A Research on Deepfake Face Detection Techniques Based on Multimodal Biometric Cross – Verification

Haozhe Wu

School of Computer and Software, Hohai University, Nanjing, Jiangsu, China

Abstract. The rapid advancement of deepfake technology poses severe threats to social security and information authenticity, as traditional single-modal detection methods face bottlenecks due to their vulnerability to circumvention. This paper systematically reviews deepfake face detection techniques based on multimodal biometric cross-verification, analyzing theoretical foundations, technical approaches, datasets, and challenges. Theoretically, it integrates visual features (facial micro-expressions, corneal specular highlights), auditory features (speech spectra, lip-sync consistency), and physiological signals (heart rate rhythms, facial blood flow), leveraging modal complementarity and consistency verification mechanisms to capture cross-modal forgery traces. Technically, it summarizes feature extraction methods such as CNN-based texture analysis, spectrogram modeling, and near-infrared imaging, and compares early fusion, late weighted voting fusion, and attention-guided dynamic fusion strategies—where attention mechanisms significantly enhance sensitivity to complex cues. It also organizes multimodal datasets (e.g., IAV-DF, DECRO) and evaluation metrics (accuracy, F1-score), providing standardized benchmarks. Although multimodal detection has improved robustness, it still faces challenges such as high-fidelity forgeries threatening modal consistency and inadequate adaptability to complex scenarios. Future research should focus on fine-grained biometric mining, lightweight model deployment, interpretability enhancement, and the improvement of regulations and technical standards to curb misuse and promote legitimate applications in digital security.

1 Introduction

The rapid advancement of deepfake technology in the fields of artificial intelligence and multimedia in recent years has exerted a profound impact on social security, judicial forensics, and other critical domains. Capable of generating highly realistic fake videos, audios, and images, this technology has been misused for malicious activities such as spreading false news and perpetrating identity fraud, severely threatening information authenticity and the societal trust foundation [1]. For example, in April 2024, the French Renaissance Party utilized the open-source tool DeepFaceLab and 2,000 hours of President

2206050109@hhu.edu.cn

Macron's speech materials crawled from the parliamentary official website to create a fake video claiming to "abolish the retirement system." During the forgery process, the organization analyzed the president's micro-expressions, body language, and speech rhythm to construct a highly realistic virtual image, which achieved 12 million views through algorithmic recommendations on Platform X. This directly triggered mass protests in Paris, leading to traffic paralysis and violent conflicts. In March 2024, the U.S. financial system suffered its largest-ever deepfake attack: a criminal gang cloned the voiceprints and images of five top executives from multinational banks, forged a video conference of collective decision-making, and defrauded a $230 million loan. The fraudsters first trained a speech model using earnings call recordings, then constructed 3D facial models by analyzing the executives' social media photos, and finally used "DeepFaceLab + voiceprint synthesizer" to generate lip-synced false instructions. The bank's risk control system, relying solely on static face recognition, failed to detect abnormal micro-expressions and mismatched background light and shadow in the dynamic footage. An investigation by the Financial Crimes Enforcement Network (FinCEN) revealed that the gang employed a "progressive penetration" strategy: phishing for materials in the early stage, generating test videos to exploit vulnerabilities in the mid-term, and executing precise fraud in the final stage.

The spread of false news can disrupt social order, while identity impersonation may lead to personal privacy leaks and even financial fraud. Single-modal detection technologies (relying only on visual or auditory features) have gradually revealed their limitations when confronting increasingly sophisticated deepfake techniques. Single-modal methods are easily bypassed through targeted optimization—for example, forgery technologies can deceive visual or audio-only detection systems by refining corresponding modal features [2]. Therefore, the necessity of multimodal biometric cross-verification has become increasingly prominent. By integrating visual, auditory, and even physiological features (such as facial micro-expressions and speech spectra), multimodal detection technologies can more comprehensively capture forgery traces, thereby enhancing detection robustness and accuracy [3].

