike conventional unimodal approaches that isolate textual or visual analysis, this paradigm exploits cross-modal feature interdependencies to enhance trace detection precision. The academic community has developed diverse countermeasure techniques targeting synthetic content generation, with this study focusing on methodological frameworks for multimodal integration. The paper systematically categorize and analyze technical architectures for combined data-form detection, emphasizing optimized strategies for multimodal forensic analysis.

In the digital information age, multimodal data processing and analysis has emerged as a critical research focus across scientific disciplines. Human cognitive processing paradigms inherently reject isolated sensory input analysis, instead integrating concurrent multisensory data streams during environmental perception. This biological perceptual mechanism has motivated the scientific community to pioneer advanced multimodal feature analysis frameworks.

Multimodal data encompasses rich and diverse forms of information. At its core, it includes primary modalities: visual modalities such as images and videos directly depict the appearance, scenes, and dynamic changes of objects; audio modalities capture auditory information like speech and ambient sounds; while textual modalities convey precise semantic content through linguistic representation. As technology evolves and research deepens, extended modalities have increasingly gained attention. Physiological signals, including heart rate and eye movement patterns, serve as indicators of an individual's internal physiological state. Behavioral data, such as keyboard input sequences and mouse movement trajectories, reveal users' interaction patterns during human-computer engagement. These modalities collectively form a comprehensive framework for understanding complex real-world phenomena from multiple dimensions.

In practical applications, multimodal data exhibit strong interdependencies. For example, temporal synchronization ensures audio-visual coherence in multimedia analysis, while semantic consistency aligns retrieved content with query intent in image-text tasks, improving retrieval accuracy. Such intrinsic correlations enable advanced applications. Advancing multimodal detection technologies can thus drive intelligent, secure, and efficient human-computer interaction frameworks.

2 Methodology

2.1 Research on Correlations in Multimodal Data and Detection Techniques

Multimodal DeepFakes generation employs AI-driven synthesis/manipulation of cross-modal content (visual, auditory, textual), relying on cross-modal consistency and semantic alignment to produce realistic, contextually coherent outputs. Core mechanisms include:

2.1.1 Cross-modal Generation Technology

Cross-modal generation technologies, including text-to-multimodal generation, use cross-modal models like CLIP to provide semantic associations between text and images and employ language models to parse text descriptions for generating corresponding visuals or audio.

Figure 1. Contrastive pre-training [1]

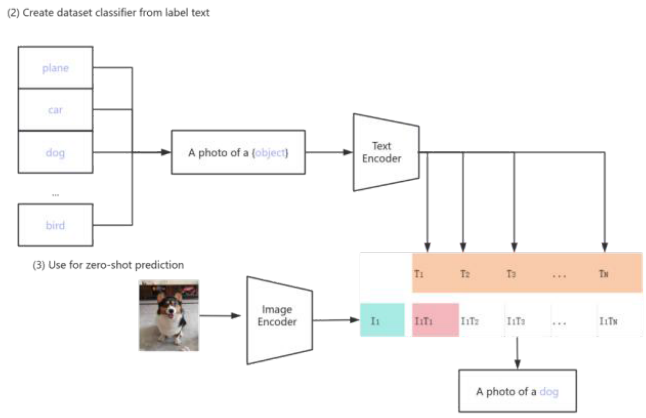

Figure 2. Create dataset classifier from label text [1]

In standard vision models, an image encoder and linear classifier are jointly trained to predict labels. CLIP instead aligns an image encoder with a text encoder through contrastive learning on matched image-text pairs.

As shown in Figure 1: Illustrates contrastive pre - training. An image is processed by an Image Encoder to generate embeddings, and text by a Text Encoder. Their embeddings form a correlation matrix.

As shown in Figure 2: Creates a dataset classifier from label text, then conducts zero - shot prediction. An image's embedding (via Image Encoder) matches with text embeddings to yield a result such as "A photo of a dog".

Or it could be audio - driven visual generation, that is, using Wav2Lip to achieve the synchronous generation of lip shape and audio, as well as AD - NeRF combined with neural radiation scenarios to generate audio - driven high - fidelity 3D face videos. Drive facial animations such as lip shapes and expressions through acoustic features.

2.1.2 Inter-modal Synchronous Control Technology

Inter-modal synchronous control technology primarily functions through temporal alignment and semantic consistency enhancement. Temporal alignment employs dynamic time warping or LSTM networks to synchronize audio-video frames, while semantic consistency leverages contrastive learning to align generated content with input modalities. For example, textual prompts like "smiling" require corresponding facial actions, as implemented in DeepFaceLab's Expression Transfer module, which maps source video expressions to target audio emotions.

2.2 Multimodal Diffusion Model Technology

Deepfake detection technology, particularly driven by diffusion models, has advanced substantially by leveraging their unique stepwise noise diffusion and reverse denoising mechanisms. These models excel in extracting latent data feature distributions, demonstrating superior performance in identifying synthetic forgeries.

In the frontier research of multimodal generation technology, diffusion models exhibit unique technical principles and remarkable performance advantages. Their core technical framework encompasses three key dimensions: diffusion process modeling, which simulates the gradual diffusion of noise and reverse denoising dynamics; cross-modal joint training, which establishes semantic correspondence across visual, auditory, and textual modalities; and conditional control, which enables guided generation through input constraints such as text descriptions or audio features. These principles collectively underpin the models' capability to learn complex data distributions and generate high-fidelity multimodal content.

