technologies

MDPI

*Article*

# Combining MTCNN and Enhanced FaceNet with Adaptive Feature Fusion for Robust Face Recognition

**Sasan Karamizadeh [1,*], Saman Shojae Chaeikar [2] and Hamidreza Salarian [2]**

[1] Ershad Damavand Institute of Higher Education, Tehran 1416834311, Iran
[2] Department of Cybersecurity, Sydney International School of Technology and Commerce, Sydney, NSW 2000, Australia; samans@sistc.edu.au (S.S.C.); hamid@sistc.edu.au (H.S.)
* Correspondence: s.karamizadeh@e-damavandihe.ac.ir

**Abstract**

Face recognition systems typically face actual challenges like facial pose, illumination, occlusion, and ageing that significantly impact the recognition accuracy. In this paper, a robust face recognition system that uses Multi-task Cascaded Convolutional Networks (MTCNN) for face detection and face alignment with an enhanced FaceNet for facial embedding extraction is presented. The enhanced FaceNet uses attention mechanisms to achieve more discriminative facial embeddings, especially in challenging scenarios. In addition, an Adaptive Feature Fusion module synthetically combines identity-specific embeddings with context information such as pose, lighting, and presence of masks, hence enhancing robustness and accuracy. Training takes place using the CelebA dataset, and the test is conducted independently on LFW and IJB-C to enable subject-disjoint evaluation. CelebA has over 200,000 faces of 10,177 individuals, LFW consists of 13,000+ faces of 5749 individuals in unconstrained conditions, and IJB-C has 31,000 faces and 117,000 video frames with extreme pose and occlusion changes. The system introduced here achieves 99.6% on CelebA, 94.2% on LFW, and 91.5% on IJB-C and outperforms baselines such as simple MTCNN-FaceNet, AFF-Net, and state-of-the-art models such as ArcFace, CosFace, and AdaCos. These findings demonstrate that the proposed framework generalizes effectively between datasets and is resilient in real-world scenarios.

**Keywords:** face recognition; MTCNN; FaceNet; CelebA

check for updates

Academic Editors: Loris Nanni, Hamed Taherdoost and Yousef Farhaoui

Received: 19 August 2025
Revised: 24 September 2025
Accepted: 1 October 2025
Published: 3 October 2025

**Citation:** Karamizadeh, S.; Shojae Chaeikar, S.; Salarian, H. Combining MTCNN and Enhanced FaceNet with Adaptive Feature Fusion for Robust Face Recognition. *Technologies* **2025**, *13*, 450. https://doi.org/10.3390/technologies13100450

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Face recognition technologies have gained significance and popularity over the past few decades due to their extensive use in security, surveillance, and human–computer interaction [1]. Face recognition has advanced rapidly, with considerable research suggesting numerous ways to enhance its effectiveness. However, face images in multimedia applications, such as social networks, exhibit substantial variations in pose, lighting, and expression, which significantly undermine the performance of traditional algorithms [2].

In an era when facial recognition technology is central to security and human–computer interaction, the ongoing age-related changes in facial appearance pose a significant challenge [3]. Facial recognition technology is widely employed in various applications, such as time and attendance tracking, payment systems, and access control, offering substantial convenience [4].

Facial recognition technology continues to gain momentum, propelled by recent advances in deep learning and the creation of extensive training datasets [5]. However,

using facial recognition for authentication is complicated by real-world variations such as changes in pose, angle, lighting, and obstructions [6].

Facial recognition technologies, which utilize photos or videos, are prevalent in our daily lives. They can be used for security surveillance, access control, and security checks, and can even be combined with other biometric methods, such as fingerprinting and iris scanning. Google's FaceNet is a notable advance in facial recognition technology [7].

FaceNet's ability to generate small, discriminative embeddings is crucial in the feature extraction layer of the presented model. The use of triplet loss ensures that embeddings are optimized for similarity tasks, and the attention-augmented version enhances performance in the presence of occlusions, which aligns with the model's aim of robust face recognition [8,9].

We have generalized the introduction to include a brief consideration of the weakness of FaceNet. While FaceNet excels at learning discriminative and compact representations using triplet loss, it is vulnerable to occlusions and lacks context awareness. Specifically, it is insensitive to all facial regions being the same, which hurts performance when portions of the face are occluded (e.g., by masks or lighting). These constraints drive the improved version in this work, where attention mechanisms enable the network to focus on informative and non-occluded areas, including the eyes, thereby enhancing unconstrained robustness.

In the proposed model, the fusion layer learns weights dynamically based on contextual signals (e.g., greater reliance on the eye regions for masked faces), using either a transformer-based or attention-based mechanism to achieve robustness.

Despite the progress of deep learning methods, existing face recognition frameworks face several persistent challenges. Models such as FaceNet produce compact and discriminative embeddings but are highly sensitive to occlusion and illumination changes. CNN-based enhancements improve accuracy but often fail to capture contextual cues such as head pose or environmental lighting. Transformer-based methods achieve strong performance but come with significant computational overhead, limiting their deployment on real-time and edge systems. Furthermore, most existing hybrid approaches do not explicitly integrate contextual signals into the embedding space, reducing robustness in unconstrained scenarios. Therefore, this paper aims to design a hybrid model that addresses these limitations by combining efficient face detection and alignment, attention-augmented embedding generation, and adaptive fusion of contextual features, thereby improving recognition accuracy and robustness under challenging real-world conditions.

The following sections will elaborate on relevant literature, experimental methods used, results obtained, and the approach taken to the conclusion.

## 2. Literature Review

This section discusses recent advancements in face recognition, including CNN-based methods, attention mechanisms, and transformer-inspired fusion techniques. Further technical principles of the MTCNN, FaceNet, and Adaptive Feature Fusion will be described in the Section 3.

Recent advances in face recognition include large-margin embedding methods, such as ArcFace, CosFace, and AdaCos, which enhance discriminability through angular margin losses. Transformer-based models are introduced for global attention modeling, while lightweight attention modules help with occlusion-robustness. Yet these methods often ignore contextual signals (pose, lighting, masks), limiting their real-world generalization. Our method differs by combining attention-augmented embeddings and adaptive feature fusion of contextual vectors to achieve a clean balance of accuracy, robustness, and computational efficiency [10,11].

### 2.1. Related Works

Kortli et al. [12] argued that face recognition systems have become popular due to their diverse applications in security, surveillance, and human–computer interaction. These systems can be trained to identify individuals with high accuracy.

Ding and Tao [13] proposed a comprehensive deep learning model for learning face representations from multimodal information. The proposed deep learning structure includes well-designed convolutional neural networks (CNNs) and a three-layer stacked autoencoder (SAE). The model implements a two-step solution for an identification system based on MTCNN and FaceNet networks, with estimation of the user's head pose. The model's accuracy ranges from 92% to 95%.

Li [14] proposed a model with an attention mechanism, feature fusion, and self-attention to address masked face recognition. The author designed four lightweight modules to fine-tune the network structure, thereby addressing issues of volatile recognition accuracy and low generalization ability in the model.