Mainstream deepfake generation techniques include generative adversarial networks (GANs), variational autoencoders (VAEs), and face-swapping technologies (e.g., FSGAN and Faceshifter). These techniques have made significant progress in creating realistic forgeries but also pose new challenges. For instance, GAN-generated fake videos are visually indistinguishable from real ones [4], while VAEs excel in audio forgery [5]. However, their abuse has driven researchers to develop more advanced detection methods. Single-modal detection technologies exhibit clear bottlenecks in cross-dataset testing and against adversarial attacks. Visual-based methods, for example, show poor generalization across datasets [6], while audio-based methods are vulnerable to audio adversarial samples [7]. Additionally, single-modal techniques struggle with multimodal forgeries, such as deepfake videos that manipulate both video and audio simultaneously. As a result, research on multimodal detection has emerged as a critical direction in the field of deepfake detection [4]. This paper elaborates on representative techniques for multimodal biometric cross-verification-based detection that have emerged in recent years, systematically enriching the progress in this domain. It also introduces adversarial attack methods targeting multimodal biometric cross-verification and provides insights into future research directions.

2 Theoretical Foundations of Multimodal Biometrics

2.1 Definition and Classification of Biometric Features

Biometric features refer to unique, measurable characteristics of an individual's physiology or behavior, commonly used for identity recognition and verification. Multimodal biometric systems enhance recognition accuracy and robustness by integrating multiple biometric modalities. Below is a classification of biometric features and their applications in multimodal cross-verification:

2.1.1 Visual Features

Visual features primarily include facial micro-expressions, corneal specular highlights, and gaze consistency. Facial micro-expressions—subtle movements of facial muscles associated with emotional states—help detect unnatural expressions in deepfake videos. Corneal specular highlights, the light reflections on the eye surface, exhibit consistent position and shape in real videos, but may show anomalies in forgeries. Gaze consistency analysis examines the alignment between eye fixation direction and head movement to identify inconsistencies in manipulated videos [8].

2.1.2 Auditory Features

Auditory features encompass speech spectral characteristics and lip-sync consistency with speaker identity. Speech spectral characteristics, reflecting the frequency distribution of sound, enable speaker identification. Lip-sync consistency involves analyzing the temporal relationship between speech and lip movements to detect mismatches in forged videos. Studies show that deepfakes often exhibit temporal misalignment between lip movements and speech, providing critical clues for multimodal detection [9].

2.1.3 Physiological Signal Features

Physiological signal features include heart rate rhythms and facial blood flow changes. Heart rate rhythms can be detected through subtle facial movements in videos, while facial blood flow changes are extracted using photoplethysmography (PPG) technology. These features are difficult to replicate in deepfake videos, making them effective for distinguishing real from forged content [10].

2.2 Multimodal Cross-Verification Mechanisms

Multimodal cross-verification mechanisms enhance the accuracy and robustness of deepfake detection by integrating information from different modalities, leveraging their complementarity and consistency. The core mechanisms of multimodal cross-verification are as follows:

2.2.1 Inter-Modal Complementarity

Inter-modal complementarity refers to the mutual supplementation of information expression across different modalities. For example, the visual modality (e.g., lip movements) and the auditory modality (e.g., speech) exhibit strong temporal correlation. Analyzing the lip-sync consistency between speech and lip movements can detect

inconsistencies in forged videos. Additionally, facial micro-expressions in the visual modality and speech emotional features in the auditory modality can validate each other, further improving detection accuracy [11].

2.2.2 Cross-Modal Consistency Analysis

Cross-modal consistency analysis involves matching and verifying features from different modalities based on identity information. For instance, by jointly using speaker recognition and face recognition models, one can validate whether speech and facial features belong to the same identity [12]. This approach effectively addresses scenarios where a single modality is tampered with in forged videos, as forgers find it difficult to replicate features across multiple modalities simultaneously.

The advantage of multimodal cross-verification lies in its ability to fully utilize complementary information across modalities and detect inconsistencies in forged content through consistency analysis. However, designing multimodal systems faces challenges such as the complexity of inter-modal feature alignment and the optimization of cross-modal fusion strategies [13]. Future research can further explore the application of deep learning in multimodal feature fusion and how to enhance system adaptability to novel forgery techniques.

By integrating visual, auditory, and physiological signal features and leveraging multimodal cross-verification mechanisms, the accuracy and robustness of deepfake detection can be significantly improved, providing a strong guarantee for digital media security.

3 Multimodal Biometric Detection Techniques

3.1 Feature Extraction Methods

3.1.1 Visual Feature Extraction

CNN-Based Texture Analysis:Convolutional neural networks (CNNs) play a critical role in visual feature extraction, particularly in texture analysis. In recent years, methods based on local abnormal texture detection have become a research focus. For example, Hu et al. (2023) proposed a method combining CNNs with local texture features, which extracts global image features through CNNs to capture the overall structure and semantic information of faces. To enhance sensitivity to forged regions, the method introduces local texture feature analysis, which can capture subtle inconsistencies introduced during forgery, such as unnatural texture transitions or local detail distortions. By fusing CNN-derived global features with local texture features, this approach enables comprehensive analysis at both global and local levels, thereby improving detection accuracy [14].