Diffusion process modeling is rooted in Markov chain theory, achieving the transformation from random noise to target multimodal data through an iterative gradual denoising process. During this process, the model progressively structures and semantizes initial random noise based on its diffusion coefficients and denoising strategies. Taking the generation of video and audio from text descriptions as an example, noise is refined across spatiotemporal dimensions, enabling each video frame and corresponding audio signal to precisely match the semantic information of the text description—simulating the natural law of evolution from disorder to order in data generation. The following figure shows a real image and four images generated by a diffusion model.

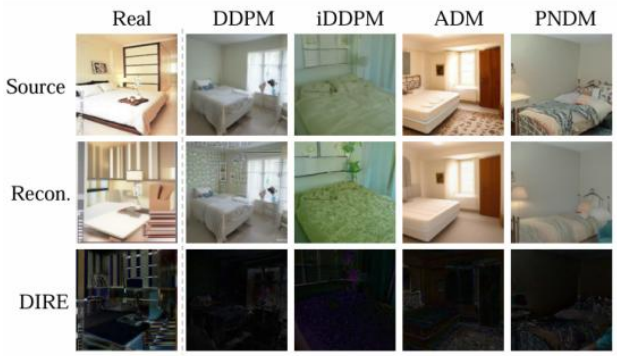

Figure 3. Comparison of images generated under different categories [2]

Figure 3 shows the comparison of real images (Real) with images generated by diffusion models like DDPM, iDDPM, ADM, and PNDM under different categories of Source, Recon., and DIRE [2].

Denoising Diffusion Probability Model (DDPM) Inspired by non-equilibrium thermodynamics, the Diffusion model [3] was proposed for the first time and achieved strong performance in image generation. They defined the Markov chain of diffusion steps [2], which slowly add Gaussian noise to the data until it degenerates into an isotropic Gaussian distribution (forward process), and then learn to reverse the diffusion process to generate samples from the noise (reverse process). The Markov chain in the forward process is defined as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\frac{\alpha_t}{\alpha_{t-1}}} x_{t-1}, (1 - \frac{\alpha_t}{\alpha_{t-1}})I) \quad (1)$$

xt is the image noise in t-step, $\alpha 1 ... \alpha T$ is a predefined plan, and T represents the total number of steps.

Cross-modal joint training aims to achieve deep alignment and fusion of multimodal features by constructing a unified latent space. A typical example is the spatio-temporal patch coding technology adopted by the Sora model, which projects data from different modalities into the same latent space. Within this space, modal features efficiently interact via attention mechanisms and nonlinear transformations of deep neural networks, enabling the model to learn cross-modal correlations and complementarity. This process generates more natural and semantically coherent multimodal content.

Conditional control mechanisms leverage pretrained models like CLIP to achieve semantic alignment between generated content and input modalities. Through large-scale contrastive learning on image-text pairs, CLIP establishes cross-modal semantic mappings that encode textual prompts into generation-guiding signals. The generation model dynamically regulates parameter update trajectories based on these signals, enforcing tripartite constraints on semantic relevance, stylistic coherence, and structural integrity throughout the synthesis process, thereby effectively mitigating semantic drift between synthesized outputs and original inputs.

Stable Diffusion 3 (SD3) enhances image/video generation quality through latent space diffusion modeling, utilizing text-guided denoising processes in low-dimensional manifolds to synthesize semantically consistent visual content for creative design applications, while Sora (OpenAI) pioneers a unified spatiotemporal diffusion framework that directly generates synchronized video-audio-text outputs by integrating dynamic interframe coherence mechanisms, audiovisual rhythm-emotion alignment modules, and precision semantic anchoring via joint latent space mapping, achieving high-fidelity multimodal

synthesis through spatiotemporal diffusion dynamics that optimize both temporal continuity and cross-modal consistency [3].

2.3 The Core Technical Framework and Background of Multimodal Detection Technology

2.3.1 Visual Modality

Convolutional Neural Network (CNN) extracts local features via convolutional kernels, reduces dimensionality for low-level feature representation through pooling layers, and implements classification via fully connected layers. In image forgery detection, it can identify anomalies like facial distortion or inconsistent lighting (as shown in Figure 4). However, it struggles to model long-range dependencies, such as global semantic coherence.

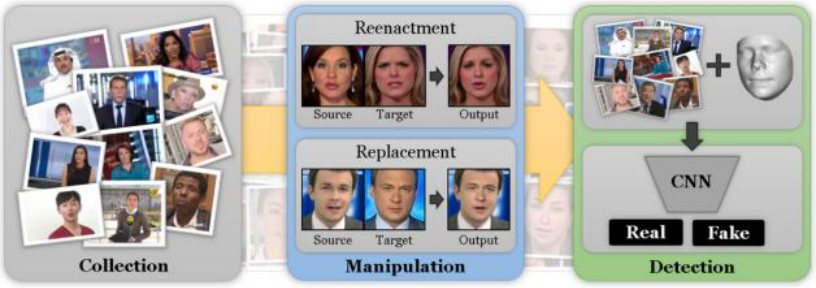

Figure 4. The process of detecting authenticity through CNN [4]

Face Forensics constitutes a benchmark facial forgery dataset that enables supervised training of deep learning-based detection frameworks. This corpus comprehensively covers four state-of-the-art manipulation methodologies— Face2Face, FaceSwap, DeepFakes, and NeuralTextures— representing seminal work in facial forgery generation [4].