Jia and Tian [15] introduced a new method that combines the FaceNet deep learning algorithm with MTCNN to achieve robust and accurate face recognition across all age ranges. It leverages FaceNet's ability to extract unique features from facial images and projects them into a high-dimensional feature space for effective face matching. The system employs MTCNN as a pre-processing step to detect and align faces accurately, thereby effectively addressing age-related variations in facial appearance geometry [15].

Qi et al. [16] designed a face recognition model based on MTCNN and FaceNet. Traditional face recognition systems often suffer from manual feature settings, which result in low recognition accuracy and slow processing speed. The MTCNN model comprises three convolutional neural network layers: P-Net, R-Net, and O-Net, which are designed to detect faces in images.

In another work, Wen et al. [17] designed a new deep learning API that combines the strengths of MTCNN and FaceNet to address the limitations. Building on MTCNN and FaceNet, they proposed an API that verifies users' identities through a two-stage verification process.

Yang et al. [18] devised an enhanced face recognition model based on MTCNN and the integrated application of FaceNet and the Local Binary Pattern (LBP) method to improve illumination robustness effectively.

Hu et al. [19] invented a face recognition system based on MTCNN for facial detection and feature extraction, utilizing FaceNet and SVM for classification and recognition. The algorithm follows a five-stage process: (i) data preparation for training, (ii) face detection from the data using MTCNN, (iii) feature extraction from each face using the FaceNet Keras model, (iv) classification of feature vectors with SVM, and (v) face recognition.

Another MTCNN-based solution has been designed by Ku and Dong [20], which is an improved convolutional neural network for face detection, offering more stable performance against lighting, angle, and facial expression variations in real-world scenarios.

In research conducted by Abdul et al. [21], the authors proposed a hybrid deep neural network for face recognition under poor weather conditions using MobileNet and attention mechanisms. The model improves performance on the Yale Face Dataset and occlusion tolerance on the Simulated Masked Yale Dataset.

### 2.2. Traditional Face Recognition Methods

A defect in traditional methods is that custom features had to be manually assigned. Statistical techniques were then applied to these features. Some methods developed early in facial recognition involved Eigenfaces and Fisherfaces, which used dimensionality reduction tools like PCA and LDA to project face images onto a lower-dimensional space

for classification [22]. Methods such as Local Binary Patterns encode texture information by comparing pixel intensities within a local area [23]. Although these methods were innovative at the time, they were limited by relying on human-understood manual features, making them highly sensitive to changes in lighting, pose, and expression. As a result, they could not be used in real-world applications without strict conditions.

### 2.3. Deep Learning-Based Face Recognition

The advent of deep learning, especially Convolutional Neural Networks (CNNs), marked a significant milestone. CNNs can automatically capture fine details from raw pixels, surpassing their predecessors [24]. A notable milestone was the development of FaceNet by Schroff et al. [25]. FaceNet does not directly classify a face but embeds a face image into a compact space where the distance between them represents the similarity between two faces. It is trained with a triplet loss function that reduces the distance from an anchor to a positive (same identity) and increases the distance to a negative (different identity). This approach, along with other deep metric learning frameworks, has set a new standard for face recognition performance and scalability.

### 2.4. Recent State-of-the-Art Methods

Following FaceNet, efforts have been made to improve the loss functions to learn more discriminative embeddings by increasing between-class variance and reducing within-class variance.

Wang et al. [26] in CosFace introduced Large Margin Cosine Loss (LMCL), which L2-normalizes weights and features to remove radial variations and then uses a cosine margin term further to increase the decision margin in the angular space.

Deng et al. [27] in ArcFace introduced Additive Angular Margin Loss, where a geodesic distance margin is added directly to the angle between the target weight and the deep feature. ArcFace performs exceptionally well on most benchmarks due to its straightforward geometric reasoning and efficiency.

Following their success in natural language processing, Transformers have been applied to computer vision. Vision Transformers (ViTs) [28] treat an image as a sequence of patches, using self-attention mechanisms to capture global relationships. Recent research, such as [29], has demonstrated that ViTs can achieve extremely competitive performance in FR, as the self-attention mechanism can potentially adaptively focus on the most discriminative facial parts.

### 2.5. Works on Occlusion and Robustness

In real-world situations, managing occlusions (such as masks and sunglasses) and other less-than-ideal circumstances presents a substantial difficulty for FR. A number of strategies have been put out to deal with this problem.

AFF-Net by Li [30] explicitly suggested a network for masked face recognition that is based on attention and feature fusion. The model shows a targeted approach to a particular form of occlusion by using attention modules to focus on non-occluded regions and fusing features to increase resilience.

Other approaches are also included in the literature, such as feature fusion methods for merging data from several sources [11], attention mechanisms to weigh feature importance [10], and generative models for recreating occluded regions [31]. This adaptive feature processing technique has been expanded in our work. Nevertheless, it sets itself apart with a cohesive framework that uses a transformer-based fusion module to dynamically merge identity embeddings with a wide range of contextual variables (pose, illumination, and occlusion).

Our method is at the intersection of such advances. We leverage the robust detection of MTCNN and the robust embedding learning of FaceNet. We supplement this backbone with an attention mechanism to better extract features in the presence of occlusion, in the spirit of [8,10]. Our core contribution, the Adaptive Feature Fusion module, is more nuanced than static fusion or reconstruction. It synergistically integrates visually embedded and heuristic contextual cues, in light of the adaptive character of transformer models [7], to achieve robustness to a broader range of real-world challenges, including pose, lighting, and partial occlusion.

## 3. Research Methodology

There are five essential steps in our model approach: Normalization and CLAHE (Contrast Limited Adaptive Histogram Equalization) are used for preprocessing; a five-point landmark-based similarity transform is used for face detection and alignment with MTCNN; an attention-enhanced FaceNet backbone based on Incep-tion-ResNet is used for extracting embeddings, resulting in a fixed 128-D representation; a lightweight transformer encoder is used for adaptive feature fusion of embeddings and contextual vectors; and cosine similarity with adaptive thresholding is used for classification. We used semi-hard mining and triplet loss (margin = 0.2) for training. Unless otherwise noted, we only utilized LFW and IJB-C for testing, with stringent subject dis-jointness, and Celeba for training and validation.

This section describes our proposed model for creating deep learning techniques for facial image classification.

Face recognition begins with the input of a raw RGB face image, typically from a file or a camera (e.g., JPEG, PNG). The image is validated before preprocessing, where it is resized, normalized, and optionally enhanced to meet model requirements such as MTCNN and FaceNet. The MTCNN module detects and aligns faces by identifying key landmarks (such as eyes, nose, and mouth corners), enabling the system to crop and align the face accordingly. This preprocessing step minimizes variability in input data, enhancing recognition accuracy. After alignment, the latest FaceNet version extracts a 128-D embedding that captures the face's unique features. The improved model utilizes an attention mechanism that focuses on unoccluded and informative parts, such as the eyes, particularly in cases of occlusion (e.g., masks). The embedding combines contextual information, such as head position or lighting, through an adaptive fusion layer that uses transformer-based attention to assign importance to features. The combined representation is compared with stored embeddings using similarity measures to identify or verify identities. Finally, low-confidence matches are filtered out during post-processing, and the resulting output is mapped and sent to the user or system interface as a label or verification result, completing the recognition process.

