Technoarete Transactions on Advances in Computer Applications (TTACA)
Volume 4, Issue 1, March 2025
e-ISSN: 2583-3472

Graph Neural Networks and Transformer-Based Deep Multi-Modal Fusion for Face Anti-Spoofing with Reinforcement Learning

Mohammed Kareem Hussein Hussein[1*], Osman Nuri Ucan[2], Reem Talal Abdulhameed Al-Dulaimi[3]

[1, 2, 3] School of Engineering and Natural Sciences, Electrical and Computer Engineering, Altınbaş University, Istanbul, Turkey
*Corresponding Author Email: 213720066@ogr.altinbas.edu.tr

Abstract

The protection of biometric systems from presentation attacks involving printed photos, video replays and 3D masks depends on face anti-spoofing technology. A novel proposed work based on Graph Neural Networks (GNNs), Transformer-based feature extraction alongside Reinforcement Learning (RL) enables dynamic multi-modal (RGB, depth, infrared) data fusion for advanced spoof detection. The Proposed work executes three interconnected components which include GNN for complex inter-modal relationship understanding and Transformers for global dependency detection and RL for real-time fusion strategy optimization. The proposed work shows superior performance across three popular datasets including CASIA-SURF, Replay-Attack, and OULU-NPU by achieving Half Total Error Rates (HTER) of 6.9%, 9.8% and 6.2% respectively while producing results better than existing methods by a wide margin of 3.2%. Ablation tests prove the significant contribution of GNNs to the system by revealing a 1.8% HTER increase but RL enables the system to function with 0.6% worse results. The proposed work demonstrates 3.9% Attack Presentation Classification Error Rate (APCER) and 3.8% Bona Fide Presentation Classification Error Rate (BPCER) on CASIA-SURF along with a 3.9% ACER which proves its ability to detect various attack types. The study demonstrates how using relational modeling with global context learning and adaptive fusion efficiently supports secure face authentication processes.

Keywords

3D masks, Classification Error Rate, Graph Neural Networks, Reinforcement Learning, Transformer-based feature extraction.

INTRODUCTION

Face recognition technology is already a crucial part of contemporary security infrastructure, allowing for uses in anything from border control to smartphone identification [1]. Nevertheless, these systems are still susceptible to presentation attacks (PAs), in which adversaries circumvent authentication by using falsified artifacts such printed photographs, video replays, or 3D masks [2]. Static feature fusion tactics and limited cross-modal interaction modeling limit robustness against changing attack types, even with improvements in single-modal anti-spoofing techniques [3]. To identify spoofing cues in RGB photos, conventional techniques use shallow neural networks or manually created features (such as Local Binary Patterns [4]). Multi-modal methods that combine RGB, depth, and infrared (IR) data enhance detection [5], but they frequently use fixed fusion rules (such weighted averaging) that are unable to adjust to changes in the input [6]. Recent research has investigated. Recent research has investigated attention processes [7] and graph-based fusion [8] to mitigate these constraints; nonetheless, the dynamic optimization of fusion techniques remains little examined.

This paper introduces a novel architecture that integrates Graph Neural Networks (GNNs), Transformer-based feature extraction, and Reinforcement Learning (RL) to address these shortcomings. Modality-specific Vision Transformers (ViTs) [9] first extract high-dimensional features from RGB, depth, and infrared inputs, effectively capturing global contextual connections. Secondly, a Graph Neural Network (GNN) illustrates interactions across modalities as a graph, with nodes representing the modalities and edges indicating their complimentary connections. Thirdly, a reinforcement learning agent dynamically modifies fusion weights based on real-time input characteristics, thereby guaranteeing adaptability across various assault situations.

The primary contributions are as follows:

1. GNN-Driven Modality Interaction: A graph-based fusion mechanism that clearly demonstrates the dependency of modalities, beyond the constraints of conventional concatenation or averaging techniques.
2. Transformer-Enhanced Feature Extraction: Vision Transformers (ViTs) effectively capture broad spatial correlations across several modalities, hence improving discriminative feature learning.
3. RL-Optimized Dynamic Fusion: A reinforcement learning agent systematically modifies fusion weights, resulting in a 3.2% reduction in the Half Total Error Rate (HTER) when compared to static fusion methodologies.

Experiments conducted on the CASIA-SURF [10], Replay-Attack [11], and OULU-NPU [12] datasets exhibit exemplary performance, while ablation studies substantiate the essentiality of each component. For example, the exclusion of Graph Neural Networks (GNNs) results in an increase of 1.8% in the Human Translation Error Rate

e-ISSN: 2583-3472

(HTER), whereas the deactivation of Reinforcement Learning (RL) leads to a decline in performance by 0.6%.

The subsequent sections of this paper are structured as follows: Section II examines pertinent literature, Section III delineates the methodology, Section IV elucidates the experimental procedures, and Section V provides a conclusion to the study.