3.1.2 Auditory Feature Extraction

Spectrogram Analysis and Speaker Embedding:In auditory feature extraction, spectrogram analysis and speaker embedding techniques have been widely applied. Xue and Zhou (2022) proposed a voice spoofing detection method based on physiological-physical feature fusion, which effectively identifies speech synthesis and voice conversion attacks by analyzing spectrograms and speaker embedding features. Despite its significant achievements, this method has limitations: first, the extraction of physiological features relies on high-quality

speech signals, which may be constrained in low signal-to-noise ratio environments; second, its high computational resource requirements may limit real-time deployment [15].

3.1.3 Physiological Signal Modeling

Near-Infrared (NIR) Imaging and Remote Photoplethysmography (rPPG): Physiological signal modeling is increasingly emerging in multimodal biometric detection. Abd El-Rahiem et al. (2021) proposed a deep fusion multimodal biometric system based on electrocardiogram (ECG) and finger vein features, aiming to address the vulnerability of traditional single-modal biometric systems to attacks (e.g., spoofing) and significantly enhance system security and reliability. The system integrates ECG and finger vein biometrics, leveraging deep learning and signal processing techniques to achieve efficient identity authentication [16]. This research not only advances multimodal biometric technology but also opens new possibilities for related application scenarios.

3.2 Multimodal Fusion Strategies

3.2.1 Early Fusion

Direct Concatenation of Multimodal Features as Input to Deep Learning Models:Early fusion is a technique that directly concatenates multimodal features as input to deep learning models, with the core idea of fusing data from different modalities at an early stage of feature extraction to fully utilize inter-modal complementary information. In the field of deepfake face detection, techniques based on multimodal biometric cross-verification have gradually become a research hotspot. A review of their development shows that early multimodal fusion initially focused on independent processing of single-modal features, with fusion typically occurring at the later stage of feature extraction. However, with the development of deep learning, researchers began to explore the potential of early fusion. Konrad Gadzicki et al. (2020) compared the performance of early fusion and late fusion in human activity recognition (HAR) tasks through experiments and found that early fusion significantly improves model performance in certain scenarios [17]. This study provides an important theoretical and technical foundation for applying early multimodal fusion in deepfake detection. Future research can further explore the optimization of fusion strategies and the expansion of application scenarios based on this foundation.

Shiv Shankar et al. proposed a progressive fusion method, an innovative technique for optimizing feature extraction and fusion in multimodal data processing, which significantly enhances the accuracy and robustness of deepfake detection. This method provides a new technical path for deepfake detection: its stage-by-stage fusion and multimodal interaction mechanisms improve detection performance and offer important references for research in multimodal learning. With further technical optimization and expansion, this approach is expected to play a greater role in more fields in the future [18]. As a key technique, early fusion leverages direct concatenation of multimodal features to fully exploit inter-modal complementarity, significantly enhancing detection performance. With the further development of generative AI and cognitive computing technologies, this field is poised to make breakthroughs in robustness, generalization, and practical applications.

3.2.2 Late Fusion

Weighted Voting of Independent Modal Detection Results:Late fusion is a commonly used technique in multimodal biometric cross-verification, especially in deepfake face detection,

which improves detection accuracy and robustness through weighted voting of independent modal detection results. In 2023, Mulin Tian et al. proposed an unsupervised multimodal deepfake detection method that leverages late fusion by mining intra- and cross-modal inconsistencies. This approach achieves efficient deepfake detection without labeled data, demonstrating the potential of late fusion in unsupervised learning. Specifically, it identifies forged content by analyzing inconsistencies between modalities (e.g., visual and audio), eliminating the need for labeled data and effectively detecting deepfakes under unsupervised conditions—particularly suitable for newly emerged forgery techniques, which highlights the strong adaptability of late fusion in unsupervised scenarios [19].

