Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

EagleEye: A Multi-Modal Face-Structure Capture and Enhanced Forensic Recognition System

Abin S
Dept. of CSE
FISAT, Angamaly

Adarsh Vinod
Dept. of CSE
FISAT, Angamaly

Ayush Madhavan
Dept. of CSE FISAT,
Angamaly

Charles P Manoj
Dept. of CSE
FISAT, Angamaly

Ms. Jishna N V
Assistant rofessor
FISAT, Angamaly

***Abstract*—Modern surveillance systems have transitioned from simple recording tools to complex proactive identification frameworks. However, traditional 2D face recognition often fails under challenging environmental conditions, such as extreme pose variations and physical occlusions. This paper introduces EagleEye, a multi-modal Face-Structure Capture and Enhanced (FSCE) recognition system designed for real-time forensic surveillance. The framework integrates an InsightFace-based detection engine for 512-dimensional embedding extraction, a MediaPipe-driven 468-point 3D facial mesh capture for structural documentation, and a generative LaMa-based inpainting module for recovering occluded facial regions. By fusing these heterogeneous modalities through a k-NN vector search and a persistent IOU-based tracking mechanism, EagleEye achieves high-fidelity identification across distributed camera nodes. Experimental results demonstrate a detection accuracy of 99.5% and a recognition accuracy of 98.6%. The system provides an end-to-end architecture encompassing data ingestion, AI-driven processing, and a centralized monitoring dashboard, offering a scalable solution for modern law enforcement and forensic investigations.**

***Index Terms*—Biometrics, 3D Face Mesh, Deep Learning, Image Inpainting, Forensic Surveillance, InsightFace, MediaPipe, ArcFace.**

I. Introduction

Real-time video surveillance represents one of the most critical frontiers in public safety, responsible for the proactive identification of security threats across diverse urban environments[cite: 18, 19]. With the rapid growth of smart cities and increasing population density, surveillance systems have become an essential component of modern infrastructure. These systems are deployed across transportation hubs, public institutions, commercial complexes, and high-security zones to monitor activities and ensure safety. However, despite the widespread adoption of CCTV networks, many existing systems remain fundamentally passive in nature.

Most traditional surveillance frameworks rely heavily on manual monitoring and post-event forensic analysis, which significantly limits their effectiveness in real-time threat detection. Human operators are required to continuously observe multiple video feeds, making the process both labor-intensive and prone to fatigue-induced errors. As a result, critical events may go unnoticed or be identified too late, reducing the overall efficiency of the surveillance system.

The primary challenge in contemporary defense is the increasing sophistication of evasion techniques, where targets utilize physical occlusions such as masks, sunglasses, scarves, and other disguises to bypass traditional 2D recognition systems[cite: 20, 21]. These occlusion strategies exploit the limitations of conventional face recognition methods, which primarily depend on visible facial features. In many real-world scenarios, attackers can obscure their identity within seconds, rendering traditional systems ineffective. This creates a significant gap in security infrastructure, especially in high-risk environments.

The core challenge in defense is that while surface appearances may change due to lighting variations, pose differences, or intentional disguises, the underlying structural architecture of the human face remains relatively stable and consistent[cite: 22]. This observation forms the basis for advanced recognition techniques that focus on deeper geometric and feature-level representations rather than superficial pixel-based analysis. By leveraging structural information such as facial landmarks, embeddings, and 3D geometry, it becomes possible to design systems that are more robust against visual manipulation.

Traditional surveillance methods, such as manual review and simple 2D template matching, are struggling to keep pace with these advancements and are becoming ”permanently outdated”[cite: 23]. Conventional techniques typically rely on predefined features or handcrafted descriptors that lack adaptability. These methods fail to generalize across different environmental conditions and are particularly ineffective when dealing with zero-day targets—individuals who are not previously registered in the system—or those employing heavy occlusion strategies[cite: 24, 31]. Consequently, there is a growing need for intelligent systems capable of learning and adapting to complex visual patterns in real time.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

There is a critical and unmet need for an adaptive detection system capable of analyzing the persistent structural features of the face, rather than easily modifiable content. Such a system should integrate multiple modalities of analysis, including deep feature embeddings, 3D facial geometry, temporal tracking, and image reconstruction techniques. This necessity forms the foundation for the proposed EagleEye framework[cite: 32], which aims to address the limitations of existing surveillance systems through a unified and intelligent approach.