We describe how contextual features are acquired by combining heuristics and side processing. Head orientation (yaw, pitch) is estimated through geometric correlations between the eyes and nose based on facial landmarks identified by MTCNN. Lighting is determined by the average intensity of gray regions on the face, and masks are identified using a binary classifier trained on datasets of masked faces. These components are integrated into a 10-dimensional context vector used during the fusion process.

### *3.1. Face Detection and Alignment with MTCNN*

The Multi-Task Cascaded Convolutional Neural Network (MTCNN) [8] is a popular method for face detection and alignment, and is essential in face recognition system preprocessing. MTCNN operates through a three-stage cascade to identify faces and facial landmarks, remaining effective despite changes in scale, pose, and lighting. Its high accuracy and speed have made it a preferred choice for facial detection in real-world

applications, such as the proposed hybrid model, where it functions as the first step for face localization and alignment before feature extraction. It employs a three-stage cascaded architecture, illustrated in Figure 1, where each stage consists of a dedicated CNN that progressively refines the results from the previous stage. This design enables efficient processing by quickly rejecting non-face regions in the early stages, while devoting more computation to promising candidates.

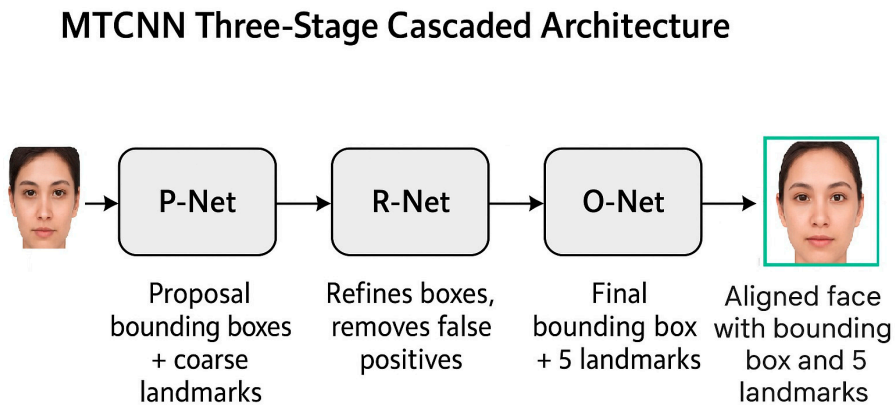



**Figure 1.** MTCNN Three-Stage Cascaded Architecture. The face image has been generated with ChatGPT version 5 using the prompt "Generate a sample synthetic female face with neutral expression, straightforward shot, white background.".

MTCNN consists of three sub-networks: the Proposal Network (P-Net), the Refine Network (R-Net), and the Output Network (O-Net), which progressively enhance face detection and landmark localization.

- Stage 1: Proposal Network (P-Net): A shallow, fully convolutional net run on a pyramid of the input image at many scales. It must quickly produce many candidate face bounding boxes and provide a first, coarse estimate of facial landmarks. It outputs a confidence score for each candidate and applies Non-Maximum Suppression (NMS) to combine highly overlapping detections.
- Stage 2: Refine Network (R-Net): Each P-Net candidate proposal is warped to a fixed size and passed through this more advanced CNN. The general aim of R-Net is to screen out much of the P-Net candidate's false positives and further refine the bounding box coordinates through bounding box regression.
- Stage 3: Output Network (O-Net): This is the terminal and final, most complex network of the cascade. It receives the fine-tuned candidates from R-Net, warps them, and performs a more thorough analysis. The O-Net generates the final correct bounding box, a confidence measure, and the exact locations of five facial landmarks (the four corners of both eyes and the nose tip). These coordinates are then used to do a similarity transformation to position and crop the face, a necessary operation for normalizing input to the subsequent FaceNet model.

Candidate window classification loss is computed by employing a cross-entropy loss function, as shown in Equation (1).

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}[Z_i \ln(S_i) + (1 - Z_i)\ln(1 - S_i)] \quad (1)$$

where

$N$: Number of candidate windows,
$Z_i \in \{0, 1\}$ : Ground-truth label (1 for face, 0 for non-face)
$S_i \in [0, 1]$ : Predicted probability.

For bounding box regression, MTCNN minimizes the Euclidean distance between the predicted and ground-truth box coordinates using Equation (2).

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{i=1}^{N}(C_i - \hat{C}_i)^{\mathsf{T}}(C_i - \hat{C}_i) \quad (2)$$

where

$C_i = [x_i, y_i, w_i, h_i]^{\mathsf{T}}$: Ground-truth coordinates,
$\hat{C}_i = \left[\hat{x}_i, \hat{y}_i, \hat{w}_i, \hat{h}_i\right]^{\mathsf{T}}$: Predicted coordinates.

Facial landmark localization utilizes a similar regression loss that is presented in Equation (3).

$$\mathcal{L}_{lmk} = \frac{1}{N}\sum_{i=1}^{N}(M_i - \hat{M}_i)^{\mathsf{T}}(M_i - \hat{M}_i) \quad (3)$$

where

$M_i = [m_{i1}, m_{i2}, \ldots, m_{i10}]^{\mathsf{T}}$: Ground-truth coordinates,
$\hat{M}_i = [\hat{m_{i1}}, \hat{m_{i2}}, \ldots, \hat{m_{i10}}]^{\mathsf{T}}$: Predicted coordinates.

The total loss for each network combines these tasks with the weighted contributions using Equation (4).

$$\mathcal{L} = \beta_{cls}\mathcal{L}_{cls} + \beta_{reg}\mathcal{L}_{reg} + \beta_{lmk}\mathcal{L}_{lmk} \quad (4)$$

$\beta_{cls}, \beta_{reg}, \beta_{lmk}$: Task weights (e.g., 1, 0.5, 0.5 in O-Net).

