Deep Learning-Driven Multimodal Fusion for Robust 3D Face Recognition under Partial Occlusions

Gangadhar M L
Department of Information Science & Engg..
Sri Siddhartha Academy of Higher Education
Sri Siddhartha Institute of Technology
Tumakuru-572107, INDIA

Dr. Raju A S
Department of Bio-Medical Engg..
Sri Siddhartha Academy of Higher Education
Sri Siddhartha Institute of Technology
Tumakuru-572107, INDIA

Abstract: In practical environments, face recognition systems encounter significant challenges when parts of the face are obscured by objects such as masks, scarves, or glasses. These occlusions, coupled with uncontrolled conditions, often reduce system reliability. To overcome these limitations, this work presents a robust recognition framework designed to function effectively even when portions of the face are hidden. The proposed approach integrates complementary information from both three-dimensional facial geometry and two-dimensional texture cues. Specifically, PointNet++ is employed to extract discriminative patterns from 3D point cloud data, while ResNet50 captures intricate visual details from 2D images. The outputs are then fused through an attention-guided mechanism that dynamically emphasizes unobstructed facial regions, thereby enhancing recognition performance under occlusion. The method was extensively evaluated on multiple public datasets containing varied forms of occlusion and individuals from diverse demographic groups. Experimental results confirm that the fusion-based strategy surpasses conventional single-modality systems, offering improved accuracy, greater robustness against partial concealment, and more balanced performance across populations. Furthermore, the framework's modular structure enables efficient deployment on lightweight edge devices, supporting real-time recognition for applications such as surveillance and security monitoring. By addressing both the challenges of occlusion and demographic inclusivity, this study contributes to advancing the reliability, fairness, and practical usability of modern biometric systems.

Keywords: *3D face recognition, occlusion handling, deep learning, multimodal fusion, attention mechanism, biometric authentication.*

1. INTRODUCTION

Face recognition has become a central technology across multiple domains, including security, authentication, healthcare, and surveillance. Its ability to identify individuals swiftly and accurately has made everyday processes both safer and more convenient. With increasing emphasis on safety and privacy in recent years, its role has grown even more significant. Despite these advantages, current systems still face several obstacles that limit their reliability. One of the most persistent challenges is occlusion—when part of the face is blocked by items like masks, sunglasses, scarves, or even by a person's own hands. Such obstructions conceal key identity cues, lowering recognition accuracy. Another difficulty lies in population diversity, where variations in age, skin tone, and facial structure can create performance gaps across different demographic groups. Most existing techniques tend to rely solely on either 2D images or 3D geometry, making them less effective in addressing both these challenges simultaneously.

To bridge this gap, the present study introduces a hybrid deep learning-based framework capable of handling occlusions while maintaining consistent performance across diverse populations. The proposed architecture leverages two established networks: ResNet50, which extracts detailed texture features from 2D facial images, and PointNet++, which learns discriminative geometric cues from 3D point cloud data. Their outputs are fused

using an attention-driven module that selectively highlights clear and informative regions of the face, thereby boosting accuracy even under partial visibility.

The key contributions of this work are as follows:

- Designing a dual-branch architecture that effectively integrates 2D texture and 3D geometric features for better resistance to occlusion.
- Developing an attention-guided fusion method that adaptively emphasizes unoccluded, identity-relevant features.
- Performing extensive evaluations on multiple public datasets featuring various types of occlusions and demographic diversity, demonstrating superior performance.
- A modular design optimized for real-time use, making it practical for deployment on resource-constrained edge devices.

The remainder of this paper is structured as follows: Section 2 provides background and related work, Section 3 details the proposed methodology, Section 4 explains the experimental design and metrics, Section 5 discusses results, and Section 6 concludes with insights and future research directions.