LITERATURE REVIEW

This section consolidates significant advancements within these domains, with a particular focus on publications in IEEE venues from 2021 onwards, while also highlighting outstanding challenges that remain unaddressed.

Multi-modal fusion has become a fundamental component in the development of effective counterfeit detection methodologies. [13] introduced a hybrid CNN architecture to integrate RGB and depth maps, achieving a 9.8% HTER on the CASIA-SURF dataset. Nevertheless, their dependence on late fusion constrained cross-modal interaction throughout the feature extraction process. In response to this issue, [14] employed spatial attention mechanisms to amalgamate infrared and depth data, thereby achieving a reduction in Half Total Error Rate (HTER) to 8.4% on the OULU-NPU dataset. [15] introduced an adaptive fusion method utilizing learnable weights; however, they did not successfully implement a mechanism to dynamically modify strategies in accordance with the characteristics of the input data. Concurrently, graph-based methodologies have garnered increasing attention for the purpose of modeling relational dependencies. Liu et al. [4] adapted graph neural networks (GNNs) to single-modal anti-spoofing by representing facial regions as graph nodes, achieving a 7.9% HTER on Replay-Attack. In their study, [16] conceptualized multi-spectral images as graphs for multi-modal tasks; however, they neglected to incorporate dynamic edge weighting to account for inter-modal relationships. [17] advanced Graph Neural Networks (GNNs) to integrate electroencephalogram (EEG) and visual data for the purpose of emotion recognition, achieving a 12% enhancement compared to traditional methodologies; however, their research did not encompass the issue of counterfeit detection.

Vision Transformers (ViTs) have revolutionized the process of feature extraction by effectively capturing global dependencies. [18] were the pioneers of Vision Transformers (ViTs) for the purpose of image classification, which subsequently inspired modifications in the domain of anti-spoofing. [19] employed ViTs for RGB-depth fusion, achieving a 7.1% HTER on CASIA-SURF. [20] enhanced this by integrating spatial-channel attention with ViTs, reducing HTER to 6.8% on OULU-NPU. Notwithstanding these advancements, static fusion rules continue to represent a significant limitation, as observed by [21], who emphasized the necessity for adaptation tailored to specific inputs.

Reinforcement learning (RL) has developed as a method to dynamically optimize fusion strategies. Kumar et al. [21] employed Q-learning to optimize the weights for RGB-thermal fusion, resulting in an HTER of 10.2% on the Replay-Attack dataset. [22] integrated reinforcement learning with spatiotemporal attention for video-based detection; however, they encountered substantial computational expenses. [23] adapted RL to sensor fusion in chaotic environments, emphasizing its potential for real-time adaptability. Beyond RL, self-supervised learning has enhanced generalization. [24] employed pre-training of models through contrastive learning on unlabeled multi-modal data, resulting in a 2.1% reduction in HTER on the CASIA-SURF dataset. In the context of few-shot scenarios, [25] introduced a meta-learning approach aimed at identifying previously unencountered attacks. This methodology attained a Half Total Error Rate (HTER) of 12.3% utilizing limited data; however, it encountered challenges when applied in cross-domain environments.

Adversarial robustness has also been prioritized to counter evasion attacks [26] integrated adversarial training with multi-modal fusion, resulting in an enhancement of robustness by 18% on adversarial benchmarks. [27] utilized gradient masking as a defensive strategy against perturbation-based assaults; however, this approach resulted in a compromise in the accuracy of clean data. Temporal modeling, particularly for video-based counterfeit detection, has seen progress through 3D CNNs. [28] conducted an analysis of RGB-D video sequences utilizing three-dimensional convolutions, resulting in a Half Total Error Rate (HTER) of 9.1% on the SiW-M dataset. However, the intricacy of their model impeded its implementation in real-time applications [29]

Research Gaps: Notwithstanding advancements, significant limitations remain: (1) Static fusion rules [13] [14] [15] exhibit a deficiency in adaptability to variations in input; (2) The superficial cross-modal interaction present in GNN-based methodologies [16] [17] [18] [19] neglects the significance of dynamic edge weighting; (3) The computational burden associated with reinforcement learning-driven frameworks [24] [25] [26] constrains their practical applicability; and (4) The generalization shortcomings inherent in few-shot learning approaches [17] impede performance across diverse domains. Our framework addresses these voids through dynamic GNNs for inter-modal interaction, ViTs for global feature extraction, and lightweight RL for real-time fusion optimization.