In 2024, Yizhe Zhu et al. introduced a deepfake detection method based on spatial-frequency feature fusion, using late fusion to perform weighted voting on features from spatial and frequency domains. This method excels in processing compressed images and videos, showcasing the application potential of late fusion in complex scenarios. Its core lies in leveraging Contrastive Spatio-Temporal Distilling (CSTD) technology to combine spatial and frequency-domain features, then conducting weighted voting through late fusion to achieve robust detection in low-quality compressed deepfake videos [20]. By addressing the inability of existing methods to fully utilize spatio-temporal inconsistencies in low-quality videos, this approach further demonstrates the utility of late fusion in complex environments.

Late fusion technology has evolved through stages of initial exploration, dynamic weighting, dual-domain fusion, and unsupervised learning in deepfake face detection via multimodal biometric cross-verification. With advancements in deep learning, late fusion shows strong potential in improving detection accuracy, robustness, and cross-modal consistency. Future research may explore its application in more complex scenarios and integration with advanced techniques like generative adversarial networks and self-supervised learning.

3.2.3 Attention-Guided Fusion

Dynamic Feature Selection Based on Self-Attention Mechanisms: Attention-guided fusion and dynamic feature selection based on self-attention mechanisms have achieved significant progress in deepfake face detection. In the early stages of deepfake detection, researchers primarily relied on single-modal feature extraction and classification. However, as forgery techniques complexified, the limitations of single-modal methods became evident, shifting focus to multimodal feature fusion. For example, Jintao Wen et al. proposed the Dynamic Interactive Multiview Memory Network (DIMMN), which enhances emotion recognition in conversations by fusing text, audio, and video modalities—providing insights for applying multimodal fusion in deepfake detection. Future research could explore its practical use in deepfake detection and integrate more modalities (e.g., depth information, physiological signals) to improve detection accuracy and robustness [21].

In 2024, attention-guided fusion technology made notable advancements in deepfake detection, particularly in audio-visual fusion and personnel verification systems. Xuebin Jing et al.'s research infused new vitality into this field by proposing an audio-visual fusion method based on interactive attention, which dynamically selects critical modal information to significantly enhance system accuracy and robustness [22]. Their innovative fusion mechanism and performance improvements not only advance academic research but also provide strong support for practical applications. With further technical optimization and expanded use cases, this approach is expected to play a larger role in multimodal fusion.

Despite these advancements, challenges remain, such as aligning and fusing multimodal data efficiently, ensuring real-time dynamic feature selection, and improving cross-domain generalization. Future work combining advanced attention mechanisms with multimodal

fusion strategies could achieve higher detection accuracy and robustness in deepfake detection. From early attempts at multimodal feature fusion to the introduction of self-attention and deepened dynamic feature selection, deepfake face detection techniques based on multimodal biometric cross-verification continue to evolve within the framework of attention-guided fusion, driving progress in both detection performance and multimodal information processing.

3.3 Adversarial Robustness Enhancement

Deepfake face detection techniques based on multimodal biometric cross-verification have seen significant progress in recent years, particularly in enhancing adversarial robustness and joint training of statistical and deep features. In 2019, Modak and Jha published a review discussing the importance of multimodal biometric fusion and its potential to boost system robustness. They highlighted that single-biometric systems are limited in the face of noise, spoofing attacks, and intra-class variations, while multimodal fusion—by combining features like fingerprints, irises, and faces—effectively improves recognition accuracy and security [23]. This review provides critical theoretical and practical guidance for researching and applying multimodal biometric fusion.

In 2024, Rashid and Rivas proposed an innovative method to enhance the adversarial robustness of multimodal image captioning models, integrating the Fast Gradient Sign Method (FGSM) with adversarial training to address model vulnerabilities under adversarial attacks. This research holds both theoretical significance and practical value, offering effective solutions for model security in real-world applications [24].

Progress in adversarial robustness enhancement for deepfake face detection via multimodal biometric cross-verification includes advancements in joint training of statistical and deep features and adversarial sample defense techniques. Future directions may involve optimizing multimodal fusion strategies, developing more efficient adversarial training methods, and exploring novel deepfake detection technologies.

4 Datasets and Evaluation Metrics

4.1 Key Datasets for Multimodal Biometric Cross-Verification in Deepfake Detection

4.1.1 IAV-DF (Indian Audio–Video DeepFake Dataset)

Proposed by Kingra et al. in 2025, IAV-DF contains 10,717 real videos and 247,678 deepfake videos, covering diverse content from India and the Middle East. Its innovation lies in enhancing detection robustness through cross-verification of multimodal biometrics (e.g., facial and speech features). Evaluation metrics include accuracy, F1-score, and AUC (Area Under Curve). Experimental results show that its detection performance in cross-cultural scenarios significantly outperforms other datasets.