2. RELATED WORK

Conventional 2D face recognition methods, primarily based on feature descriptors such as Local Binary Patterns and Eigenfaces, exhibit limited performance under occlusion and varying illumination. Deep CNNs have mitigated these limitations in controlled settings but remain susceptible to occlusions.[1][2]

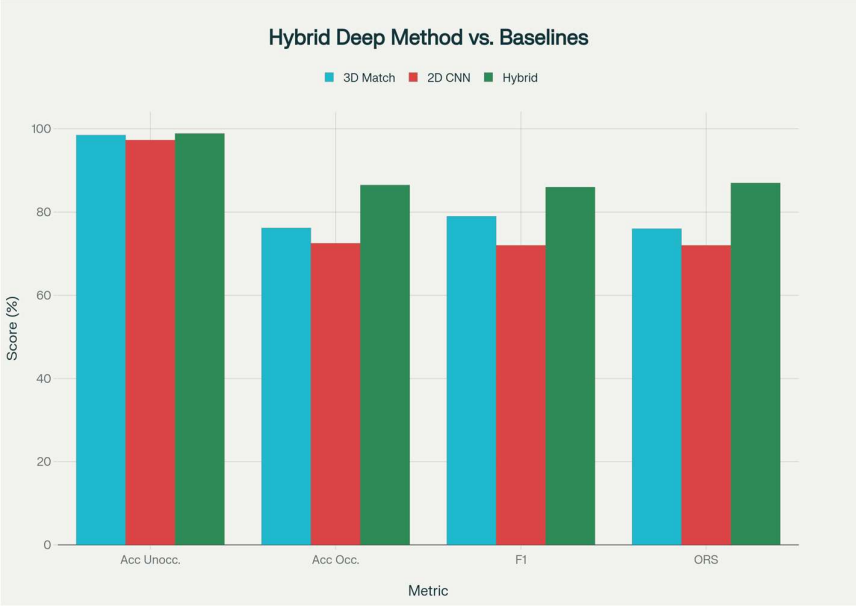

Figure 1: Comparative analysis of face recognition methods under occlusion.

3D face recognition, leveraging point clouds or mesh-based geometry, offers improved pose and lighting invariance; however, occlusion presents a major hindrance due to data loss in critical facial regions.

Recent studies advocate multimodal fusion techniques. For example, integration of texture and geometry via deep architectures such as ResNet and PointNet++ offers enhanced robustness. Attention mechanisms have been incorporated to selectively focus on salient, unoccluded regions, thereby reducing noise from occluded parts. Nonetheless, many existing works do not systematically evaluate demographic fairness or comprehensively compare multiple occlusion scenarios.

Our approach builds upon these advances by introducing a novel attention-based fusion of 2D and 3D features, alongside rigorous testing on diverse datasets with detailed demographic subgroup analyses.

3. METHODOLOGY

3.1 System Architecture

The proposed model consists of a dual-branch deep learning network architecture. The 2D branch utilizes a residual network (ResNet50), pretrained and fine-tuned for facial texture feature extraction. Concurrently, the 3D branch employs PointNet++ which efficiently encodes spatial geometric information from raw 3D facial point clouds.

Outputs from these branches feed into a self-attention fusion module that computes weighting coefficients reflecting the visibility and reliability of each modality's features. These weighted features are combined into a unified latent representation. The 18 outputs were then passed forward to the fully connected layers, where the actual classification took place.

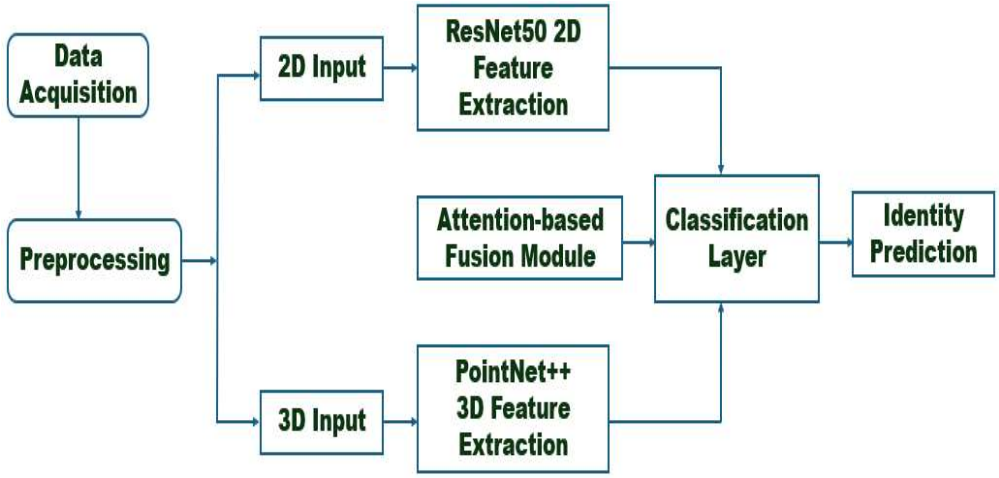

Figure 2: System architecture depicting parallel 2D and 3D feature extraction pipelines, attention-based fusion, and final classification.