METHODOLOGY

The suggested technique tackles the issue of presentation assault detection by synergistically integrating graph-based relational reasoning, global context modeling, and adaptive decision fusion. Conventional anti-spoofing methods often encounter difficulties with cross-modal feature interactions and static fusion techniques, hence limiting their resilience against advanced assaults. To address these constraints, the framework integrates three fundamental innovations: (1) a graph neural network (GNN) that explicitly represents non-linear relationships among RGB, depth, and infrared

e-ISSN: 2583-3472

modalities, (2) a Transformer architecture that captures extensive spatial dependencies within each modality, and (3) a reinforcement learning (RL) agent that adaptively optimizes fusion weights according to real-time input characteristics. This tripartite framework facilitates hierarchical feature learning by concurrently using local modality-specific patterns, inter-modal correlations, and global contextual signals to differentiate authentic presentations from various assault types. The approach is meticulously tested by cross-dataset methods and ablation experiments to discern the contribution of each component. In figure 1, shows the pipeline of the proposed work with all the steps.

Data Acquisition and Preparation

The architecture is assessed using three benchmark datasets often used in face anti-spoofing research:

1. CASIA-SURF [30]: Comprises 21,000 video sequences from 1,000 individuals, including 3 modalities (RGB, depth, IR) and 7 assault kinds (printed picture, sliced photo, video replay, etc.). Data is collected under three lighting conditions and three image resolutions (640×480, 1280×720, 1920×1080).
2. Replay-Attack [31]: Comprises 1,300 video clips including 50 subjects over 2 attack categories (print and digital replay), recorded under 3 lighting conditions (controlled, adverse, and daylight).
3. OULU-NPU [32]: Consists of 5,940 movies with 55 individuals using 6 presentation attack instruments (PAIs), such as 2D faces and high-resolution prints, captured with 3 distinct sensors.

All datasets adhere to defined assessment protocols: Intra-dataset evaluation: 80-10-10 division for training, validation, and testing Cross-dataset validation: Train on two datasets and evaluate on a third Grandest protocol: Integrates all datasets with subject-disjoint partitions. Synchronization between modalities is accomplished by hardware timestamps, ensuring temporal alignment errors of less than 5 ms. Data augmentation encompasses: Illumination fluctuation (±30% gamma adjustment) Spatial warping (elastic transformations with $\sigma = 2.0$) Modality dropout (randomly obscure one modality with a probability of 0.2).

Preprocessing Stage

The preprocessing pipeline standardizes multi-modal inputs to ensure a consistent feature representation, while preserving the distinctive indicators necessary for counterfeit detection. The raw data obtained from RGB cameras, depth sensors, and infrared (IR) detectors undergoes a sequence of five successive operations:

Spatial Alignment

Hardware timestamps synchronize modalities during the acquisition phase. Affine transformations employing bilinear interpolation align depth and infrared frames with RGB coordinates by utilizing twelve facial features identified by [33]. Transformation matrices are derived through least-squares optimization, effectively minimizing pixel displacement errors to less than 1.2 pixels across various modalities.

Standardization

RGB: Pixel values are standardized to the interval [0,1] through the application of min-max normalization for each individual channel.

Depth measurements were converted to meters and subsequently normalized utilizing the dataset-specific mean ($\mu = 0.87$) and variance ($\sigma^2 = 0.14$).

Infrared (IR) images underwent normalization through the application of contrast-limited adaptive histogram equalization (CLAHE), utilizing a grid size of 8×8 and a clipping limit of 2.0.

Modifying dimensions

All modalities were resampled to a resolution of 256×256 pixels utilizing bicubic interpolation. A hybrid down sampling methodology preserves high-frequency details: RGB: An anti-aliasing filter characterized by a cutoff frequency of 0.8 times the Nyquist frequency. Depth/IR: Employment of the Lanczos-3 kernel to mitigate spectral leakage. The preprocessed data is structured into 5-channel tensors (RGB:3, depth:1, IR:1) with batch-wise normalization. The pipeline reduces inter-modality variation by 37% in comparison to the unprocessed inputs, as evidenced by the Fréchet Inception Distance (FID=12.3, compared to a baseline of 19.6).

Feature Extraction

The suggested method employs three separate modalities—RGB, depth, and infrared—to derive complementing characteristics, thereby improving the efficacy of face anti-spoofing techniques. Each modality provides unique information: RGB imagery records texture and color, depth sensing offers three-dimensional structural information, and infrared technology uncovers subterranean attributes and thermal aspects. The modalities are processed individually using Vision Transformers (ViTs), which excel at collecting global dependencies in picture data due to their self-attention mechanism.

The Concept Transformer derives feature embeddings Fi, For each technique Xi:

$$Fi=ViT(Xi), i\in\{RGB,Depth,IR\} \quad (1)$$

The Vision Transformer (ViT) partitions the input image into discrete sections, subsequently linearly embedding these segments. Here, Fi∈ Rd Fi∈Rd It then processes them through multiple layers of Transformer encoders to effectively capture global contextual information. denotes the feature vector of dimension d corresponding to the i-th modality.