4.1.2 Deepfake Cross-lingual Evaluation Dataset (DECRO)

Developed by Ba et al. in 2023, DECRO focuses on cross-lingual deepfake detection and includes audio-visual data in multiple languages. Its multimodal features encompass facial expressions, speech spectra, and textual content. Evaluation metrics include cross-lingual

detection accuracy and recall, with experiments demonstrating high generalization capability in multilingual environments.

4.1.3 DeepfakeBench

Introduced by Yan et al. DeepfakeBench aims to provide a standardized benchmark for deepfake detection, covering various generation techniques and multimodal features (e.g., visual, audio, text). Evaluation metrics include model generalization ability, cross-dataset detection performance, and processing efficiency. By adopting unified data processing pipelines and experimental setups, DeepfakeBench significantly improves the comparability and reliability of detection models.

4.1.4 PUDD (Prototype-based Unified Framework for Deepfake Detection)

Proposed by Lopez Pellicer et al. in 2024, PUDD focuses on prototype-based learning for multimodal deepfake detection. Its dataset includes diverse generation techniques and real-scene data. Evaluation metrics include cross-dataset detection performance, model generalization, and processing efficiency. Experimental results show that PUDD achieves high accuracy and robustness when detecting unknown generation techniques and individuals.

4.2 Summary of Evaluation Metrics

Table 1. Evaluation Metrics

Accuracy	F1-Score	Area Under Curve	Cross-Dataset Detection Performance	Processing Efficiency
Measures the overall detection performance of the model, suitable for balanced datasets.	It comprehensively takes into account precision and recall, and is suitable for imbalanced datasets.	Assesses the model's classification performance across different thresholds, applicable to binary classification tasks.	Measures the model's generalization ability on unseen datasets, a critical metric for deepfake detection.	Evaluates the computational complexity and real-time detection capability of the model, applicable to practical application scenarios.

These evaluation metrics and datasets play a pivotal role in deepfake detection based on multimodal biometric cross-verification, providing a rich experimental foundation and standardized evaluation framework for research (as shown in table 1). As generation technologies continue to evolve, future efforts must focus on further optimizing dataset design and evaluation metrics to address more complex detection challenges.

5 Challenges and Future Directions

The rapid development of deepfake technology poses a severe threat to the authenticity of digital media. Deepfake detection techniques based on multimodal biometric cross-verification—by integrating visual, auditory, and other biometric features—effectively enhance detection robustness and accuracy. However, this technology faces significant challenges in practical applications while offering broad opportunities for future research.

5.1 Technical Challenges

5.1.1 Threat of High-Fidelity Forgeries to Multimodal Consistency

Advances in generative adversarial networks (GANs) and diffusion models have significantly improved the realism of deepfakes, especially in audiovisually synchronized multimodal forged videos. Forgers can precisely match visual and auditory features, making traditional single-modal detection methods ineffective [25]. For example, facial expressions and speech features in forged videos may appear highly consistent, masking forgery traces. Additionally, the dynamism and diversity of multimodal forgery techniques further complicate detection [25].

5.1.2 Adaptability to Complex Scenarios

In real-world applications, deepfake detection must handle complex scenarios such as low-light conditions and high-compression rates, where biometric feature extraction and recognition are severely constrained, degrading detection performance. For instance, capturing facial micro-expressions and physiological signals becomes difficult under low light, while high-compression data may lose critical information, further weakening model robustness [26].

5.2 Research Directions

5.2.1 Fine-Grained Biometric Feature Mining

Future research can focus on mining fine-grained biometric features, such as micro-expressions, eye movement patterns, and heart rate changes, to model their deep associations with forgery behaviors. These features are difficult to fully replicate in forgeries, making them effective detection cues. For example, analyzing the temporal dynamics of facial micro-expressions can identify unnatural behaviors in forged videos [27].

5.2.2 Lightweight Detection Models

With the proliferation of edge computing and mobile devices, lightweight detection models are in urgent demand. Research can explore optimizations based on lightweight architectures like MobileNet to enable real-time detection on resource-constrained devices. Techniques such as model pruning, quantization, and knowledge distillation can significantly reduce computational complexity while maintaining detection accuracy [28].