The Vision Transformer (ViT) achieves superior cross-modal alignment performance by partitioning images into patch sequences and employing self-attention mechanisms to capture global contextual relationships across visual data elements.

2.3.2 Audio Mode

MFCC (Mel Frequency Cepstral Coefficient) simulates human auditory perception to extract spectral features, emphasizing high-frequency and timbre attributes, and serves as a critical tool for detecting spectral anomalies in AI synthetic speech. The advancement of text-to-speech (TTS) and voice conversion (VC) technologies has amplified security risks from hyper-realistic synthetic speech. To address this, phoneme-level feature analysis frameworks have emerged, identifying synthetic speech through statistical inconsistencies in phoneme sequences. Adaptive Phone Pooling dynamically extracts phoneme-specific features from frame-level signals, enhancing detection of phoneme-level anomalies. Experimental results demonstrate that integrating this technique with pre-trained audio models significantly improves cross-dataset robustness, outperforming traditional methods and advancing synthetic speech defense systems. Figure 5 shows T-SNE visualization of clustering patterns.

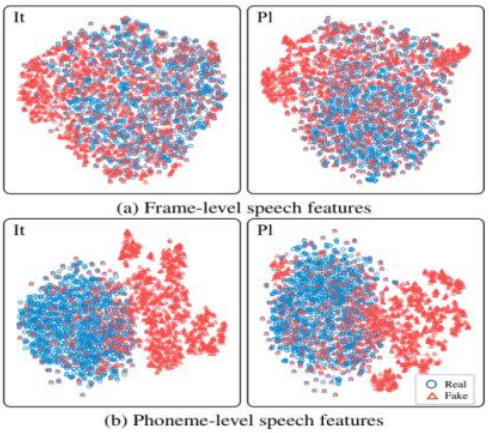

Figure 5. T-SNE Visualization of Clustering Patterns [5]

The feature extraction of the spectrogram is achieved by converting the audio signal into a time-frequency graph and inputting it as an image into a CNN for processing.

2.3.3 Text Modality

Language Models such as BERT and GPT: BERT leverages Masked Language Modeling (MLM) to learn contextual semantic associations, generating deep text embeddings that detect logical inconsistencies in false-news detection [6]. To counter multimodal disinformation (text/images/videos), researchers have developed a fusion framework combining BERT-derived text semantics, VGG19-learned visual features, and Scale-Dot Product Attention for cross-modal alignment. This approach achieves 3.1% higher accuracy on Twitter datasets than SOTA methods, with cross-domain validation confirming strong generalization for unlabeled data [6].

GPT employs autoregressive generation to model long-text coherence. When co-trained with CLIP, it aligns text-image encoder parameters, enabling natural language-driven image retrieval [7].

2.3.4 Temporal Feature Processing (for Video/Audio): RNN, LSTM, 3D-CNN:

RNN/LSTM models model temporal dependencies through a recurrent structure, where LSTM alleviates the vanishing gradient problem via its gating mechanism (input, forget, output gates). 3D-CNN extracts spatio-temporal features by incorporating a temporal dimension alongside spatial dimensions (width, height), enabling analysis of sequential visual data [8].

2.4 Multimodal Fusion Strategy

2.4.1 Cross-modal Alignment Based on Attention Mechanism

This strategy dynamically adjusts modal feature contributions via attention weights to capture fine-grained cross-modal correlations. The Cross-Modal Transformer embeds cross-modal attention layers within Transformer architectures to integrate text, visual, and auditory modalities. Key challenges include: (1) Data Misalignment (e.g., facial expressions lagging speech due to sampling rate disparities) and (2) Long-Term

Dependency Modeling (capturing temporal cross-modal correlations). The model synthesizes verbal expressions, facial features, and prosodic intonation for precise behavioral analysis, despite increased complexity from audio-visual frequency mismatches during fusion [9].

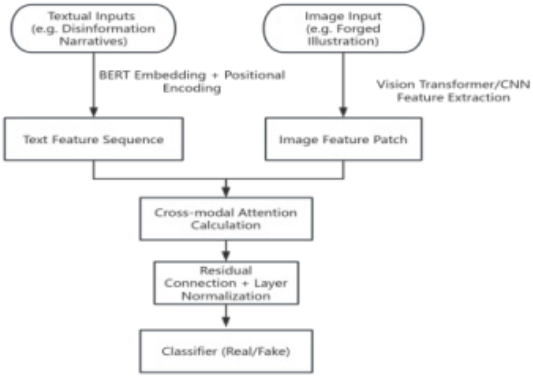

Figure 6. Cross-Modal Attention Architecture in Cross-Modal Transformer (Picture credit: Original)

Figure 6 shows the Cross-Modal Attention Architecture. CLIP aligns text and image embedding spaces via contrastive learning and enables zero-shot cross-modal retrieval. Traditional computer vision systems exhibit limited universality due to training on fixed preset categories, requiring additional data labeling and retraining to recognize unpresented concepts. This limitation can be overcome through text supervision, with image-text matching pretraining serving as an efficient and scalable approach. Trained on a large-scale dataset of 400 million internet image-text pairs, the model learns state-of-the-art (SOTA) image representations from scratch. Pretrained models can reference learned and brand-new visual concepts via natural language, enabling zero-shot transfer to diverse downstream tasks without task-specific retraining, thereby significantly expanding application scenarios and practical utility.