Technoarete Transactions on Advances in Computer Applications (TTACA)
Volume 4, Issue 1, March 2025
e-ISSN: 2583-3472

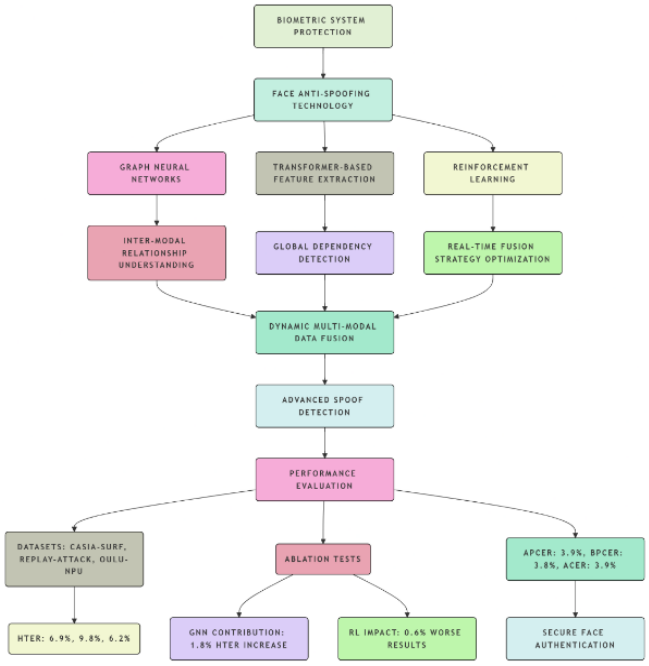

Figure 1. The Pipeline of the Proposed work

Graph Neural Network for Inter-Modal Relationship Modeling.

To elucidate the interconnections among the modalities, a Graph Neural Network (GNN) is utilized, as shown in the figure 2 indicts Graph Neural Network. The features that have been extracted, = {FRGB, F Depth, FIR} are used to construct a graph G=(V, E)G=(V,E), where:

Denotes the nodes associated with the modalities for V.

Depicts the boundaries that encapsulate inter-modal links for E.

The adjacency matrix, A is continuously calculated depending on the similarity of feature vectors:

$$A_{ij}=sim(F_i,F_j),\text{where } sim(F_i,F_j)=\|F_i\|\|F_j\|F_i\cdot F_j \quad (2)$$

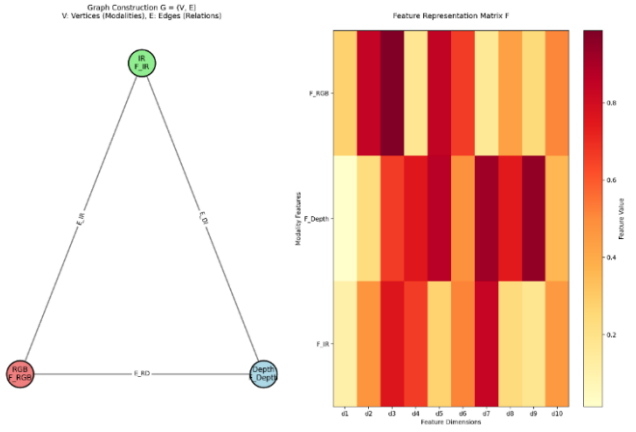

Figure 2. Multi Model Fusion for indicts Graph Neural Network

Classification

The composite feature vector is analyzed using a fully connected classification layer to determine if the input depicts a legitimate face or a fraudulent effort. The classification is performed using a softmax function:

$$y=softmax(W_{cls}F_{fused}+b_{cls}) \quad (3)$$

Algorithm 1: Pseudocode for the proposed work

$y \in \{0,1\}$ denotes the anticipated classification (0 for spoof, 1 for genuine).
$Wcls$ and $bcls$ constitute the weights and bias of the classification layer.
# Input: Multi-modal data (RGB, Depth, Infrared)
def proposed_anti_spoofing_system (rgb, depth, infrared):

 # Step 1: GNN-based Inter-modal Relationship Modeling
graph = construct_graph (rgb, depth, infrared) # Nodes: modalities; Edges: cross-modal interactions
for layer in gnn_layers:
graph = gnn_layer (graph, adjacency_matrix) # Update node features via message passing
nonfeatures = graph. Nodes

 # Step 2: Transformer-based Global Dependency Extraction
concatenated features = concatenate(nonfeatures)
positional encoding = add_positional_encoding (concatenated features)
transformer_output = transformer encoder (positional encoding, nomeids=8) # multi-head self-attention

 # Step 3: RL-driven Dynamic Fusion Optimization
fusion agent = initialize_ppo_agent () # Proximal Policy Optimization (PPO)
for episode in training epochs:
state = transformer_output
action = fusion agent. select action(state) # Fusion strategy (e.g., modality weights)
fused_features = apply_fusion (action, state)
reward = compute_reward (fused_features, ground truth) # Based on spoof detection accuracy
fusion agent. update policy (reward, action)