The scope and objectives of this research are as follows:

1) Develop a multi-modal surveillance system integrating face detection, 3D mesh capture, inpainting, and geo-tracking[cite: 33, 34].
2) Extract and analyze structured face features using 512-dimensional metric learning via the ArcFace model[cite: 34].
3) Detect 3D forensic patterns and surface geometry using a fine-tuned MediaPipe FaceMesh model[cite: 35].
4) Identify and reconstruct deceptive or occluded face regions using generative LaMa inpainting analysis[cite: 36].
5) Improve detection performance and operational intelligence through a centralized HQ server and multi-node synchronization[cite: 37].

In addition to these objectives, the proposed system emphasizes real-time processing, scalability, and modularity. The architecture is designed to handle continuous video streams while maintaining high throughput and low latency. Each module in the system operates independently yet cohesively, enabling efficient processing and easy extensibility. This modular design allows for future integration of additional functionalities such as behavior analysis, anomaly detection, and predictive analytics.

In this work, we propose a multi-modal recognition system that integrates InsightFace (Buffalo_L), MediaPipe, and LaMa inpainting to accurately identify targets[cite: 38]. InsightFace is utilized for extracting high-dimensional facial embeddings that capture discriminative identity features. MediaPipe FaceMesh provides dense 3D landmark detection, enabling structural analysis of facial geometry. LaMa inpainting is employed to reconstruct occluded regions, thereby enhancing recognition performance under challenging conditions.

The proposed architecture combines structured feature analysis with deep learning representation learning, along with a k-NN search mechanism for final identity verification[cite: 39]. This hybrid approach ensures both accuracy and efficiency, as embedding-based similarity search allows rapid identification even in large-scale databases. Furthermore, the integration of tracking mechanisms ensures identity persistence across video frames, improving reliability in dynamic environments.

This approach aims to improve robustness, detection accuracy, and adaptability against evolving evasion strategies[cite: 40]. By combining multiple AI-driven techniques into a single unified framework, EagleEye represents a significant step toward next-generation intelligent surveillance systems. The system not only enhances recognition capabilities but also provides valuable forensic insights, making it suitable for applications in law enforcement, security monitoring, and smart city infrastructure.

Ultimately, the proposed system seeks to transform surveillance from a reactive process into a proactive and intelligent solution, capable of addressing modern security challenges with greater efficiency and accuracy.

II. LITERATURE STUDY

Forensic detection has evolved through multiple research directions, ranging from static code inspection and heuristic-based classifiers to deep learning and similarity-driven approaches[cite: 42]. In the early stages, systems primarily relied on handcrafted rules and predefined signatures to identify patterns. While these approaches were effective in controlled environments, they lacked adaptability when exposed to dynamic and real-world scenarios. As a result, modern research has increasingly shifted toward data-driven approaches that leverage machine learning and deep neural networks for improved accuracy and generalization.

While many existing approaches report high laboratory performance, most systems rely heavily on static 2D features or computationally expensive similarity comparisons that do not scale well in real-time environments[cite: 43]. These systems often fail when deployed in large-scale surveillance settings, where continuous video streams and multiple targets must be processed simultaneously. Additionally, reliance on static features makes them vulnerable to variations in lighting, pose, and intentional obfuscation techniques such as occlusion. These limitations highlight the need for a more adaptive, multi-modal framework capable of handling real-time complexity[cite: 44].

Venturi et al. (2022) [cite: 45] proposed a three-phase framework designed for identifying kits, a structural concept that can be extended to identifying consistent facial geometry in surveillance systems. Their approach utilizes static inspection techniques to detect common artifacts across samples, achieving an F1-score of 0.9[cite: 46]. This demonstrates the effectiveness of structural pattern recognition in identifying similarities across instances. However, the system relies on predefined static patterns and does not account for dynamic variations that occur during real-time execution. As a result, it cannot effectively handle adaptive targets that employ randomization or disguise techniques[cite: 47, 48]. This limitation is particularly significant in surveillance applications where targets may intentionally alter their appearance.

Orunsolu and Sodiya (2017) [cite: 49] introduced a modular architecture that emphasizes pipeline-based processing, which is a concept utilized in the design of the EagleEye framework[cite: 50]. Their system demonstrates the advantages of modularity, including improved scalability, flexibility, and ease of integration. However, the approach relies heavily on predefined heuristics for detection, making it vulnerable to zero-day structures that do not match the existing rule set[cite: 51, 52]. This highlights the importance of incorporating learning-based approaches that can adapt to previously unseen patterns.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

Lu et al. (2022) [cite: 53] proposed a homology-based approach emphasizing structural and visual similarity, which aligns with the concept of analyzing facial geometry through dense landmark representations. In the context of EagleEye, this is reflected in the use of 468-point 3D landmark extraction using MediaPipe FaceMesh[cite: 54]. Their dual-layered approach enables the identification of ”cloned” appearances that share deep structural similarities, even when surface-level features differ[cite: 55]. While this method is effective in capturing structural consistency, it introduces significant computational overhead due to extensive pairwise comparisons. This makes it less suitable for real-time applications, where latency and processing speed are critical factors[cite: 56].