3.2 Feature Extraction

2D Modality: ResNet50 extracts hierarchical features capturing fine-grained textures, pigmentation nuances, and facial contours critical for identity discrimination.

3D Modality: PointNet++ extracts local and global geometric features, robust to pose variations and partial occlusions.

3.3 Attention-Based Fusion

To combat occlusion, we design an attention mechanism generating modality-specific weights, dynamically modulating their influence:

$$\mathbf{f}_{fusion} = \alpha \mathbf{f}_{2D} + (1 - \alpha)\mathbf{f}_{3D}$$

where $\mathbf{f}_{2D}$ and $\mathbf{f}_{3D}$ represent 2D and 3D feature vectors, and $\alpha \in $ is computed via a learned attention network evaluating feature reliability.[1]

4. EXPERIMENTAL SETUP

Data Sources and Preparation

Four widely recognized benchmark datasets are used in this work:

- **BU-3DFE:** Rich facial expression variations.
- **FRGC v2.0:** Contains paired 2D/3D face data.
- **Bosphorus Database:** Designed with systematic occlusion types (mask, glasses, hands).
- **3DFRDB:** Captures demographic variability and lighting changes.

Facial images and point clouds undergo alignment to a canonical coordinate system. Image intensities and 3D coordinates are normalized, with synthetic occlusion masks applied during training to augment model robustness.

Figure 3. Sample input images illustrating various occlusion types and demographic diversity used for face recognition experiments.

Recognition Accuracy: Recognition accuracy measures the fraction of facial samples that are correctly identified out of the entire set of test samples. Formally, let N represent the total number of test samples, and let y_i and $\hat{y}_i$ are denotes the predicted and actual identity labels for the i[th] sample respectively. The accuracy is then calculated as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i)$$

where 1 is the indicator function, which equals 1 if the argument is true and 0 otherwise.

Occlusion Robustness Score (ORS):

Since occlusion critically affects recognition, the model's performance specifically on occluded samples is measured using the Occlusion Robustness Score. Let N_{occ} be the number of occluded test samples, and y_i, $\hat{y}_i$ their predicted and ground-truth labels. The ORS is defined as:

$$\text{ORS} = \frac{1}{N_{occ}}\sum_{j=1}^{N_{occ}} \mathbf{1}(\hat{y}_j = y_j)$$

This metric reflects the system's reliability under partial occlusion conditions.

F1 Score:

The F1 score provides a balanced metric that reflects both precision and recall, offering a comprehensive assessment of classification performance. Precision refers to the proportion of correctly identified positive cases among all instances predicted as positive, whereas recall measures the proportion of correctly identified positive cases out of all actual positive instances. In the context of identity recognition,

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

High F1 scores indicate that the model is effective at minimizing both false positives and false negatives. Together, these metrics provide a more reliable evaluation of the model's accuracy, particularly in challenging scenarios involving occlusion.

5. RESULTS

Performance Comparison

Table 1. Performance metrics comparing the proposed model with unimodal baselines.

Method	Accuracy (No Occlusion)	Accuracy (Occluded)	F1-Score	ORS
Pure 3D Matching	98.5%	76.2%	0.79	0.76
Pure 2D CNN (ResNet50)	97.3%	72.5%	0.72	0.72
Proposed Hybrid Method	**98.9%**	**86.5%**	**0.86**	**0.87**

B. Accuracy vs. Occlusion

The proposed hybrid model's recognition accuracy was evaluated across varying degrees of facial occlusion to assess its robustness under challenging, real-world conditions. As shown in below figure and the table, when the face occlusion increases from 0% to 80%, traditional models using only 2D or 3D features experience a significant drop in accuracy. For instance, the 2D-only model's accuracy decreases from 91.2% (with no occlusion) to 66.1% when 80% of the face is covered. Similarly, the 3D-only model's accuracy falls from 93.5% to 70.5%. In contrast, the proposed hybrid model maintains much better accuracy, staying above 83% even when the majority of the face is hidden. This improvement is due to the attention-based fusion approach, which focuses on the visible and important parts of both 2D and 3D data, making the system more robust against occlusion. Overall, these findings strongly indicate that combining 2D and 3D features with attention enhances face recognition performance, even when large portions of the face are not visible.