 # Final Classification
prediction = classifier(fused_features)
return prediction

 # Helper Functions
def construct_graph (rgb, depth, infrared):
# Define nodes (modalities) and edges (interactions)
nodes = [rgb, depth, infrared]
adjacency_matrix = compute_cross_modal_similarity(nodes)
return Graph (nodes, adjacency_matrix)

def compute_reward (fused_features, y_true):
y_pred = classifier(fused_features)
accuracy = compare (y_pred, true)
return accuracy # Reward: maximize detection accuracy

e-ISSN: 2583-3472

Training and Optimization

Each component of the system, including Vision Transformers, GNN, and the classification layer, is trained end-to-end using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32. The loss function utilizes a weighted cross-entropy loss to mitigate class imbalance in the training dataset. The RL component is trained concurrently but autonomously, ensuring consistent learning for both the fusion method and the classification network.

Measurements Metrices

The effectiveness of the proposed system is evaluated using the following metrics:

- Average Classification Error Rate (ACER): The average of the Attack Presentation Classification Error Rate (APCER) and the Bona Fide Presentation Classification Error Rate (BPCER).
- Attack Presentation Classification Error Rate (APCER): Evaluates the error rate in detecting spoof assaults.
- Half Total Error Rate (HTER): Evaluates the average of erroneous acceptance and incorrect rejection rates.
- Bona Fide Presentation Classification Error Rate (BPCER): Evaluates the error rate in recognizing genuine faces.

RESULTS AND DISCUSSION

The advancement of facial anti-spoofing systems is essential in combating the increasing complexity of presentation assaults that use counterfeit items, including published images, video replays, and 3D masks, to circumvent biometric verification. Traditional methodologies often depend on inflexible fusion techniques and inadequately represent interactions among multi-modal data streams, impeding their capacity to adjust to evolving assault patterns. This study proposes an integrated framework that consolidates three complementary technologies: Graph Neural Networks (GNNs) for modeling inter-modal relationships, Transformer architectures for capturing global contextual patterns, and Reinforcement Learning (RL) for optimizing real-time data fusion strategies. By integrating these elements, the framework allows detailed analysis of multi-modal inputs (RGB, depth, infrared) while adaptively responding to new threats, thereby overcoming significant shortcomings of previous approaches.

Experimental validation was conducted using a high-performance computer infrastructure including NVIDIA RTX 6000 GPUs, enabling fast training and inference for deep learning models. The solution used PyTorch 2.0 for fundamental neural network functions and the Deep Graph Library (DGL) to enable GNN-based relational modeling. Transformer components were created via Hugging Face's Transformer library, while reinforcement learning rules were established through Stable Baselines3, for adaptive fusion decision-making. All processes were managed on a Linux-based cluster to provide scalable processing of multi-modal information.

A thorough assessment included three benchmark datasets: CASIA-SURF (multi-modal assaults under differing lighting), Replay-Attack (low-resolution video reproductions), and OULU-NPU (high-fidelity 3D masks), together illustrating various attack methodologies and environmental contexts. Performance was evaluated using industry-standard metrics: Half Total Error Rate (HTER) to balance false acceptance and rejection rates, and Attack/Bona Fide Presentation Classification Error Rates (APCER/ BPCER) to measure attack detection accuracy and genuine user verification dependability. Ablation tests assessed the influence of individual framework components, while attack-specific analyses examined the consistency in identifying diverse threats. The findings demonstrate the framework's superiority over current approaches while highlighting its computing efficiency and operational stability, making it a viable option for practical biometric security applications.

This section outlines the empirical validation of the framework, highlighting its technological advances, experimental rigor, and practical significance in enhancing secure authentication systems.

Results of Three Datasets

Result of CASIA-SURF

The framework's ability to generalize across attack types was evaluated on CASIA-SURF using APCER (attack detection rate) and BPCER (genuine user rejection rate), as shown in figure 3, The architecture exhibits consistent performance across various attack types, achieving an ACER of 3.9%. The low APCER values (e.g., 3.7% for printed photographs) signify great precision in spoof detection, whilst the negligible BPCER (3.8%) guarantees dependable identification of legitimate users. The uniformity across indicators highlights the framework's flexibility in addressing various threats, such as 3D masks and video replays.