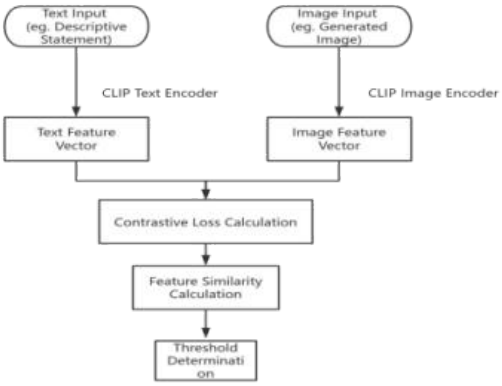

Figure 7. CLIP-Guided Contrastive Learning Flowchart (Picture credit: Original)

As shown in Figure 7, the above model application analyzes text-image consistency for false news detection (e.g., a "fire scene" text description conflicting with an image lacking smoke) and enables medical image diagnosis by combining CT images with pathological

reports to generate diagnostic conclusions. Current visual-text contrastive learning models like CLIP focus on matching paired images with text embeddings while distinguishing embeddings of unmatched pairs to enhance feature representation transferability and support zero-shot prediction.

2.4.2 Multimodal Correlation Modeling of Graph Neural Network (GNN)

Graph neural networks achieve information aggregation by using multimodal features as nodes and inter-modal relationships as edges, with graph convolution (e.g., MMGFN constructs multimodal heterogeneous graphs to capture text-image-user node associations in social rumor propagation). Dialogue Emotion Recognition (ERC), a key task in emotional dialogue systems, requires understanding user emotions to generate empathetic responses. Current research is mostly limited to text modality speaker/context information or simple feature concatenation of limited multimodal data. To address this, the MMGCN model employs a multimodal fusion graph convolutional network to efficiently leverage multimodal and long-context information. It mines inter-modal dependencies and models speaker interactions and individual characteristics via speaker information, offering an innovative emotion recognition solution and enhancing emotional dialogue system performance in complex scenarios [10]. Figure 8 shows the GNN multimodal key modeling flowchart.

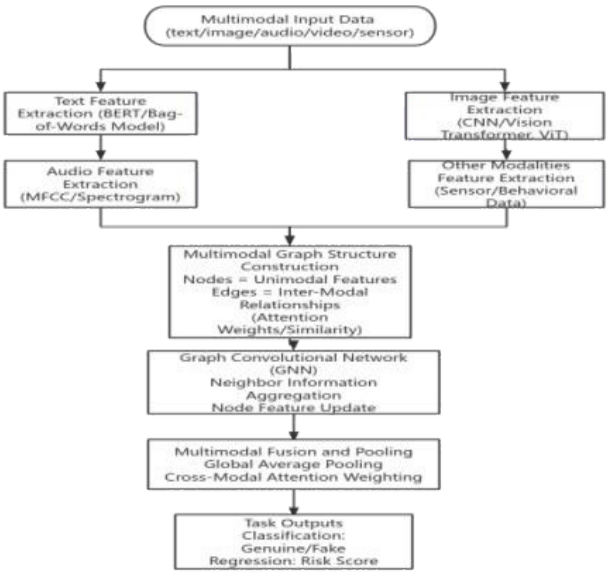

Figure 8. Graph Neural Network (GNN) Multimodal Key Modeling Flowchart (Picture credit: Original)

2.4.3 Cross-modal Representation Learning Based on Contrastive Learning

In the field of cross-modal representation learning, contrastive learning constructs a cross-modal shared semantic space by maximizing the semantic similarity of positive sample pairs and minimizing the feature differences of negative sample pairs; the core of this method lies in using the contrastive loss function to drive the model to learn cross-modal general semantic representations, break modal barriers, and promote efficient information fusion. Representative models include CLIP, which maps images and texts to a unified

semantic space through contrastive learning to achieve cross-modal data comparability and lay a technical foundation for image-text alignment tasks [2], and ALIGN, which performs contrastive training based on large-scale noisy image-text pairs, innovatively utilizes unstructured text to supervise data, significantly enhances the robustness of cross-modal retrieval against noise interference, and demonstrates excellent adaptability in complex scenarios [11].

Natural language processing has shifted from manual annotation to unlabeled text training via self-supervised learning, but vision and visual-language representation still depend on costly professionally annotated datasets, limiting research scale and progress. Recent studies have overcome this bottleneck: using the Conceptual Captions dataset (over one billion noisy image-text pairs), a simple dual-encoder architecture with contrastive loss achieves effective visual-language alignment without complex data filtering.

In video content review, contrastive learning enables Deepfake detection by modeling normal cross-modal correlation patterns. For example, in forged speech videos, the model learns audio-visual modality synchronization and uses lip-sound misaligned samples as negative pairs for contrastive training, thereby identifying audio-visual inconsistencies and providing semantic-level support for video authenticity detection.

2.4.4 Tensor Fusion

The Tensor Fusion Network (TFN), a representative architecture in this domain, employs tensor decomposition theory—specifically Tucker decomposition—to address computational efficiency by performing low-rank approximation on high-order fusion tensors. This approach retains cross-modal interaction key information while reducing exponential computational complexity to polynomial order. The optimization strategy ensures both model engineering feasibility and theoretical support for real-time multimodal data processing, facilitating its deployment in large-scale application scenarios.