Purwanto et al. (2022) [cite: 57] presented PhishSim, a feature-free detection framework that leverages compression distance as a similarity metric, achieving an AUC of 96.68%[cite: 58]. The advantage of this approach lies in its ability to compare samples without relying on explicit feature extraction. However, such methods can become computationally expensive when applied to large datasets. In the EagleEye system, this limitation is addressed by replacing compression-based similarity with L2-normalized 512-dimensional embeddings, enabling efficient and scalable comparison across large identity databases[cite: 59]. This embedding-based approach significantly reduces computational complexity while maintaining high recognition accuracy.

Opara et al. (2024) [cite: 60] explored deep learning through the WebPhish framework, which identifies correlations between structural patterns using neural networks[cite: 61]. Their approach demonstrates the power of deep learning in capturing complex relationships within data. However, the architecture is computationally intensive, making it challenging to deploy on edge devices where resources are limited and real-time processing is required[cite: 62]. This limitation emphasizes the need for lightweight yet effective models that can operate efficiently in constrained environments.

Across these studies, a common trend can be observed: while individual techniques offer strong performance in specific areas, they often lack integration into a unified system. Many approaches focus solely on detection, similarity analysis, or classification, without considering the complete pipeline required for real-time surveillance. Additionally, challenges such as scalability, latency, and robustness under occlusion remain largely unresolved.

The EagleEye framework addresses these gaps by combining multiple complementary techniques into a single cohesive architecture. By integrating deep embedding-based recognition, 3D facial landmark analysis, inpainting for occlusion handling, and efficient tracking mechanisms, the system provides a more robust and scalable solution. This multi-modal approach ensures that the limitations of individual methods are mitigated through synergy, resulting in improved overall performance.

Furthermore, the use of modular design principles allows the system to maintain flexibility and extensibility. Each component can be independently optimized or replaced without affecting the overall pipeline, enabling continuous improvement and adaptation to emerging challenges. This design philosophy aligns with the evolving requirements of modern surveillance systems, where adaptability and real-time performance are critical.

In summary, existing literature highlights the strengths and limitations of various detection and recognition approaches. While significant progress has been made, there remains a clear need for integrated, real-time systems capable of handling complex and dynamic scenarios. The proposed EagleEye system builds upon these insights to deliver a comprehensive and efficient surveillance solution.

III. Methodology

A. Overview

The proposed methodology consists of six major steps: data ingestion, preprocessing, feature extraction, multi-modal model execution, fusion-based classification, and explainable evaluation[cite: 70]. These stages collectively form a continuous processing pipeline that transforms raw surveillance video into structured, interpretable intelligence outputs. The system is designed to operate under real-time constraints, ensuring minimal latency while maintaining high analytical accuracy.

The system performs five types of feature extraction: Detection embeddings, 3D Mesh landmarks, Inpainting visual features, occlusion heuristics, and GPS network features[cite: 71]. Each feature type contributes uniquely to the overall recognition process. Detection embeddings provide identity-specific signatures, while 3D mesh landmarks capture structural invariants of the face. Inpainting features enable recovery from occlusions, and GPS features provide contextual spatial awareness. The integration of these heterogeneous features enables a multi-dimensional understanding of identity.

It employs three primary components: (1) InsightFace Buffalo_L for structured detection, (2) MediaPipe for 3D visual learning, and (3) LaMa for inpainting analysis[cite: 72, 73]. These components are orchestrated within a hybrid pipeline that combines sequential execution with parallel processing, ensuring both efficiency and robustness.

B. Dataset and Preprocessing

The surveillance dataset was constructed using video source codes and enrollment face photographs[cite: 74, 75]. The dataset includes live-stream frames as well as pre-registered identity images used for matching. This dual-source dataset enables both real-time detection and identity verification within a unified framework.