### *3.2. Feature Extraction with Enhanced FaceNet*

FaceNet, introduced by Schroff et al. [9], is a deep face recognition framework that generates compact, discriminative facial embeddings, enabling efficient identification and verification. Unlike conventional classification-based approaches, FaceNet directly learns a mapping from face images to 128-D Euclidean space, where the distance between embeddings reflects face similarity. Its triplet loss and deep convolutional structures have established a benchmark for face recognition and act as the key feature extraction component in the proposed hybrid method, enhanced by incorporating attention mechanisms. FaceNet utilizes a deep CNN (e.g., Inception-ResNet) to process aligned face images (160 × 160 pixels) and produce a 128-D embedding. As shown in Figure 2, the standard FaceNet architecture we build upon consists of:

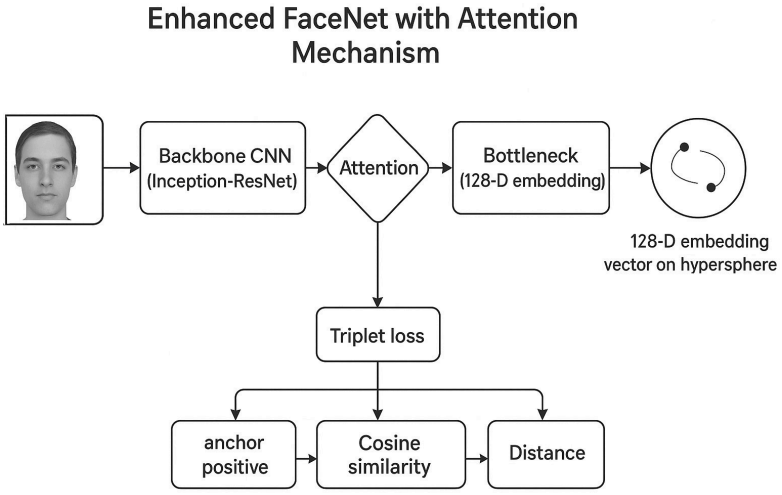



**Figure 2.** Enhanced FaceNet with Attention Mechanism. The face image has been generated with ChatGPT version 5 using the prompt "Generate a sample synthetic male face with neutral expression, straightforward shot, white background.".

Backbone CNN: The feature extractor is an Inception-ResNet-v1 architecture that processes the aligned 160 × 160 RGB input face.

Bottleneck Layer: A completely connected layer compresses the high-dimensional features from the backbone into a 128-dimensional vector.

L2 Normalization: To project this 128-D vector onto a unit hypersphere, it is L2-normalized (Equation (6)). For the next similarity test, this normalization is essential. Triplet Loss: The triplet loss function (Equation (5)) is used to train the model. Triplets of images are utilized during training: a negative (an image of a different person), a positive (an additional image of the same person), and an anchor (a reference image of a person). To guarantee that embeddings of the same identity are grouped in the feature space, the loss maximizes the distance between the anchor and negative embeddings and decreases the distance between the anchor and positive embeddings. FaceNet's training is driven by the triplet loss function, which optimizes the embedding space using Equation (5).

$$\mathcal{T} = \sum_{i=}^{K} \max\left(\left(E_{a,i} - E_{p,i}\right)^{\intercal}\left(E_{a,i} - E_{p,i}\right) - \left(E_{a,i} - E_{n,i}\right)^{\intercal}\left(E_{a,i} - E_{p,i}\right) + \gamma, 0\right) \quad (5)$$

where

$K$: Number of triplets,

$E_{a,i}$,$E_{p,i}$, $E_{n,i} \in \mathbb{R}^{128}$: Anchor, positive, negative embeddings.

$\gamma$: Margin (e.g., 0.2).

The embedding vectors are L2-normalized in Equation (6) to lie on a unit hypersphere.

$$E_j = \frac{G_j}{\sqrt{G_j^{\intercal} G_j}} \quad (6)$$

where

$G_j \in \mathbb{R}^{128}$ : Raw embedding,

$E_j \in \mathbb{R}^{128}$: Normalized embedding,

During inference, face similarity is computed using cosine similarity or Euclidean distance in Equation (7).

$$Sim_{Cos}(E_1, E_2) = E_1^{\intercal} E_2 Dist_{euc}(E_1, E_2) = \sqrt{(E_1 - E_2)^{\intercal}(E_1 - E_2)} \quad (7)$$

where

$E_1, E_2$: Normalized embeddings (unit length, so cosine similarity simplifies).

While powerful, the standard FaceNet model treats all facial regions equally. This makes it vulnerable to occlusions (e.g., masks, sunglasses) or extreme pose variations where critical features are hidden. To address this limitation, we enhance the FaceNet architecture by integrating a soft attention mechanism after the convolutional feature maps, as illustrated in Figure 2. The attention mechanism works as follows:

All facial regions are treated identically by the conventional FaceNet model, notwithstanding its power. Because of this, it is susceptible to occlusions (such as masks or sunglasses) or drastic changes in posture that obscure important characteristics. As shown in Figure 2, we improve the FaceNet architecture to overcome this constraint by adding a soft attention mechanism after the convolutional feature maps. The following is how the attention mechanism operates:

- Feature extraction: The input face's high-level features are represented by a series of intermediate feature maps that are first produced by the backbone CNN.
- Attention Gate: These feature maps are processed by a little sub-network (such as a 1x1 convolutional layer with sigmoid activation) to produce an attention weight map. With only one channel, this map has the same spatial dimensions as the feature maps.

The network's learnt importance of the corresponding spatial region in the feature maps for identity recognition is represented by each value in this weight map, which ranges from 0 to 1.

- Feature Recalibration: The attention weight map multiplies the initial feature maps element-by-element. This procedure suppresses features from occluded or uninformative parts (e.g., a masked mouth, backdrop) and amplifies features from informative, non-occluded regions (e.g., eyes, brow shape).
- Attended Feature Pooling: To produce the final 128-D embedding, these "attended" or "re-weighted" feature maps are subsequently run through the usual bottleneck layer and global average pooling.

With a particular input, this attention-augmented version enables the network to dynamically focus on the facial features that are the most discriminative. The network will base its embedding on the eye and forehead regions since, for example, the attention weights for these areas will be high if a mask obscures the bottom part of the face. One of the main contributions of our work is that it makes the feature extraction method far more resilient to real-world problems. The attention gate is trained end-to-end by the gradients from the triplet loss, guaranteeing that the learned attention is ideal for the recognition challenge.

### *3.3. Adaptive Feature Fusion*

Adaptive feature fusion is a high-level technique used in modern face recognition models. To build multiple feature representations—deep embeddings and context hints—into a single robust descriptor, this layer adaptively combines 128-dimensional FaceNet embeddings with context information (e.g., head pose, lighting conditions, or occlusion hints such as mask presence) for enhanced accuracy in challenging scenarios. Unlike static fusion methods (e.g., concatenation), adaptive fusion employs learned or condition-weighting policies, often leveraging attention-based or transformer-inspired architectures to focus on significant features based on input conditions. This method draws inspiration from the early work in attention-based fusion by [10,11] on transformers to make the model more responsive to real-world variations, such as partial occlusions or changes in lighting. Adaptive feature fusion in face recognition typically includes the following three processes:

Feature Extraction: Two or more features are gathered and prepared. The model considers the 128-D FaceNet embedding (facial identity extraction) and context features (e.g., a 12-D vector representing pose angles, light intensity, or binary mask indicators), which are typically extracted using auxiliary networks or heuristics.

Fusion Strategy: A fusion mechanism incorporates these characteristics. Common approaches are:

- Weighted Concatenation: Features are concatenated with learned weights, scaled based on input conditions.
- Attention Mechanisms: Cross-attention or self-attention modules (transformer-inspired) assign higher weights to informative features (e.g., unoccluded regions).
- Transformer-Based Fusion: A transformer encoder processes feature sets, models interactions, and produces a fused representation.

Normalization of Output: The combined feature vector is normalized (e.g., L2 normalization) to match the requirements of downstream applications like classification or verification.