Multimodal sentiment analysis overcomes the limitations of traditional monolingual sentiment analysis by integrating multimodal information in multilingual contexts. To model the dynamic intra- and inter-modal relationships required by this task, the tensor fusion network was developed. It enables end-to-end learning of these relationships and is optimized for processing complex multimodal data, such as oral expressions, gestures, and voices in online videos, to meet the demands of intricate multimodal interactions [12].

On the publicly available CMU-MOSI multimodal corpus, this method has achieved state-of-the-art emotion classification and regression results, demonstrating strong effectiveness. Qualitative analysis of the model highlights that the temporal attention layer plays a critical role in emotion prediction despite noise in acoustic and visual modalities. Further analysis confirmed that gated multimodal embedding is equally effective in selectively filtering such noisy modalities.

2.4.5 Lightweight Fusion Based on Knowledge Distillation

In cross-modal fusion research, knowledge distillation technology offers new insights for lightweight model deployment: 1. Distill-M3:employs hierarchical distillation to transfer cross-modal attention and semantic alignment knowledge from multimodal BERT to a lightweight single-modal network. By leveraging collaborative feature decoupling and projection from the teacher model, the student model reconstructs cross-modal reasoning using single-modal inputs, significantly reducing parameter count. 2. TinyCLIP：integrates dynamic channel pruning with quantized-awareness distillation to compress CLIP’s cross-modal matching capabilities into mobile architectures. Through two-stage optimization of modal comparison distillation and task distillation, it enables real-time content review on

mobile devices. 3. The MIND Framework: addresses scarce multimodal data scenarios (e.g., healthcare) by transferring knowledge from multi-source single-modal pre-trained teacher models to small multimodal student models, enhancing fusion performance under low-data conditions via diversified representation distillation. These approaches balance model efficiency with cross-modal task accuracy through teacher-student collaborative learning, advancing the implementation of lightweight multimodal technologies. Figure 9 shows knowledge distillation-based lightweight fusion framework.

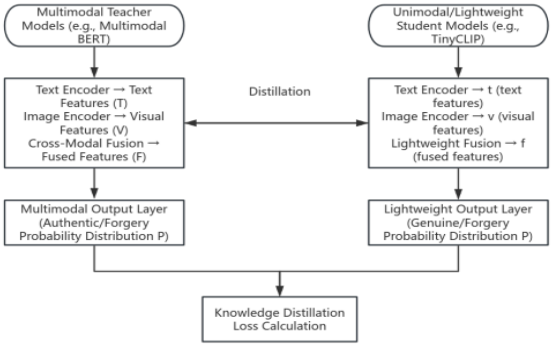

Figure 9. Knowledge Distillation-Based Lightweight Fusion Framework (Picture credit: Original)

2.5 Comparison of Multimodal Detection Datasets, Evaluation Criteria and Single-modal Defects

2.5.1 The Limitations of Single-modal Detection

Single-modal visual analysis has significant limitations: relying solely on visual information makes it susceptible to local forgeries (e.g., preserved real backgrounds in face-swapping videos) or physical inconsistencies in GAN-generated images (e.g., conflicting lighting directions), undermining detection reliability. Testing on FaceForensics++ shows that the single-modal ResNet model achieved an AUC of 0.76 for face-swapping detection, while a multimodal fusion model improved this to 0.93 by integrating multi-source information, demonstrating the effectiveness of multimodal fusion in compensating for the limitations of the visual modality [4]. The audio modality is notably affected by environmental noise and adversarial attacks: background noise often masks spectral anomalies in synthetic speech (e.g., MFCC feature failures), and adversarial noise can bypass voiceprint liveness detection. The WaveFake experiment indicates that the detection accuracy of the single-modal LSTM model drops by up to 35% under noise interference, highlighting its insufficient robustness in complex scenarios [13]. Text modalities face bottlenecks in semantic parsing and cross-lingual scenarios: AI-generated paraphrased texts (e.g., fake news produced by GPT-3) can evade traditional keyword-based detection, while cross-lingual adversarial attacks (e.g., semantic distortions in Chinese-English translation) significantly reduce detection accuracy. Grover tests reveal that the single-modal BERT model achieves an F1-Score of only 0.62 when detecting paraphrased fake news, reflecting its challenges in balancing precision and recall and revealing the text modality's limitations in handling complex semantics [14].

2.5.2 Mainstream Multimodal Datasets

Multimedia forgery detection faces inherent limitations in individual modalities: visual analysis struggles with local face-swapping artifacts (e.g., single-modal ResNet achieving an AUC of 0.76 on FaceForensics++) and physical inconsistencies in GAN-generated content (e.g., lighting mismatches), while audio detection is hindered by environmental noise masking spectral anomalies (e.g., LSTM accuracy dropping 35% under WaveFake noise) and adversarial voiceprint attacks. Text detection confronts challenges from semantic rewriting (e.g., BERT achieving an F1-Score of 0.62 in Grover's fake news detection) and cross-lingual adversarial distortions. Addressing these gaps, FakeAVCeleb—the first audio-visual synchronous forgery dataset—integrates 20,000 labeled videos with audio generated via Wav2Lip and AD-NeRF, enabling systematic analysis of cross-modal inconsistencies (e.g., lip-sync errors, audio-visual semantic mismatches). As a critical benchmark for multimodal detection, it facilitates the development of collaborative models that leverage complementary visual-audio features, advancing research into coherent cross-modal representation learning and robustness against synthetic media manipulations [15]. Table 1 shows the detection results of different methods.