5.2.3 Interpretability Enhancement

Improving model interpretability is another critical direction. Techniques like heatmap regression and 3D key-point detection can precisely localize forged regions and provide intuitive explanations for detection results. Heatmaps, for example, visualize abnormal areas in forged videos, while 3D key-point detection reveals geometric inconsistencies in facial structures [29].

6 Conclusion

With the rapid development of technologies like generative adversarial networks (GANs) and diffusion models, the realism of audiovisual content generated by deepfake technology has significantly improved, expanding its applications in entertainment, education, and other fields. However, it also poses severe challenges to social trust, judicial forensics, and digital security. Traditional single-modal detection methods (relying solely on visual or auditory features) struggle to address increasingly complex forgery attacks, while multimodal biometric cross-verification—integrating visual, auditory, and physiological features—significantly enhances detection robustness and accuracy by exploiting inter-modal complementarity and consistency.

This paper systematically reviews the core advancements in multimodal biometric detection techniques: theoretically, it elaborates on the extraction and fusion mechanisms of visual features (e.g., facial micro-expressions, corneal specular highlights), auditory features (e.g., speech spectra, lip-sync consistency), and physiological signals (e.g., heart rate rhythms, facial blood flow). Technically, it summarizes feature extraction methods such as CNN-based texture analysis, spectrogram modeling, and physiological signal detection, and compares the advantages and disadvantages of early fusion, late fusion, and attention-guided fusion strategies. Additionally, it introduces cutting-edge datasets like IAV-DF and DECRO, as well as key evaluation metrics such as accuracy, F1-score, and cross-dataset generalization, providing standardized benchmarks for research.

Despite significant progress, multimodal detection faces challenges from high-fidelity forgeries threatening modal consistency and inadequate adaptability to complex scenarios like low light and high compression. Future research should focus on deep mining of fine-grained biometrics (e.g., eye movement patterns, hemodynamic signals), lightweight model deployment (e.g., MobileNet optimization), and interpretability enhancement (e.g., heatmap-based forgery localization). Meanwhile, there is an urgent need to improve relevant laws, regulations, and ethical norms and establish technical standards for multimodal detection to curb misuse and promote its legitimate applications in identity authentication, content security, and other domains.

Through this survey, we aim to advance the development of multimodal biometric cross-verification techniques, provide more comprehensive solutions for deepfake detection, and call on all societal sectors to address the potential risks of technology misuse, jointly safeguarding truth and credibility in the digital era.

References

1. Kaur, A., Noori Hoshyar, A., Saikrishna, V., Firmin, S., Xia, F.: ‘Deepfake video detection: challenges and opportunities.’ Artificial Intelligence Review, 2024, 57(6)
2. Li, Q., Gao, M., Zhang, G., Zhai, W., Chen, J., Jeon, G.: ‘Towards Multimodal Disinformation Detection by Vision-language Knowledge Interaction.’ Information Fusion, 2024, 102, 102037
3. Thasiyabi, V.A., Koshy, R., Satheesh, S.: ‘Biometric fusion: Combining multimodal and multi algorithmic approach.’ In: Proc. of the International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES). 2016, 618–620
4. Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., Cao, X., Ren, K.: ‘AVoiD-DF: Audio-Visual Joint Learning for Detecting Deepfake.’ IEEE Transactions on Information Forensics and Security, 2023, 18, 2015–2029