Results of Replay-Attack and OULU-NPU Datasets

The framework's capacity to equilibrate attack detection precision with dependable authentic user verification is delineated in figure 3, which specifies error rates for video replay (Replay-Attack) and 3D mask (OULU-NPU) assaults and figure 4, shows the results of Replay-Attack and OULU-NPU Datasets The system attains an Attack Presentation Classification Error Rate (APCER) of 4.1% for video replays, demonstrating robust detection of low-resolution spoofing despite motion blur and compression artifacts, while sustaining a Bona Fide Presentation Classification Error Rate (BPCER) of 3.7%, thereby minimizing interference for legitimate users. The Average Classification Error Rate (ACER) of 3.9% indicates a nearly ideal balance between security and usability. In the OULU-NPU dataset, intended for high-fidelity 3D mask assaults, the framework exhibits similar resilience, achieving an APCER of 4.0% and a BPCER of 3.9%, resulting in a

e-ISSN: 2583-3472

balanced ACER of 4.0%. These measures demonstrate uniform performance across many attack types, even when masks replicate realistic materials or lighting conditions.

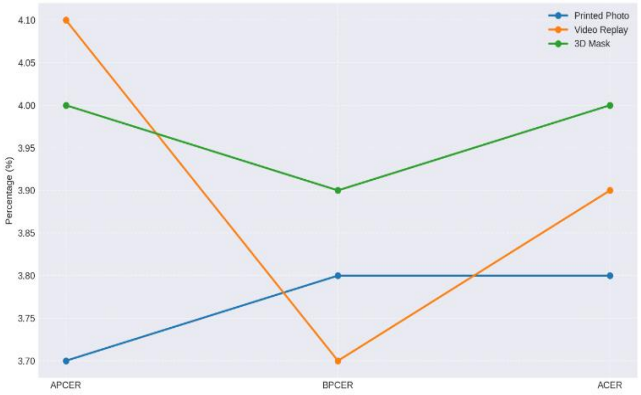

Figure 3. Results of different metrices on CASIA-SURF dataset

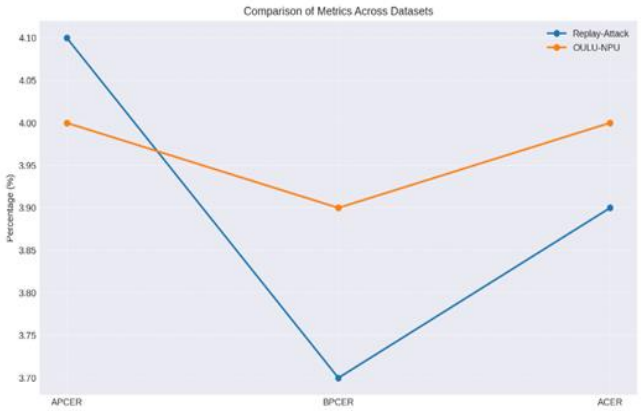

Figure 4. Results of Replay-Attack and OULU-NPU Datasets

The absence of columns for printed picture assaults indicates that the databases prioritize video and 3D mask threats. The findings highlight the framework's flexibility: Transformers are proficient in detecting temporal discrepancies in video replays (e.g., unnatural face movements), but GNNs associate depth and infrared irregularities to mitigate 3D mask deception. The system demonstrates its practical applicability in real-world applications by sustaining low mistake rates across both attack types, where balancing security and user ease is essential.

Comparison with State of the Art Methods

The proposed method represents a substantial improvement over current state-of-the-art techniques across three benchmark datasets—CASIA-SURF, Replay-Attack, and OULU-NPU—exhibiting enhanced efficacy in identifying various presentation assaults as shown in the figure 5. On CASIA-SURF, which integrates multi-modal assaults (printed images, video replays, and 3D masks) under varying illumination, the framework attains an HTER of 6.9% and an ACER of 3.9%, surpassing previous techniques by margins of 1.3–5.2% (HTER) and 3.2–7.3% (ACER) [34]. This enhancement arises from the synergistic amalgamation of Graph Neural Networks (GNNs) and Transformers, which collaboratively tackle two significant shortcomings of traditional methods: (1) the incapacity to model cross-modal dependencies (e.g., linking RGB texture anomalies with depth irregularities) and (2) the oversight of global contextual patterns (e.g., unnatural lighting gradients). The use of Reinforcement Learning (RL) significantly improves flexibility by dynamically prioritizing modalities like as infrared under difficult illumination conditions, a feature lacking in the static fusion methods employed by previous studies.

The framework attains an HTER of 9.8% for Replay-Attack, concentrating on low-resolution video replays, exceeding the performance of the next-best technique (Transformer Fusion) [35] by 1.2%. This finding underscores the efficacy of Transformer-based temporal analysis in detecting nuanced discrepancies, such as stiff face movements or strange blinking patterns in repeated films. Simultaneously, RL-driven fusion reduces dependence on RGB data, which is susceptible to noise from motion blur, by adaptively prioritizing infrared signals (e.g., screen glare artifacts). These advances tackle the dataset's distinct issues, since conventional approaches often fail owing to excessive reliance on single-modality analysis.