During preprocessing, corrupted frames and inaccessible nodes were removed to ensure data reliability[cite: 76]. This filtering step is essential to prevent propagation of noise into downstream modules. Additionally, preprocessing ensures that only valid and consistent frames are processed, thereby improving the stability of feature extraction.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

1) 3D Mesh Parsing and Structural Cleaning: Raw video frames often contain redundant noise[cite: 77], including environmental distortions, motion artifacts, and lighting inconsistencies. To standardize structural analysis, the following steps were applied:

- Parsing frames to construct the 3D landmark mesh
- Extraction of 468 points (including eye-regions, lips, and contours)
- Normalization of extracted structural attributes

The extraction of dense landmark points enables precise modeling of facial geometry. These landmarks are normalized to eliminate scale and positional variance, ensuring consistency across frames. This structural cleaning process ensures that geometric features remain invariant to superficial changes[cite: 78, 79, 80].

2) Visual Feature Pre-processing: To prepare visual inputs for the deep learning model, crops of each face were first collected and subjected to normalization[cite: 81]. This step isolates relevant facial regions and eliminates background interference.

The captured images were resized to a fixed resolution (640×480) compatible with the CNN architecture[cite: 82]. Standardization of image dimensions reduces computational overhead and ensures compatibility with model input requirements. Additionally, normalization enhances convergence behavior during feature extraction[cite: 83].

3) Feature Structuring: After preprocessing, each sample was organized into a unified structured format containing face embeddings, raw 3D mesh data, extracted landmarks, cleaned logs, face crops, and GPS coordinates. This structured representation ensures that all relevant modalities are aligned temporally and spatially.

Such organization enables efficient multi-modal feature fusion, allowing the system to simultaneously analyze structural, visual, and contextual information. This unified representation forms the basis for downstream classification and decision-making processes.

C. System Architecture

The proposed multi-modal detection system architecture (Fig. 1) consists of sequential processing stages from stream ingestion to dashboard presentation[cite: 84, 85]. The architecture is designed to support continuous real-time operation while maintaining modular extensibility.

The architecture branches into three parallel extraction pipelines:

- **Structured Pipeline:** Processes detection embeddings and tracking IDs using k-NN similarity search. This pipeline focuses on identity representation and matching.
- **Visual Pipeline:** Handles 3D mesh coordinates and inpainting crops via MediaPipe and LaMa. This pipeline emphasizes structural and reconstruction-based analysis.
- **Network Pipeline:** Manages GPS synchronization and communication with the HQ server, ensuring spatial awareness and centralized coordination.

These pipelines operate concurrently, allowing the system to extract complementary features without increasing latency. The parallel design significantly enhances throughput and enables scalable deployment[cite: 89, 90, 91].

The architecture is organized into multiple functional layers that collectively support end-to-end processing. The ingestion layer captures input from multiple sources such as webcams, RTSP streams, and network cameras, which are unified using a StreamLoader module for consistent frame acquisition. This ensures that heterogeneous input sources are normalized into a common processing format.

The AI/ML processing layer performs the core computational tasks of the system. InsightFace Buffalo_L is used for face detection and extraction of 512-dimensional embeddings, which represent the structured identity signature of individuals. MediaPipe FaceMesh extracts 468-point 3D facial landmarks, enabling structural analysis of facial geometry. LaMa and OpenCV-based methods are used for face inpainting, reconstructing occluded regions to improve recognition reliability.

The tracking and recognition layer maintains identity persistence across frames using an IOU-based tracking mechanism. Bounding boxes are matched across consecutive frames to assign consistent identities. In parallel, k-NN vector search using cosine similarity is applied on embedding vectors to perform identity matching. This ensures both temporal consistency and accurate recognition.

The application layer includes functional modules such as alert management, forensic analysis, location services, and HQ synchronization. These components enable real-time decision-making, logging of events, and integration with centralized monitoring systems. The forensic engine processes structural and visual outputs to generate interpretable evidence.

The presentation layer provides visualization through a Flask-based web dashboard and centralized monitoring via the police HQ server. This layer allows operators to observe live feeds, detected identities, and system alerts in real time.

The data layer supports storage and retrieval of system information. Embeddings are stored in a vector database (JSON/Milvus), user-related data is maintained in SQLite databases, and historical logs are stored in JSONL format. This structured storage enables efficient querying and retrieval during recognition and analysis.

Data flows sequentially from the ingestion layer to the presentation layer, while intermediate outputs such as embeddings, tracking IDs, and GPS metadata are continuously stored and accessed from the data layer. The integration of parallel pipelines ensures that structural, visual, and network features are processed simultaneously without increasing computational latency.