Table 1. Quantitative Deepfake detection using the FakeAVCeleb dataset

Dataset	Real Videos	Fake Videos	Total Videos	Rights Cleared	Agreeing subjects	Total Subjects	Synthesis Methods	Real Audio	Deepfake Audio	Fine-grained labeling
UADFV [16]	49	49	98	No	0	1	1	No	No	No
DeepfakeTIMIT [17]	640	320	960	No	0	2	2	No	Yes	No
FF++ [18]	1000	4000	5000	No	0	4	4	No	No	No
Celeb-DF [19]	590	5639	6229	No	0	1	1	No	No	No
GoogleDFD [20]	363	3000	3363	Yes	28	5	5	No	No	No

Multimedia forgery detection faces inherent limitations across visual, audio, and text modalities: in visual analysis, single-modal convolutional neural networks (e.g., ResNet) struggle to detect local forgeries (e.g., face-swapping with real backgrounds) and physical inconsistencies in GAN-generated content (e.g., illumination reflection violations), achieving an AUC of only 0.76 for local face-swapping detection on

FaceForensics++significantly lower than the 0.93 AUC of multimodal fusion models. Text modalities suffer from semantic ambiguity, as AI-generated text evades keyword detection via rewriting and cross-lingual attacks degrade non-native detection accuracy; for example, single-modal BERT achieves an F1-Score of just 0.62 in detecting rewritten fake news in Grover model tests [21].

The Korean Deepfake Dataset (KoDF) is a large-scale multimodal dataset designed for deepfake detection, integrating video, audio, and multilingual text (including 10,000 Korean/English bilingual forged news videos). Specializing in Asian facial features and cross-lingual attack scenarios (e.g., Korean speech with English subtitles), it addresses the shortage of cross-cultural forgery data in prior datasets and enhances detection complexity by introducing cross-lingual inconsistencies, providing critical cases for researching cross-lingual information consistency and multimodal fusion in deepfake detection [22].

Multimodal Face News (MM-Fake): The Multimodal Face News (MM-Fake) is a pivotal dataset dedicated to multimodal fake news research, distinguished by its modal diversity and large-scale data composition. Integrating text, image, and social context modalities, the dataset comprises 50,000 fake news instances with detailed annotations of semantic conflicts, covering scenarios such as misleading illustrations and tampered screenshots. These rich data resources and granular annotations provide robust support for in-depth analysis of multimodal semantic inconsistencies in fake news. MM-Fake's unique strengths lie in its simulation of social media dissemination chains—including forwarding and commenting behaviors—to highly replicate the spread pathways and real-world dissemination environments of fake news on social platforms, enabling researchers to conduct studies within authentic contextual frameworks. Additionally, the dataset supports multimodal association detection, facilitating exploration of latent correlations among text, images, and social contexts to enable more comprehensive and accurate fake news identification [23].

2.6 Multimodal Detection and Evaluation Criteria

2.6.1 Core Performance Indicators

In multimodal data processing and forgery detection, core performance indicators include two primary dimensions: cross-modal accuracy and inter-modal consistency. Cross-modal accuracy is commonly measured using metrics such as AUC-ROC (e.g., assessing a model's overall ability to distinguish real from forged data in the DFDC competition) and F1-Score (e.g., balancing precision and recall in financial risk control scenarios). Inter-modal consistency is quantified via lip-syncing error (LSE) to detect audiovisual mismatches in face-swapped videos or through CLIP semantic similarity to evaluate text-image semantic alignment (e.g., identifying textual-visual contradictions in fake news). These metrics facilitate multi-faceted performance evaluation and fine-grained model optimization [24].

2.6.2 Robustness Test

In the performance evaluation of systems like multimodal forgery detection models, robustness testing is critical, with noise interference testing serving as a key component. By introducing Gaussian noise to data, this testing simulates random interference factors in real-world applications, assessing whether the model can maintain accurate detection under noisy conditions. Adversarial attack tests focus on evaluating a model's defensive capabilities through generating adversarial samples via methods such as Projected Gradient

Descent (PGD) [20] and FGSM (Fast Gradient Sign Method), which are designed to mislead the model into erroneous judgments. FGSM operates as a simple one-step scheme to maximize the interior of the saddle point formula, while the more powerful multi-step variant PGD performs projected gradient descent on a negative loss function: this iterative process optimizes a min-max problem at the loss function's saddle point, generating robust adversarial examples by minimizing model accuracy while maximizing perturbation effectiveness [25].

$$x + \varepsilon \, sgn(\nabla_x L(\theta, x, y)). \quad (2)$$

FGSM is an attack against bounded attackers. The adversarial example is calculated as:

$$x^{t+1} = \Pi_{x+s}(x^t + \alpha \, sgn(\nabla_x L(\theta, x, y))). \quad (3)$$

Feeding adversarial samples into models for testing enables researchers to evaluate performance under targeted attacks, uncover weaknesses and vulnerabilities, and inform improvements to model robustness and security. Ultimately, this process ensures stable operation in real-world environments characterized by complex, dynamic, and potentially malicious attacks.

2.6.3 Computational Efficiency

In multimodal data processing and detection models, computational efficiency is primarily measured by inference delay and FLOPs: real-time applications like video forgery detection require inference delays ⩽150ms (in line with standards such as the Microsoft Azure API) to ensure rapid response, while edge device deployment necessitates models maintaining computational complexity (FLOPs) under 5G to adapt to resource-constrained environments, reduce energy consumption, and improve operational efficiency. These metrics guide the balance between model performance and resource usage, enabling practical implementations in real-time detection, edge computing, and similar scenarios [1].