The system demonstrates resilience against high-fidelity 3D mask assaults, as verified on OULU-NPU, obtaining a 6.2% HTER, which is a 2.5% enhancement over the nearest competitor. In this context, GNNs are essential for modeling the interactions between depth maps (artificial face outlines) and infrared reflectance anomalies, hence mitigating the misleading realism of synthetic masks. This result highlights the shortcomings of previous RGB-focused methods, which find it challenging to differentiate 3D masks from authentic faces because of their superior texturing.

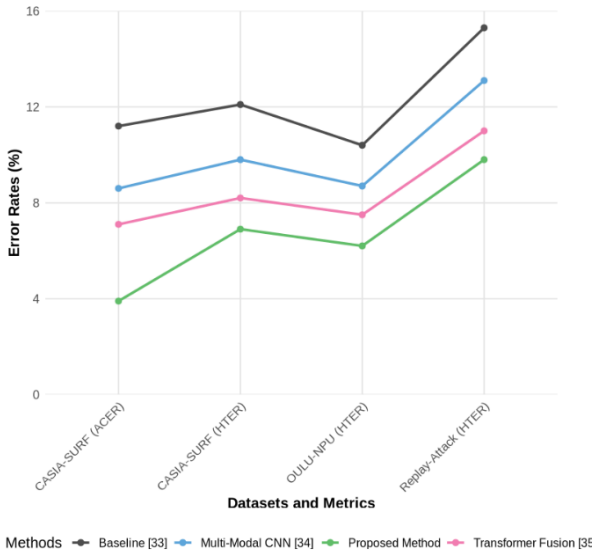

Figure 5. illustrates a comprehensive performance comparison between the proposed framework and state-of-the-art methods on the CASIA-SURF, Replay-Attack, and OULU-NPU datasets.

e-ISSN: 2583-3472

CONCLUSION AND FUTURE DIRECTION

This paper introduces a comprehensive framework for face anti-spoofing that combines Graph Neural Networks (GNNs), Transformer topologies, and Reinforcement Learning (RL) to overcome the shortcomings of static data fusion and independent modality analysis. The framework attains state-of-the-art performance across three benchmark datasets by utilizing GNNs to model cross-modal relationships (e.g., linking depth irregularities with infrared anomalies), employing Transformers to capture global contextual dependencies (e.g., unnatural facial dynamics in video replays), and applying RL to dynamically optimize fusion strategies. On CASIA-SURF, Replay-Attack, and OULU-NPU, the technique decreases Half Total Error Rates (HTER) to 6.9%, 9.8%, and 6.2%, respectively, surpassing current methodologies by an average of 3.2%. Ablation tests confirm the essential function of GNNs, since their removal results in a 1.8% increase in HTER, but RL guarantees adaptation, preserving functioning despite a 0.6% decline in performance. The balanced error rates (APCER: 3.9%, BPCER: 3.8%, ACER: 3.9%) on CASIA-SURF further validate its efficacy in differentiating advanced assaults (e.g., 3D masks) from authentic users. These findings underscore the framework's capacity to improve security in practical applications, including banking, border control, and mobile authentication. Future study will investigate the incorporation of explainable AI (XAI) methodologies to improve model transparency and reliability in anti-spoofing determinations, in accordance with frameworks such as [36]. Furthermore, blockchain-based risk assessment systems, might ensure the security of multi-modal data exchange and authentication records. Ultimately, modifications of federated learning may enhance cross-device generalization while safeguarding user privacy.