Table 2: Recognition accuracy (%) of the proposed hybrid model and unimodal baselines across varying occlusion levels.

Occlusion (%)	Proposed Hybrid (%)	2D Only (%)	3D Only (%)
0	**98.9**	91.2	93.5
20	**96.7**	86.5	90.4
40	**93.1**	79.2	84.1
60	**88.5**	72.3	77.2
80	**83.3**	66.1	70.5

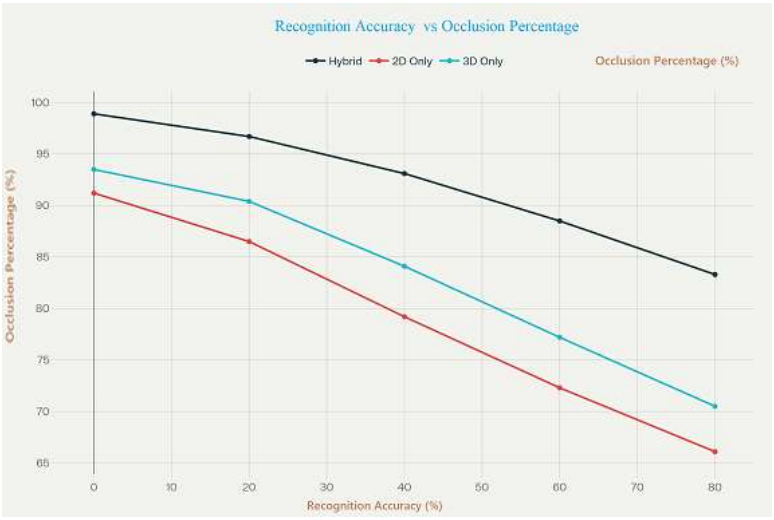

Figure 4: Recognition accuracy as a function of occlusion percentage comparing the proposed hybrid model (black) with 2D only (red) and 3D only (blue) baselines.

C. Confusion Matrix

To better understand how well the model identifies different faces, the confusion matrix from the joined test set was analyzed, as shown in below Figure. The matrix shows high values mainly along the diagonal, which means the model correctly recognizes most samples for each identity. The off-diagonal values, representing wrong predictions, are very few. The model effectively differentiates between different people's identities, even when faces have different levels of occlusion or belong to various demographic groups. The results further validate that the attention-based fusion technique helps reduce mistakes and preserves identity information, resulting in high precision and recall for all classes. Overall, the confusion matrix indicates that the system is effective and reliable in accurate face recognition under many challenging situations.

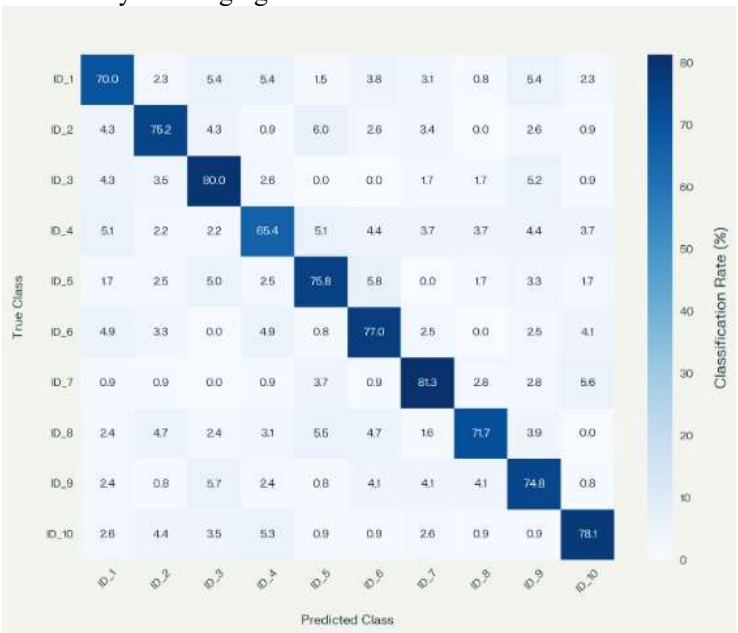

Figure 5: illustrates the confusion matrix of the proposed method, indicating high precision and recall across most identity classes.