3 Results

3.1 Challenges and Frontier Directions of Multimodal Detection

3.1.1 Information Redundancy and Conflict among Modalities

In multimodal data processing, information redundancy and conflicts across modalities represent core challenges that degrade model performance. Due to inherent differences in data characteristics and sources, modalities often exhibit redundant or conflicting information—for example, background noise in videos can interfere with audio feature extraction, leading to ineffective cross-modal associations during model training and impeding the capture of meaningful semantic relationships. Experimental results indicate that redundant modal features can reduce model accuracy by up to 12%. To address this, dynamic weighted fusion strategies quantify the reliability and importance of modal information to dynamically adjust fusion weights: this approach amplifies the semantic contribution of high-value modalities while suppressing interference from redundant or conflicting data. By optimizing multimodal information utilization, such strategies enhance

model robustness and judgment accuracy in complex scenarios, ensuring more effective modeling of cross-modal semantic relationships [26].

3.1.2 Dynamic Adversarial Environment

The rapid evolution of generative models has created a dynamic adversarial landscape, posing core challenges for multimodal detection research. Frontier models like Sora and Stable Diffusion 3 continuously refine multimodal content generation, producing fake images/videos with unprecedented visual realism, semantic consistency, and physical law compliance—blurring the boundary between authentic and forged content. This forces detection models into an ongoing "generation-detection" arms race, where generative advancements quickly obsolete existing detection strategies. For example, diffusion models now outperform traditional GANs in generating content adhering to physical laws (e.g., consistent lighting, realistic motion dynamics), rendering legacy methods—reliant on GAN-specific structural "fingerprints"—ineffective, as diffusion-generated content lacks such anomalies. This mismatch highlights a critical vulnerability: detection models struggle to adapt to rapid generative technology shifts, underscoring the need for agile, future-proof frameworks capable of dynamically countering emerging adversarial threats [27].

3.1.3 Low-quality Data and Cross-domain Generalization Issues

In multimodal detection, low-quality data and cross-domain generalization challenges significantly impede model performance and real-world engineering implementation. Real-world data often contains inherent noise and uneven quality: compression artifacts distort image/video content, while low-light conditions reduce brightness and obscure details, directly compromising data utility. Meanwhile, substantial distribution mismatches exist between training and test datasets—training data typically originates from controlled, ideal environments, whereas test data encompasses messy, real-world scenarios, limiting models' ability to transfer learned knowledge effectively and constraining generalization. For example, in social media contexts, low-resolution user-uploaded videos lack fine-grained details, causing detection models to fail in feature extraction and leading to severe performance degradation. Academic research has tackled these issues through specialized datasets like DFDC [28], which are critical for enhancing models' capacity to process low-quality data and improve cross-domain generalization. By simulating real-world noise and diverse distribution shifts, DFDC enables models to learn robust representations that bridge the gap between idealized training conditions and complex, unpredictable deployment environments, fostering more reliable multimodal detection systems.

3.2 Frontier Research Directions

3.2.1 Explainability Testing

In multimodal forgery detection, interpretability technologies are pivotal for enhancing system transparency and credibility, enabling precise forged content identification through specialized approaches. Attention visualization, a core technique, pinpoints cross-modal forgery traces by focusing on critical regions in multimodal data: in video analysis, it uses attention heatmaps to highlight anomalies like audiovisual asynchrony or lip-movement-speech mismatches, aiding researchers in locating forgeries and interpreting model decision logic. Physical law analysis, another key pathway, verifies compliance of GAN-generated content with real-world constraints (e.g., optical reflection, fluid mechanics), treating

inconsistencies—such as object light/shadow effects conflicting with environmental optical laws—as discriminative forgery features in image detection. These techniques bridge model predictions and human understanding, ensuring detection systems are both accurate and explainable—essential for building trust in high-stakes fields like digital forensics and content security.

3.2.2 The Dynamic Adversarial Attack-Defense Game

In multimodal forgery detection, the interplay between detection models and forgery technologies drives a dynamic adversarial game, fostering technological innovation and theoretical progress. Continuous learning strategies leverage online updates to track emerging generation techniques: by ingesting real-time data and applying streaming algorithms, models adapt parameters dynamically to recognize evolving forgery patterns (e.g., content from Sora or Stable Diffusion 3), ensuring sustained accuracy and resilience amid rapid advancements. The Projected Gradient Descent (PGD) adversarial training paradigm, valued for efficiency, generates high-fidelity adversarial samples through iterative gradient computations and norm-constrained perturbations, prompting models to learn robust, perturbation-resistant features. This enhances resilience against sophisticated attacks while advancing both theoretical frameworks and engineering capabilities to address escalating forgery complexities.

4 Conclusion

This study systematically examines key advancements in multimodal forgery detection across four core areas: dataset architecture, evaluation frameworks, algorithmic innovations, and technical challenges. Modern datasets emphasize dynamic scenarios and cross-cultural diversity to enhance training efficacy, while evaluation methods have evolved from static analysis to holistic end-to-end attack chain assessment, incorporating human perceptual metrics for validity. Multimodal fusion architectures address unimodal limitations by integrating visual, audio, and textual features through coordinated processing. Innovations like federated learning (ensuring secure cross-institutional collaboration via differential privacy) and self-supervised contrastive learning (boosting few-shot adaptability) alongside adversarial training (enhancing robustness) drive progress. Interpretability techniques (attention visualization, physical constraint analysis) combined with continuous learning facilitate deployment in dynamic environments. By leveraging cross-modal semantic correlation and spatiotemporal modeling, multimodal detection surpasses unimodal approaches in noise robustness, cross-cultural applicability, and complex scenario generalization. Integrating decentralized learning and self-supervised frameworks, these systems transition from label-dependent training to efficient universal verification, marking a fundamental paradigm shift in combating advanced synthetic media threats.