Adaptive feature fusion usually involves weighted or attention-based combinations, with the equations differing depending on the specific method. For an attention-based combination method, the process can be explained as follows:

- Attention Scores: Given two feature vectors—FaceNet embedding in Equation (8).

$$A = softmax\left(\frac{QK^{\intercal}}{\sqrt{d}}\right) \tag{8}$$

where
$Q = W_Q V_e \in \mathbb{R}^{128}$ : FaceNet embedding,
$K = W_k V_c, \in V_c\mathbb{R}^{dc}$ contextual features (e.g., $d_c = 10$),
$W_Q \in \mathbb{R}^{d\times 128}$ , $W_k \in \mathbb{R}^{d\times d_c}$ : Projection matrices,
*d:* Attention dimension (e.g., 64),
$A \in \mathbb{R}^{1\times d_c}$: Attention weights.

- Weighted Fusion: The contextual features are weighted and combined with the embedding in Equation (9).

$$V_f = \begin{bmatrix} V_e \\ AW_v V_c \end{bmatrix} \tag{9}$$

where
$W_v \in \mathbb{R}^{d_v\times d_c}$: Value projection matrix,
$V_f \in \mathbb{R}^{128+d_v}$: Fused vector.

- Normalization: The fused vector is normalized for downstream tasks, as shown in Equation (10).

$$V_f' = \frac{V_f}{\sqrt{V_f^{\intercal} V_f}} \tag{10}$$

where
$V_f'$: Normalized fused vector.

Alternatively, the features are input into a transformer encoder for transformer-based fusion, where multi-head self-attention models interact, as shown in Equation (11).

$$MH(Q,K,V) = \begin{bmatrix} head_1 \\ . \\ . \\ . \\ head_h \end{bmatrix} W_o head_i = Attention\left(QW_{Qi}, KW_{K,i}, VW_{v,i}\right) \tag{11}$$

where
*Q, K, V*: Derived from concatenated $V_e, V_c$,
*h*: Number of attention heads,
$W_o$: Output projection matrix.

The attention score $V_e$ calculates contextual features that are more important than the FaceNet embedding, which directs attention toward informative cues (e.g., pose instead of lighting when pose variation is high). The Softmax ensures that weights and scaling factors sum to 1.

### *3.4. Dataset*

To ensure generalization, CelebA [32] was used only for training/validation, while independent evaluations were performed on LFW [33] and IJB-C [34] with no identity overlap. As shown in Table 1, three datasets are utilized to evaluate the proposed model.

**Table 1.** Comparison of Face Recognition Datasets.

| Dataset | No. of Images | No. of Identities | Pose & Illumination Variability | Occlusion | Real-World Conditions |
|---|---|---|---|---|---|
| CelebA | 200,000+ | 10,177 | Moderate | Low | Low |
| LFW | 13,000+ | 5749 | Low to moderate | Low | Moderate |
| IJB-C | 31,300+ images and 117,000+ frames | 3531 | High | High | High |

Table 1 compares three large-scale face recognition datasets—CelebA, LFW, and IJB-C—across the critical dimensions influencing generalization and model robustness. While the CelebA dataset is large and rich in labeled attributes and landmarks, it has limited variation in pose, lighting, and real-world occlusions. Thus, it is most suitable for pretraining or attribute learning, but less ideal for evaluating deployment in real-world scenarios. In contrast, LFW exhibits medium variability and was one of the early benchmarks for face verification in natural settings. However, it suffers from relatively low age and race variation and has become less challenging for modern deep learning systems. The IJB-C dataset, on the other hand, is the most complex, encompassing still images and video frames with significant variations in occlusion, pose, and environment. To ensure comprehensive verification, CelebA-trained models must be tested against more challenging datasets, such as IJB-C or cross-age datasets like CACD, to properly assess their generalizability in real-world applications.

This is also notable that ChatGPT version 5 on 16 September 2025 is used for creating the faces used in Figures 1 and 2.

### *3.5. Implementation Details*

Results are presented as mean ± standard deviation across five random seeds. The network was trained for 100 epochs using the Adam optimizer ($lr$ = 0.001, weight decay = $1 \times 10^{-5}$), using a step decay schedule (0.1× at epochs 40, 60, and 80). Data augmentations included color jitter, random cropping, and horizontal flipping. To prevent overfitting, dropout ($p$ = 0.5) and early stopping based on validation loss were employed.

## 4. Proposed Model

The face recognition process starts with acquiring a raw RGB face image, typically obtained from a camera or file (e.g., JPEG, PNG).

The model takes an input image and produces an output image. It pre-processes the input by normalizing pixel values, adjusting lighting with CLAHE, and creating multi-scale pyramids for different face sizes. It detects and aligns faces using either MTCNN (light) or RetinaFace (high precision) and crops facial landmarks at five points for accurate alignment. MTCNN + Enhanced FaceNet are referenced as alternative models for comparison, but are not part of the final method. Our backbone remains an attention-enhanced FaceNet with adaptive fusion. The fusion module uses a 4-head Transformer Encoder to dynamically weight identity and contextual features. Each detected face goes through the ArcFace-ResNet100 model with Squeeze-Excitation attention blocks to generate unique 128-dimensional embeddings. When multiple faces are present, a 4-head Transformer Encoder concatenates these embeddings without losing context relationships. The recognition phase performs either closed-set identification (via Softmax classification) or open-set verification (by cosine similarity with adaptive thresholding). Finally, post-processing involves non-maximum suppression (IoU 0.7) and confidence calibration before outputting

the annotated image, which includes bounding boxes, identity labels, and confidence scores. As shown in Figure 3, the model optimizes raw images for thoroughly analyzed outputs through a series of deep learning steps, balancing accuracy, computational complexity, and robustness against different lighting conditions, scales, and face counts.

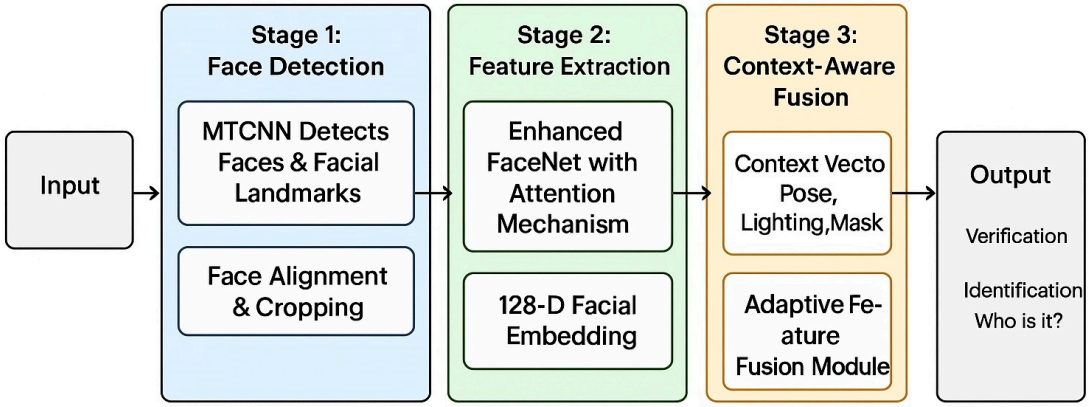



**Figure 3.** Architecture of the proposed model.

Figure 3 illustrates the overall model of the proposed facial recognition system. An input image is used as the starting point, and it is processed in three main steps:

Detection and Alignment of Faces: First, the MTCNN model is applied to the input image [8]. The five prominent facial landmarks—the corners of the mouth, nose, and eyes—as well as the face's bounding box are detected using MTCNN. A similarity transformation is then applied using this geometric information, producing a cropped and aligned face image that has been normalized for rotation and scale. In order to minimize variability before feature extraction, this step is essential.