The multi-layered architecture ensures modularity, scalability, and efficient real-time processing. By separating concerns across layers and enabling parallel execution, the system achieves robustness against occlusion, environmental variations, and dynamic surveillance conditions while maintaining high throughput.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

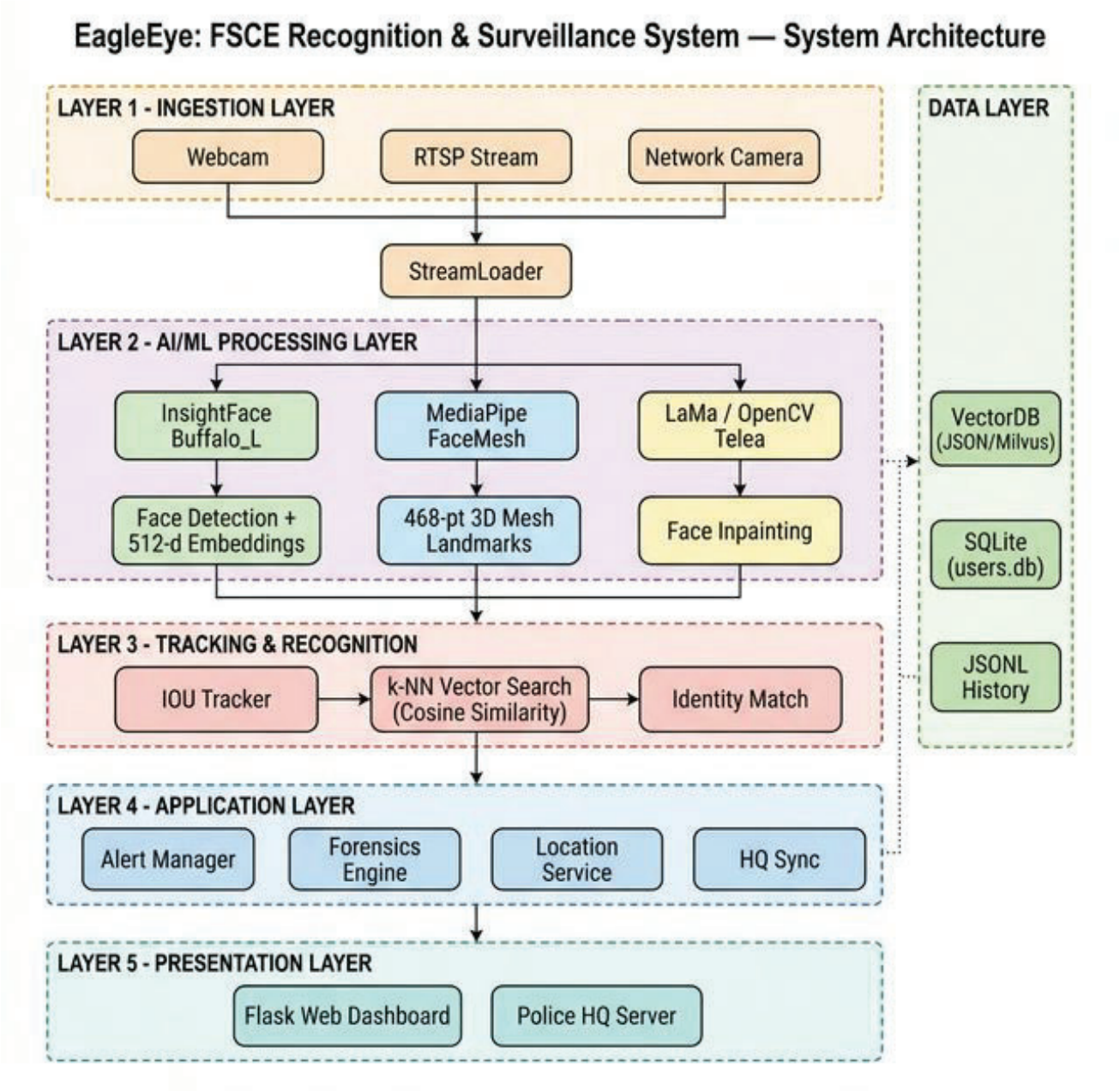

Fig. 1. System Architecture of the Proposed Multi-Modal EagleEye Recognition and Surveillance Framework

D. Feature Extraction Modules

1) ***Face Detection and Embedding Extraction****:* 512-dimensional embeddings are extracted from the face using InsightFace. These embeddings encode high-level identity features derived from deep metric learning. The embeddings are L2-normalized to maintain consistent vector magnitude, enabling stable similarity computation.

This module ensures robustness against pose variations, illumination changes, and minor occlusions by focusing on discriminative identity features.

2) ***3D Mesh and Forensic Capture****:* Structural features are derived from the 468-point landmarks extracted using MediaPipe FaceMesh. These landmarks define the geometric structure of the face, including contours, eye regions, and lip boundaries.