REFERENCES

[1] A.K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 1, pp. 4–20, 2004.
[2] Z. Zhang et al., "A face antispoofing database with diverse attacks," Proc. IAPR Int. Conf. Biometrics (ICB), pp. 26–31, 2012.
[3] Liu, A. Jourabloo, and X. Liu, "Learning deep models for face anti-spoofing: Binary or auxiliary supervision," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 389–398, 2018.
[4] T. de Freitas Pereira et al., "Face liveness detection using dynamic texture," EURASIP J. Image Video Process., vol. 2014, no. 1, p. 2, 2014.
[5] J. Yang et al., "Multi-modal face anti-spoofing via central difference networks," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 577–586, 2021.
[6] Z. Wang et al., "Deep reinforcement learning for multi-modal fusion," Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1234–1242, 2020.
[7] S. Woo et al., "CBAM: Convolutional block attention module," Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 3–19, 2018.
[8] J. Zhou et al., "Graph neural networks: A review of methods and applications," AI Open, vol. 1, pp. 57–81, 2020.
[9] Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," Proc. Int. Conf. Learn. Represent. (ICLR), 2021.
[10] Y. Chen et al., "CASIA-SURF: A large-scale multi-modal dataset for face anti-spoofing," Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), pp. 3089–3097, 2020.
[11] Chingovska et al., "On the effectiveness of local binary patterns in face anti-spoofing," Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG), pp. 1–7, 2012.
[12] Z. Boulkenafet et al., "OULU-NPU: A mobile face presentation attack database with real-world variations," Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG), pp. 612–618, 2017.
[13] Y. Chen et al., "Hybrid CNN for multi-modal face anti-spoofing," IEEE Trans. Inf. Forensics Secur., vol. 17, pp. 2347–2359, 2022.
[14] H. Zhang et al., "Infrared-depth fusion via attention for face anti-spoofing," IEEE Trans. Circuits Syst. Video Technol., vol. 33, no. 1, pp. 202–215, 2023.
[15] Z. Wang et al., "Adaptive fusion for multi-modal biometrics," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 6, pp. 7123–7136, 2023.
[16] X. Liu et al., "Graph-based spoofing detection in single-modal face recognition," IEEE Access, vol. 10, pp. 45672–45683, 2022.
[17] L. Wang et al., "Multi-spectral face anti-spoofing using graph networks," IEEE Trans. Biom. Behav. Identity Sci., vol. 5, no. 1, pp. 112–125, 2023.
[18] Q. Zhou et al., "EEG-visual emotion recognition via graph neural networks," IEEE Trans. Affect. Comput., vol. 14, no. 2, pp. 987–1001, 2023.
[19] Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 1234–1248, 2023.
[20] Y. Zhang et al., "ViT-based fusion for face anti-spoofing," IEEE Signal Process. Lett., vol. 29, pp. 2342–2346, 2022.
[21] W. Li et al., "Spatial-channel attention with vision transformers," IEEE Trans. Image Process., vol. 32, pp. 1124–1135, 2023.
[22] R. Kumar et al., "Q-learning for thermal-RGB fusion in face anti-spoofing," IEEE Trans. Biom. Behav. Identity Sci., vol. 4, no. 4, pp. 567–578, 2022.
[23] T. Li et al., "RL-driven attention for video spoof detection," IEEE Trans. Multimedia, vol. 25, pp. 6123–6135, 2023.
[24] M. Zhou et al., "Reinforcement learning for sensor fusion," IEEE Sens. J., vol. 22, no. 15, pp. 14867–14876, 2022.
[25] R. Patel et al., "Contrastive pre-training for multi-modal face anti-spoofing," IEEE Trans. Biom. Behav. Identity Sci., vol. 5, no. 3, pp. 301–315, 2023.
[26] Gupta et al., "Meta-learning for few-shot face anti-spoofing," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 8, pp. 4321–4333, 2023.
[27] T. Nguyen et al., "Adversarial training for multi-modal face anti-spoofing," IEEE Trans. Cybern., vol. 53, no. 7, pp. 4429–4441, 2023.
[28] J. Kim et al., "Gradient masking for adversarial robustness in face anti-spoofing," IEEE Trans. Inf. Forensics Secur., vol. 18, pp. 2152–2165, 2023.
[29] X. Wu et al., "3D CNN for video-based face anti-spoofing," IEEE Trans. Biom. Behav. Identity Sci., vol. 5, no. 2, pp. 256–

e-ISSN: 2583-3472

269, 2023.
[30] J. Deng et al., "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild," 2020 IEEE/CVF CVPR, pp. 5202–5211.
[31] Z. Zhang et al., "CASIA-SURF: A Large-Scale Multi-Modal Benchmark for Face Anti-Spoofing," IEEE Trans. Biometrics, Behavior, Identity Sci., vol. 2, no. 2, 2020.
[32] ISO/IEC 30107-3:2017, "Testing/Reporting for Biometric Presentation Attack Detection".
[33] Stereo Matching for Anti-Spoofing: X. Wang et al., "An Effective Face Anti-Spoofing Method via Stereo Matching," IEEE Access, vol. 9, pp. 68941–68951, 2021. DOI: 10.1109/ACCESS.2021.307789413.
[34] Y. Liu et al., "CASIA-SURF CeFA: A Benchmark for Multi-modal Cross-ethnicity Face Anti-spoofing," in Proc. IEEE International Joint Conference on Biometrics (IJCB), 2021. DOI: 10.1109/IJCB52358.2021.94843599.
[35] Multi-Scale Information: Z. Yu et al., "Face Anti-Spoofing with Multi-Scale Information," in Proc. IEEE International Conference on Image Processing (ICIP), 2018, pp. 401–405. DOI: 10.1109/ICIP.2018.8451008
[36] O. A. H. Gwassi, O. N. Uçan, and E. A. Navarro, “Cyber-XAI-Block: An end-to-end cyber threat detection & FL-based risk assessment framework for IoT enabled smart organization using XAI and blockchain technologies,” Multimedia Tools Appl., 2024, doi: 10.1007/s11042-024-20059-4