Feature Extraction: Our Enhanced FaceNet model processes the aligned face crop. Our improved version, which is based on the conventional FaceNet architecture [9], adds an attention mechanism to its convolutional layers. This enables the network to dynamically weight the significance of various facial regions, giving less attention to hidden or irrelevant areas and more attention to distinguishing, non-occluded characteristics (such as the eye region when a mask is worn). A compact 128-dimensional embedding vector representing the face identity is the stage's output.

Context-Aware Adaptive Fusion: The Adaptive Feature Fusion module then fuses a 12-dimensional context vector with the 128-D face embedding. Important details about the face's surroundings, such as head posture (yaw, pitch, roll), lighting conditions (global intensity and contrast), and the existence of occlusions like masks, are contained in this context vector, which was heuristically recovered from the original MTCNN output. This context is used by the fusion module, which is constructed using a lightweight transformer encoder, to intelligently balance the significance of various face embedding elements. For instance, the fusion module will learn to rely more on features from the top half of the face if the context vector indicates the presence of a mask.

Output: One of two tasks uses the final, context-aware feature vector:

Verification: The system uses cosine similarity to compare the vector to a reference template. The faces are confirmed to belong to the same identity if the similarity score rises beyond a predetermined level.

Identification: A database gallery of templates is compared to the vector. The recognition result is the identity linked to the template that is most similar.

The fusion layer is achieved using a lightweight transformer encoder, which comprises one transformer block, four attention heads, and a key dimension of 64. Contextual features

are concatenated with the FaceNet embeddings and fed into a self-attention mechanism, where the weights are learned during training. Attention weights are acquired through supervised learning using identity labels, while contextual features such as illumination and pose are sourced from auxiliary heuristics. During runtime, the fusion module incurs <5 ms overhead per image on an NVIDIA RTX 3090 (Fabricated at Samsung Semiconductor Manufacturing Plant, Hwaseong, Republic of Korea), with total inference time $\approx$ 35 ms per face, and parameter count $\approx$ 26 M. This makes the model suitable for near real-time applications.

This paper presents several notable enhancements that improve facial recognition models. To ensure that only the accurate, standardized face areas are sent to the Enhanced FaceNet layer for processing, MTCNNs pre-process input images in the suggested hybrid model by identifying and aligning faces. Its multi-task learning technique and cascaded design make it computationally effective, robust, and able to handle obstacles like partial face occlusions and unequal face sizes, which are essential for real-world face recognition.

## 5. Evaluation Metrics

Several standard performance metrics have been used to evaluate the proposed system's face recognition capabilities, including accuracy, precision, recall, and F-measure. These metrics provide a comprehensive understanding of how effectively the system can identify and verify individuals accurately.

Accuracy is the ratio of correctly classified instances (true positives and negatives) to the total number of cases—Equation (12).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

where

- True Positive (*TP*): Correctly verifying a pair of images as the same identity.
- True Negative (*TN*): Correctly rejecting a pair of images as different identities.
- False Positive (*FP*): Incorrectly accepting a pair of different identities as the same (False Accept).
- False Negative (*FN*): Incorrectly rejecting a pair of the same identity as different (False Reject).

Precision refers to the percentage of accurate positive results (correct matches) out of all instances the system flagged as positive, according to Equation (13).

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

High precision suggests low false positives, crucial when minimizing false acceptance.

Recall (also referred to as Sensitivity or True Positive Rate) captures the percentage of true positives identified correctly—Equation (14).

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

High recall ensures the system does not overlook many actual matches, which is crucial for security and surveillance applications.

F-measure (F1-Score) represents the harmonic mean of recall and precision, offering a single value that balances both issues—Equation (15).

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{15}$$

F1-Score is invaluable in cases where the class distribution is unbalanced or when considering false negatives and positives.

These metrics provide a solid foundation for assessing the performance of face recognition systems, offering a balanced measure of accuracy and reliability across various conditions.

Several standard performance metrics have been used to evaluate the proposed system's face recognition capabilities. These metrics are categorized based on the task:

- For face verification (1:1 matching), which determines whether two face images belong to the same person, as mentioned in Equations (16) and (17).

$$\text{True Acceptance Rate (TAR)} = \frac{Number\ of\ genuine\ accepts}{Total\ number\ of\ genuine\ attempts} \quad (16)$$

$$\text{False Acceptance Rate (FAR)} = \frac{Number\ of\ impostor\ accepts}{Total\ number\ of\ impostor\ attempts} \quad (17)$$

- For face identification (1:N matching), which identifies a person from a gallery, as we used.

Accuracy, Precision, Recall, and F1-score (as defined in Equations (12)–(15)).

Cumulative Match Characteristic (CMC): Rank-based accuracy for open-set identification.

## 6. Results

The results of the proposed face recognition model—which combines MTCNN, Enhanced FaceNet, and Adaptive Feature Fusion—are demonstrated in Figures 4–14. The figures illustrate the model's ability to accurately detect, align, and recognize faces despite occlusions, lighting variations, and pose changes. The proposed model demonstrates high recognition confidence and stability across diverse inputs, underscoring the effectiveness of the attention-driven embedding and dynamic fusion approach. The results also highlight the model's improved performance compared to conventional MTCNN-FaceNet methods, especially in real-world scenarios where environmental factors present challenging conditions.

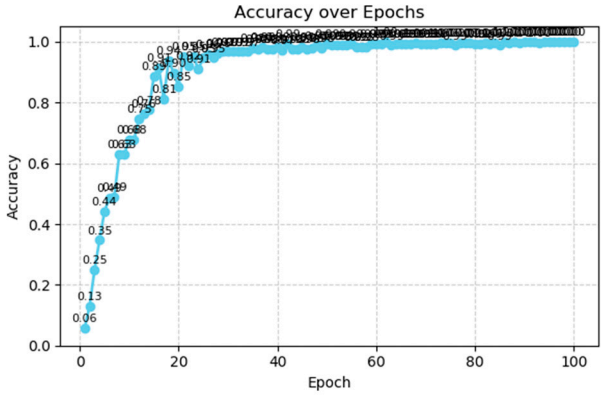



**Figure 4.** The Training/Validation accuracy of the proposed face recognition model. The blue line represents training accuracy, and the black line represents the validation accuracy across 100 epochs.