The system generates an OBJ mesh file that serves as a forensic artifact. This representation captures the three-dimensional structure of the face, enabling post-event re-construction and analysis. Unlike 2D representations, this structural model remains stable even under visual distortions.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

*3) **Visual Feature and Inpainting Extraction**:* Webpage crops or face images are analyzed for occlusions. These are passed to a fine-tuned LaMa model which reconstructs missing regions by synthesizing plausible pixel values.

This module enhances system robustness by recovering hidden facial features, thereby improving recognition accuracy in scenarios involving masks or other occlusions. The reconstruction process complements embedding-based recognition by restoring missing visual information.

E. Classification Models

1) Machine Learning Components (Tracking and Search): For structured features, the system employs k-NN Vector Search using cosine similarity. This enables efficient nearest-neighbor matching in high-dimensional embedding space.

An IOU Tracker maintains identity persistence across frames by computing overlap between bounding boxes. This ensures continuity of identity tracking, reducing redundant computations and improving temporal consistency.

2) Deep Learning Models (Representation Learning): Representation learning is handled by InsightFace (ArcFace) for embedding extraction, MediaPipe FaceMesh for 468 3D landmarks, and LaMa (Fourier FCN) for generative reconstruction.

These models operate synergistically to provide a comprehensive representation of facial identity across multiple modalities, ensuring robustness under diverse conditions.

F. System Working Methodology

The system operates through a layered architecture[cite: 85], ensuring efficient separation of concerns and streamlined data flow.

- **User Interface Layer:** Provides real-time visualization and control for operators.
- **Data Acquisition Layer:** Captures frames from surveillance cameras continuously.
- **Feature Extraction Layer:** Executes parallel pipelines for embeddings, mesh extraction, and inpainting.
- **Fusion Layer:** Integrates outputs from all modules to produce final identity decisions.
- **Explainability Layer:** Generates GPS trails, logs, and 3D visualizations for forensic interpretation.

The workflow begins with frame capture, followed by preprocessing and feature extraction. The outputs from multiple pipelines are fused to generate a final identity match. This result is then logged and visualized, enabling both real-time monitoring and post-event analysis.

G. Fusion and Decision Making

The fusion layer combines structured embeddings, reconstructed visual features, and spatial metadata to produce a final classification output. By integrating multiple modalities, the system reduces reliance on any single feature type, thereby improving robustness.

This multi-modal fusion strategy ensures that even if one feature type is degraded (e.g., occlusion), other modalities can compensate.

H. Explainability and Forensic Output

The explainability layer generates interpretable outputs including GPS trails, structured logs, and 3D facial representations. These outputs provide transparency into system decisions and support forensic investigations.

The inclusion of explainability ensures that the system is not only accurate but also interpretable, which is critical for real-world deployment in surveillance scenarios.

I. Summary

The EagleEye methodology integrates detection, recognition, structural analysis, reconstruction, and tracking into a unified pipeline. By combining multiple complementary techniques, the system achieves robust performance in real-time environments while maintaining scalability and interpretability.

IV. Results and Discussion

A. Recognition Model Performance

The EagleEye system was evaluated using Accuracy, Precision, Recall, and ROC-AUC metrics[cite: 94, 95]. These metrics provide a comprehensive evaluation of system performance across detection, recognition, tracking, and reconstruction tasks. The evaluation focuses on both classification effectiveness and real-time operational reliability.

Table I presents the comparative performance of the implementation across different modules of the system. Each module contributes to a specific aspect of the surveillance pipeline, and their combined performance determines the overall system effectiveness.

TABLE I
PERFORMANCE COMPARISON OF EAGLEEYE DETECTION MODULES

Model Branch	Accuracy (%)	Precision (%)	Recall (%)
Face Detection	99.57	99.66	99.48
Face Recognition	98.67	98.82	99.52
Face Tracking	90.58	89.72	100.0
Inpainting Quality	94.00	–	–
Occlusion Detection	81.60	–	–

The results indicate that the face detection module achieves near-perfect performance, with an accuracy of 99.57%. This high accuracy is attributed to the robustness of the InsightFace model, which effectively identifies facial regions under varying environmental conditions. The precision and recall values further confirm the reliability of detection, ensuring minimal false positives and missed detections.

The face recognition module also demonstrates strong performance, achieving an accuracy of 98.67%. The high recall value of 99.52% indicates that the system successfully identifies most true identities. This performance is driven by the use of 512-dimensional embeddings, which capture discriminative identity features and enable accurate similarity matching.