References

1. Radford, A. et al.: ‘Learning Transferable Visual Models From Natural Language Supervision’, arXiv:2103.00020 [cs.CV]. 2021
2. Ho, J. et al.: ‘Denoising Diffusion Probabilistic Models’, arXiv:2006.11239 [cs.LG]. 2020
3. Wang, Z. D. et al.: ‘DIRE for Diffusion-Generated Image Detection’, arXiv:2303.09295 [cs.CV]. 2023

4. Rössler, A. et al.: ‘FaceForensics++: Learning to Detect Manipulated Facial Images’, arXiv:1901.08971 [cs.CV]. 2019
5. Li, M. et al.: ‘A Comparative Study on Physical and Perceptual Features for Deepfake Audio Detection’, https://doi.org/10.1145/3552466.3556523. 2022
6. Tuan, N. M. D. and Minh, P. Q. N.: ‘Multimodal Fusion with BERT and Attention Mechanism for Fake News Detection’, arXiv:2104.11476 [cs.CL]. 2021
7. Zhou, Y. et al.: ‘Multimodal Fake News Detection via CLIP - Guided Learning’, arXiv:2205.14304 [cs.CV]. 2022
8. Tsai, Y. H. et al.: ‘Multimodal Transformer for Unaligned Multimodal Language Sequences’, arXiv:1906.00295 [cs.CL]. 2019
9. Wang, Z. et al.: ‘MedCLIP: Contrastive Learning from Unpaired Medical Images and Text’, arXiv:2210.10163 [cs.CV]. 2022
10. Hu, J. et al.: ‘MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation’, arXiv:2107.06779 [cs.CL]. 2021
11. Jia, C. et al.: ‘Scaling Up Visual and Vision - Language Representation Learning With Noisy Text Supervision’, arXiv:2102.05918 [cs.CV]. 2021
12. Hu, M. et al.: ‘Knowledge distillation from multi-modal to mono-modal segmentation networks’, arXiv:2106.09564 [cs.CV]. 2021
13. Khalid, H. et al.: ‘FakeAVCeleb: A Novel Audio - Video Multimodal Deepfake Dataset’, arXiv:2108.05080 [cs.CV]. 2021
14. Zhang, H., Lin, L. Y., Fang, Q., & Alioto, M.: ‘On-Chip Laser Voltage Probing Attack Detection with 100% Area Coverage at Above/Below the Bandgap Wavelength and Fully-Automated Design’, 2021
15. Qin, Y. et al.: ‘WebCPM: Interactive Web Search for Chinese Long - form Question Answering’, arXiv:2305.06849 [cs.CL]. 2023
16. Korshunov, P., & Marcel, S.: ‘Deepfakes: a new threat to face recognition? assessment and detection.’ arXiv preprint arXiv:1812.08685. 2018
17. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M.: ‘Faceforensics++: Learning to detect manipulated facial images.’ In Proceedings of the IEEE/CVF International Conference on Computer Vision, 1–11. 2019
18. Li, Y. Z., Yang, X., Sun, P., Qi, H. G., & Lyu, S. W.: ‘Celeb-df: A large-scale challenging dataset for deepfake forensics.’ In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3207–3216. 2020
19. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M.: ‘Faceforensics++: Learning to detect manipulated facial images’. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1-11. 2019
20. Jiang, L., Li, R., Wu, W., Qian, C., & Loy, C. C.: ‘Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection.’ In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2889–2898. 2020
21. Haliassos, A. et al.: ‘Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection’, arXiv:2201.07131 [cs.CV]. 2022
22. Dolhansky, B. et al.: ‘The DeepFake Detection Challenge (DFDC) Dataset’, arXiv:2006.07397 [cs.CV]. 2020
23. Zhou, Y. et al.: ‘Multimodal Fake News Detection via CLIP - Guided Learning’, arXiv:2205.14304 [cs.CV]. 2022

24. Yang, X., Li, Y. Z., & Lyu, S. W.: 'Exposing deep fakes using inconsistent head poses', In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8261–8265. IEEE, 2019
25. Madry, A. et al.: 'Towards Deep Learning Models Resistant to Adversarial Attacks', arXiv:1706.06083 [stat.ML]. 2017
26. Wang, L., Zhang, C., Xu, H., Xu, Y., Xu, X., & Wang, S.: 'Cross-modal Contrastive Learning for Multimodal Fake News Detection.' arXiv:2302.14057 [cs.LG]. 2023
27. Tolosana, R., Romero-Tapiador, S., Fierrez, J., & Vera-Rodriguez, R.: 'DeepFakes Evolution: Analysis of Facial Regions and Fake Detection Performance.' arXiv:2004.07532 [cs.CV]. https://doi.org/10.48550/arXiv.2004.07532. 2020
28. Haliassos, A., Mira, R., Petridis, S., & Pantic, M.: 'Leveraging Real Talking Faces via Self-Supervision for Robust Forgery Detection.' arXiv:2201.07131 [cs.CV]. 2022