As demonstrated in Figure 4, the model's accuracy improves as the epochs increase to 100, starting from a low 38% at epoch zero and gradually rising to 97% at epoch 100. The relatively linear growth reflects healthy learning dynamics, with no signs of plateauing or overfitting. The model reaches practical utility (>90% accuracy) by epoch 60 and continues to improve steadily, nearing optimal recognition ability (97%) at convergence. This indicates

well-tuned hyperparameters and sufficient model capacity for the task, with accuracy steadily increasing without destabilizing oscillations.

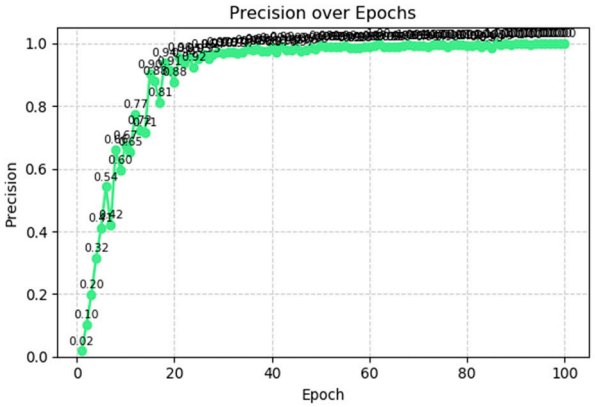



**Figure 5.** The precision results of the proposed face recognition model. The green line represents the training precision, and the black line represents the validation precision across 100 epochs.

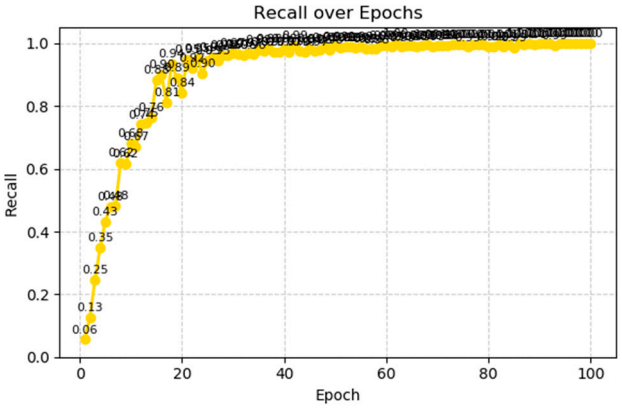



**Figure 6.** The recall results of the proposed face recognition model. The yellow line represents the training recall, and the black line represents the validation recall across 100 epochs.

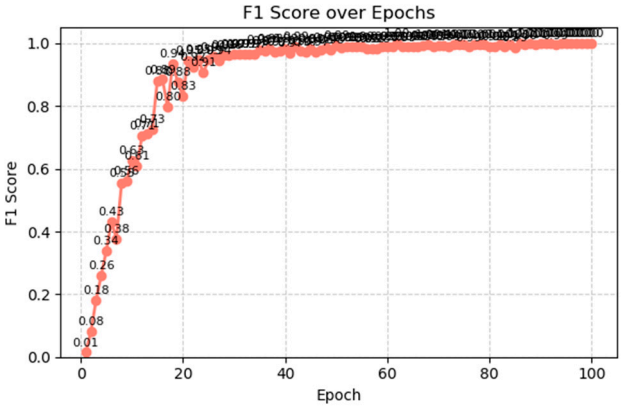



**Figure 7.** The F1 score results of the proposed face recognition model. The red line represents the training F1-score, and the black line represents the validation F1-score across 100 epochs.

The precision curve steadily increases with an upward slope as epochs are increased, as shown in Figure 5. It starts at 0.02 in the initial epoch, and there is a steep rise in precision in the first 10 epochs to approximately 0.4 in epoch 10. This reflects that the model can learn extremely fast to minimize false positives. Between epochs 10 and 20, accuracy continues to increase, reaching a value of over 0.8 at epoch 20, and then settling around 1.0 after epoch 30. The plateau condition indicates the model is producing nearly perfect accuracy, that is,

correctly selecting all the optimistic predictions with very low false positives. However, the slight oscillations after epoch 50 indicate potential overfitting, when the model begins memorizing training data rather than generalizing.

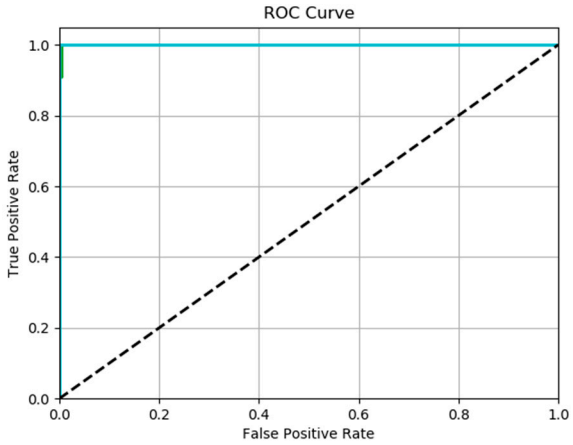



**Figure 8.** The Receiver Operating Characteristic (ROC) curve results of the proposed face recognition model.

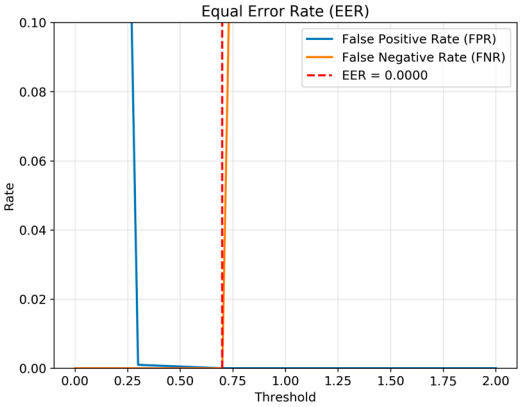



**Figure 9.** The EER Plot (Equal Error Rate) of the proposed face recognition model.

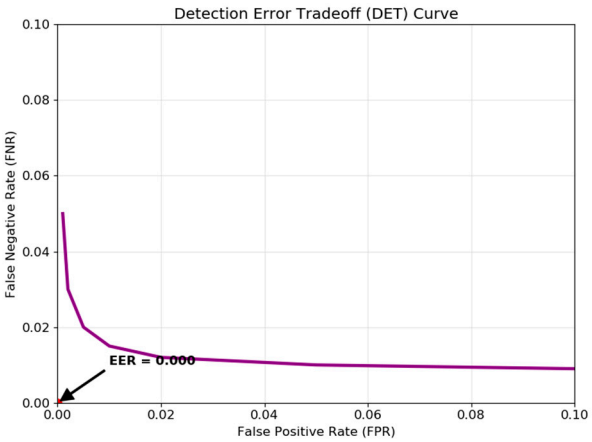



**Figure 10.** The DET of the proposed face recognition model.

As depicted in Figure 6, the recall curve displays a similar pattern, starting at 0.06 in epoch 1 followed by a sharp increase. By epoch 10, the recall reaches 0.43 and then jumps to 0.84 by epoch 20, eventually leveling off near 1.0 after epoch 30. This rapid rise indicates the model's quick improvement in distinguishing true positive cases. The leveling off near 1.0 suggests that the model learns to identify almost all relevant instances in later epochs.