5. Kumari, C.H.L., Prasad, K.V.: 'Video forgery detection and localization using optimized attention squeezenet adversarial network.' Multimedia Tools and Applications, 2024, 83(40), 87697–87725
6. Salah, Z., Abu Elsoud, E.: 'Enhancing Network Security: A Machine Learning-Based Approach for Detecting and Mitigating Krack and Kr00k Attacks in IEEE 802.11.' Future Internet, 2023, 15(8), 269
7. Zhang, Y., Xiao, G., Bai, B., Wang, Z., Sun, C., Tu, Y.: 'An Optimized Transfer Attack Framework Towards Multi-Modal Machine Learning.' In: Proc. of the International Conference on Data-Driven Optimization of Complex Systems (DOCS). 2022, 1–6
8. Thuseethan, S., Rajasegarar, S., Yearwood, J.: 'Deep3DCANN: A Deep 3DCNN-ANN framework for spontaneous micro-expression recognition.' Information Sciences, 2023, 630, 341–355
9. Khadar, K.V.A., Sunil Kumar, R.K., Sameer, V.V.: 'Speaker diarization based on X vector extracted from time-delay neural networks (TDNN) using agglomerative hierarchical clustering in noisy environment.' International Journal of Speech Technology, 2024, 28(1), 13–26
10. Zheng, K., Shen, J., Sun, G., Li, H., Li, Y.: 'Shielding facial physiological information in video.' Mathematical Biosciences and Engineering, 2022, 19(5), 5153–5168
11. Ayetiran, E.F., Özgöbek, Ö.: 'An inter-modal attention-based deep learning framework using unified modality for multimodal fake news, hate speech and offensive language detection.' Information Systems, 2024, 123, 102378
12. Xie, D., Deng, C., Li, C., Liu, X., Tao, D.: 'Multi-Task Consistency-Preserving Adversarial Hashing for Cross-Modal Retrieval.' IEEE Transactions on Image Processing, 2020, 29, 3626–3637
13. Wu, L., Long, Y., Gao, C., Wang, Z., Zhang, Y.: 'MFIR: Multimodal fusion and inconsistency reasoning for explainable fake news detection.' Information Fusion, 2023, 100, 101944
14. Hu, Z., Hu, L.: 'Visual Loop Closure Detection Based on SqueezeNet Multi-layer Feature Fusion and Adaptive Range Matching Algorithm.' Journal of Intelligent & Robotic Systems, 2023, 108(3)
15. Xue, J., Zhou, H.: 'Physiological-physical feature fusion for automatic voice spoofing detection.' Frontiers of Computer Science, 2022, 17(2)
16. Abd El-Rahiem, B., Hammad, M.: 'A Multi-fusion IoT Authentication System Based on Internal Deep Fusion of ECG Signals.' In: Studies in Big Data. Springer International Publishing, 2021, 53–79
17. Gadzicki, K., Khamsehashari, R., Zetzsche, C.: 'Early vs Late Fusion in Multimodal Convolutional Neural Networks.' In: Proc. of the IEEE International Conference on Information Fusion (FUSION). 2020, 1–6
18. Torkamani, M., Shankar, S., Rooshenas, A., Wallis, P.: 'Differential Equation Units: Learning Functional Forms of Activation Functions from Data.' In: Proc. of the AAAI Conference on Artificial Intelligence. 2020, 34(04), 6030–6037
19. Tian, M., Khayatkhoei, M., Mathai, J., AbdAlmageed, W.: 'Unsupervised Multimodal Deepfake Detection Using Intra- and Cross-Modal Inconsistencies.' arXiv, 2023
20. Song, K., Zhu, Y., Liu, B., Yan, Q., Elgammal, A., Yang, X.: 'MoMA: Multimodal LLM Adapter for Fast Personalized Image Generation.' arXiv, 2024

21. Wen, J., Jiang, D., Tu, G., Liu, C., Cambria, E.: 'Dynamic interactive multiview memory network for emotion recognition in conversation.' Information Fusion, 2023, 91, 123–133
22. Jing, X., Wang, Y., Li, D., Pan, W.: 'Melon ripeness detection by an improved object detection algorithm for resource constrained environments.' Plant Methods, 2024, 20(1)
23. Modak, S.K.S., Jha, V.K.: 'Multibiometric fusion strategy and its applications: A review.' Information Fusion, 2019, 49, 174–204
24. Rashid, M.B., Rivas, P.: 'AI Safety in Practice: Enhancing Adversarial Robustness in Multimodal Image Captioning.' arXiv, 2024
25. Hu, R., Zhou, S., Tang, Z.R., Chang, S., Huang, Q., Liu, Y., Han, W., Wu, E.Q.: 'DMMAN: A two-stage audio–visual fusion framework for sound separation and event localization.' Neural Networks, 2021, 133, 229–239
26. Yu, P., Xia, Z., Fei, J., Lu, Y.: 'A Survey on Deepfake Video Detection.' IET Biometrics, 2021, 10(6), 607–624
27. Sundararajan, K., Woodard, D.L.: 'Deep Learning for Biometrics.' ACM Computing Surveys, 2018, 51(3), 1–34
28. Hu, W., Dong, X., Liu, N., Chen, Y.: 'LUMDE: Light-Weight Unsupervised Monocular Depth Estimation via Knowledge Distillation.' Applied Sciences, 2022, 12(24), 12593
29. Dong, J., Wang, Y., Lai, J., Xie, X.: 'Restricted Black-Box Adversarial Attack Against DeepFake Face Swapping.' IEEE Transactions on Information Forensics and Security, 2023, 18, 2596–2608