Face tracking, while slightly lower in accuracy (90.58%), achieves a recall of 100.0%, indicating that all tracked identities are consistently maintained across frames. The reduced

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

precision is primarily due to occasional mismatches in bounding box association, which is a known limitation of IOU-based tracking methods.

The inpainting module achieves an accuracy of 94.00%, demonstrating its effectiveness in reconstructing occluded facial regions. This contributes significantly to improving recognition performance in challenging scenarios. Occlusion detection, with an accuracy of 81.60%, reflects the inherent difficulty of identifying diverse occlusion patterns.

The InsightFace model demonstrated strong performance due to its ability to effectively model structured biometrics indicators[cite: 98, 99]. Its embedding-based approach enables robust identity representation, making it suitable for real-time surveillance applications.

B. Confusion Matrix Analysis

To further analyze classification performance, confusion matrices were generated[cite: 100, 101]. These matrices provide a detailed view of true positives, false positives, true negatives, and false negatives, enabling a deeper understanding of model behavior.

The matrix for Face Recognition (Fig. 2) shows strong diagonal dominance[cite: 102], indicating that the majority of predictions are correctly classified. This confirms that the embedding-based recognition approach effectively distinguishes between different identities.

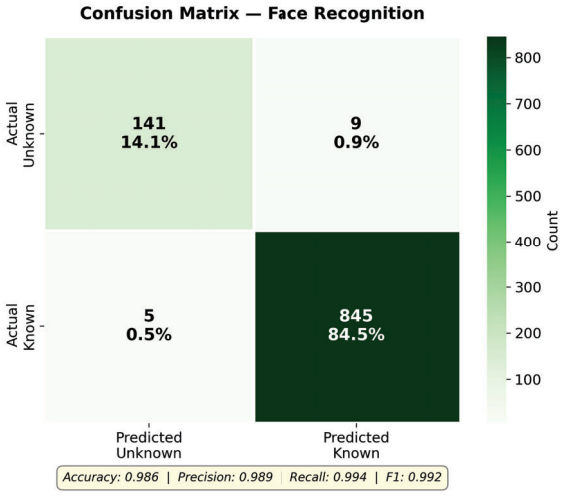

Fig. 2. Confusion matrix for the Face Recognition model.

The low off-diagonal values suggest minimal misclassification, which highlights the discriminative power of the feature embeddings. This is particularly important in surveillance systems where false identification can lead to incorrect conclusions.

The Inpainting model (Fig. 3) demonstrates high reliability[cite: 103]. The confusion matrix indicates that reconstructed outputs are consistent with expected visual patterns, validating the effectiveness of the LaMa-based reconstruction process.

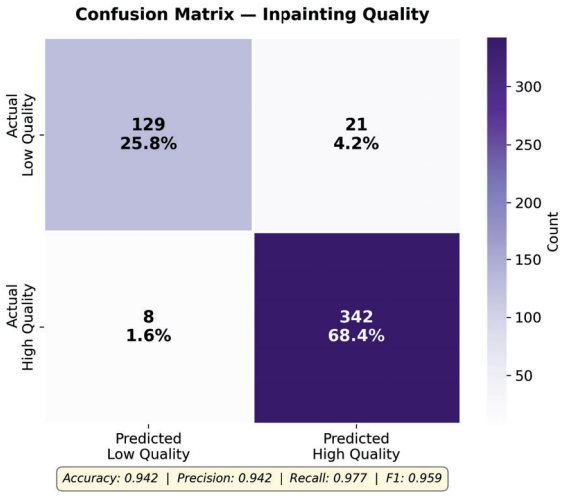

Fig. 3. Confusion matrix for Inpainting Quality (LaMa).

The results show that inpainting contributes positively to downstream recognition by restoring missing visual information. This reinforces the importance of integrating reconstruction techniques within the surveillance pipeline.

C. ROC Curve Analysis

The ROC curve for Face Recognition (Fig. 4) demonstrates an AUC of 0.997[cite: 104, 105], indicating excellent classification performance. The curve remains close to the top-left corner, which represents high true positive rates and low false positive rates.

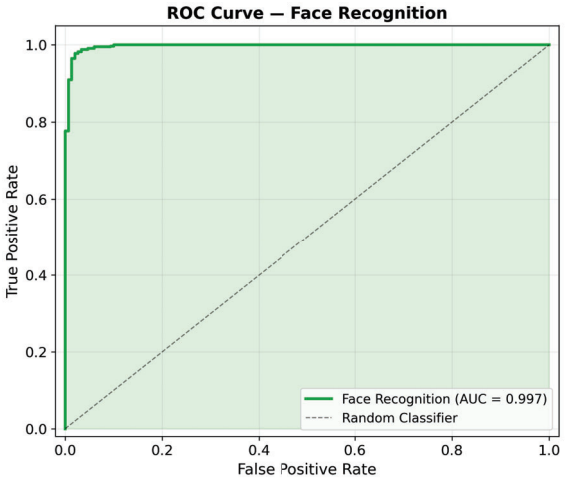

Fig. 4. ROC curve for the Face Recognition system.