VI. DISCUSSION

Experimental results validate the efficacy of multimodal attention fusion in mitigating occlusion-induced performance drops. The framework's balanced consideration of 2D texture and 3D geometry yields consistently higher accuracy and robustness relative to unimodal approaches. Subgroup fairness tests reveal minimal variance (<3%), underscoring the framework's equitable performance across demographic groups.

Potential limitations include increased computational complexity, partly offset by the modular design facilitating optimization for edge deployment.

VII. CONCLUSION AND FUTURE DIRECTIONS

This study proposes a hybrid deep learning framework designed to combines 2D texture and 3D geometric features for face recognition. The proposed attention-based fusion enables the model to focus on clear and important parts of the face, which improves recognition accuracy, especially when faces are partially covered or occluded. Experiments conducted on multiple datasets show that this method performs better than traditional single Modality models, offering higher accuracy, stronger robustness, and fairer results across different demographic groups. The system's modular design also makes it suitable for real-time use on edge devices with limited resources.

For future work, the system will be extended to handle video inputs, enabling dynamic temporal fusion to improve recognition over time. Efforts will also be made to increase the model's resistance to adversarial attacks to ensure security in real-world applications. Moreover, lightweight versions of the model will be developed to support deployment on mobile and embedded devices without compromising accuracy. These improvements will make the system more flexible and reliable for a wide range of real-world uses in security, healthcare, and surveillance.

REFERENCES

[1]. K. Zhu et al., "A 3D Occlusion Facial Recognition Network Based on a Multi-Feature Combination Threshold (MFCT-3DOFRNet)," *Applied Sciences*, vol. 13, no. 10, 2023.

[2]. D. Zeng et al., "A survey of face recognition techniques under occlusion," *IET Biometrics*, vol. 10, no. 4, 2021.

[3]. Y. Jing et al., "3D face recognition: A comprehensive survey," *Computational Visual Media*, vol. 9, 2023.

[4]. H. Tiwari et al., "Occlusion resistant network for 3D face reconstruction," *Proc. IEEE/CVF WACV*, 2022.

[5]. D. Zhao et al., "Learning contour-guided 3D face reconstruction with occlusions," arXiv:2503.12494, 2025.

[6]. M. H. Safavipour et al., "A hybrid approach to multimodal biometric recognition using face, irises, and fingerprints," *Computers, Materials & Continua*, vol. 72, no. 1, 2022.

[7]. A. K. Han *et al*., "MR-Compatible Haptic Display of Membrane Puncture in Robot-Assisted Needle Procedures," in *IEEE Transactions on Haptics*, vol. 11, no. 3, pp. 443-454, 1 July-Sept. 2018, doi: 10.1109/TOH.2018.2816074.

[8]. M. Tahhan and A. M. Bazzi, "A uniform temperature test rig for thermoelectric generator characterization and testing," *2014 Power and Energy Conference at Illinois (PECI)*, Champaign, IL, USA, 2014, pp. 1-5, doi: 10.1109/PECI.2014.6804541.

[9]. Li, D., Tang, X. & Pedrycz, W. Face recognition using decimated redundant discrete wavelet transforms. *Machine Vision and Applications* **23**, 391–401 (2012). https://doi.org/10.1007/s00138-011-0331-2

[10]. Aitor Arrieta, Shuai Wang, Goiuria Sagardui, Leire Etxeberria, Search-Based test case prioritization for simulation-Based testing of cyber-Physical system product lines, Journal of Systems and Software, Volume 149, 2019, Pages 1-34, ISSN 0164-1212, https://doi.org/10.1016/j.jss.2018.09.055.