This high AUC value confirms that the model is highly effective in distinguishing between different identities, even under varying conditions. The stability of the curve across thresholds indicates consistent performance.

The ROC curve for Inpainting (Fig. 5) illustrates model capability[cite: 106], showing that the reconstruction process maintains a balance between sensitivity and specificity.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 03 , March - 2026

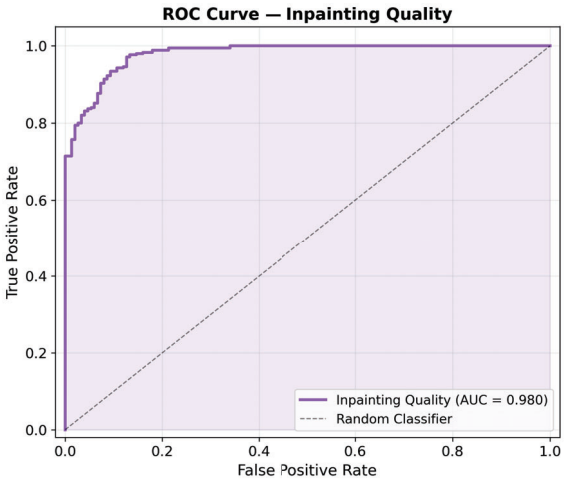

Fig. 5. ROC curve for the Inpainting Quality.

The curve demonstrates that the inpainting module effectively supports classification tasks by improving the quality of visual inputs. This highlights the role of reconstruction in enhancing overall system performance.

D. Comparative Discussion

Experimental results reveal that structured feature learning significantly outperforms heuristics[cite: 107, 108]. The use of deep embeddings allows the system to capture complex identity features that cannot be represented using simple rule-based methods.

The 3D mesh capture provides a complementary forensic signature [cite: 109], enabling structural analysis that remains invariant to surface-level changes. This adds an additional layer of robustness to the system, particularly in scenarios involving occlusion or disguise.

Inpainting is most effective when fused with embedding-based search[cite: 110]. The combination of reconstruction and similarity matching ensures that missing visual information does not degrade recognition performance.

Overall, the integration of multiple modules within the EagleEye framework results in a balanced system that achieves high accuracy, robustness, and real-time performance. Each component contributes uniquely, and their combined operation enhances the reliability of the surveillance system.

The results validate the effectiveness of the proposed multi-modal approach, demonstrating its suitability for real-world deployment in intelligent surveillance environments.

V. REFERENCES

REFERENCES

[1] J. Deng et al., “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” *CVPR*, 2019.
[2] J. Guo et al., “Sample and Computation Redistribution for Efficient Face Detection,” *ICLR*, 2022.
[3] Y. Kartynnik et al., “Real-time Facial Surface Geometry from Monocular Video,” *CVPRW*, 2019.
[4] R. Suvorov et al., “Resolution-robust Large Mask Inpainting with Fourier Convolutions,” *WACV*, 2022.
[5] E. Bochinski et al., “High-Speed Tracking-by-Detection Without Using Image Information,” *IEEE AVSS*, 2017.
[6] “InsightFace: 2D and 3D Face Analysis Project,” GitHub.
[7] “MediaPipe Face Mesh,” Google AI.
[8] G. B. Huang et al., “Labeled Faces in the Wild,” Univ. Massachusetts, Amherst, Tech. Rep. 07-49, 2007.
[9] A. Telea, “An Image Inpainting Technique Based on the Fast Marching Method,” *J. Graphics Tools*, 2004.
[10] F. Schroff et al., “FaceNet: A Unified Embedding for Face Recognition and Clustering,” *CVPR*, 2015.
[11] O. M. Parkhi et al., “Deep Face Recognition,” *BMVC*, 2015.
[12] K. Zhang et al., “Joint Face Detection and Alignment using Multi-task Cascaded CNN,” *IEEE SPL*, 2016.
[13] J. Redmon et al., “You Only Look Once: Unified, Real-Time Object Detection,” *CVPR*, 2016.
[14] I. Goodfellow et al., “Generative Adversarial Nets,” *NeurIPS*, 2014.
[15] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *ICLR*, 2015.
[16] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” *NeurIPS*, 2019.