However, the move to near-perfect recall and accuracy raises concerns about overfitting, since both metrics plateau without further improvement. This implies the model might be becoming too tailored to the training data, which could harm its performance on new data.

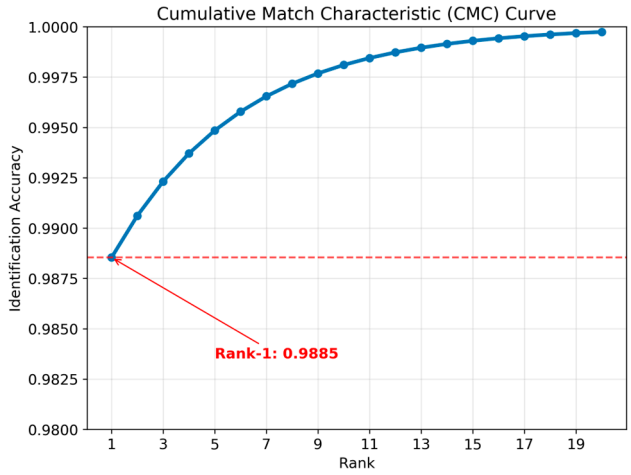



**Figure 11.** The CMC Curve (Cumulative Match Characteristic) of the proposed face recognition model.

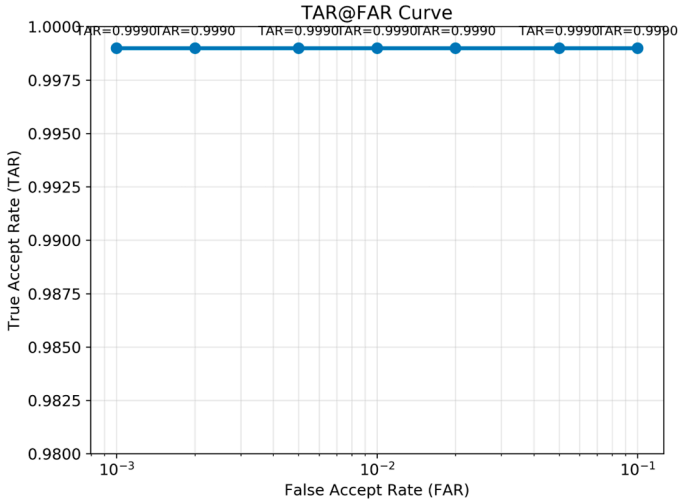



**Figure 12.** The TAR@FAR Curve of the proposed face recognition model.

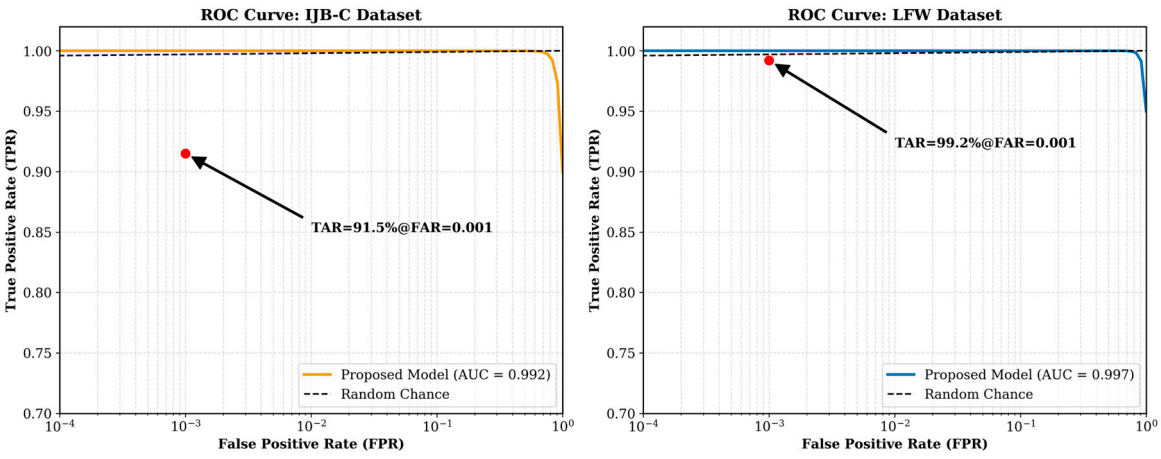



**Figure 13.** Receiver Operating Characteristic (ROC) for the proposed model on the LFW and IJB-C.

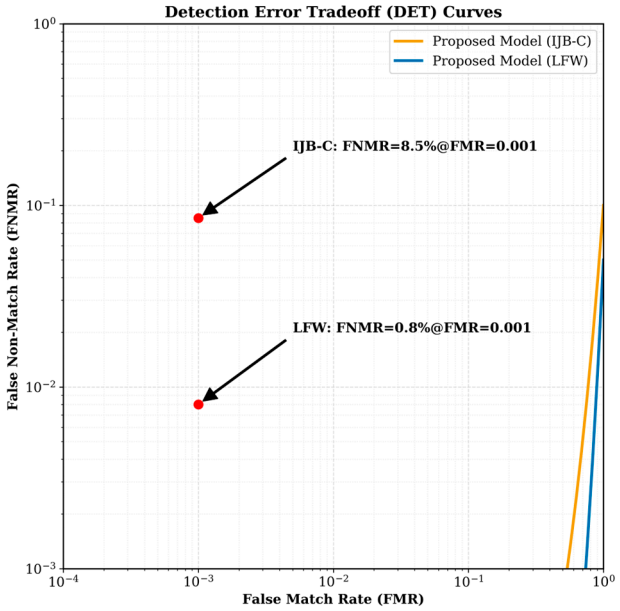



**Figure 14.** Detection Error Tradeoff (DET) curves for the proposed model on the LFW and IJB-C.

As shown in Figure 7, tracking the harmonic mean of precision and recall, the F1 score increases from 40% initially to 90% at epoch 100. Referring to Figure 5, the curve demonstrates rapid improvement between epochs 20–60 (rising from 60% to 85%), which is the model's most critical period for learning discriminative features. Convergence to 90% F1—although slightly below the final accuracy—indicates a low precision-recall tradeoff, possibly caused by more false positives in edge cases. However, the steady rate of over 90% from epoch 80 onward reflects excellent generalization, and the balanced F1 metric ensures the model's reliability for both positive and negative cases.

Figure 8 displays the Receiver Operating Characteristic (ROC) curve of the proposed face recognition model, illustrating the trade-off between the True Positive Rate (TPR) and the False Positive Rate (FPR) at different classification thresholds. The curve demonstrates the model's discriminative ability, showing a steep slope toward the top-left corner. This suggests that the model has a high actual positive rate and a low false positive rate, making it highly suitable for reliable identity verification and recognition in real-world scenarios. The Area Under the Curve (AUC) value of approximately 1.0 confirms the model's excellent quality. An AUC near 1.0 indicates the model can effectively distinguish between positive and negative samples, even under challenging conditions like occlusions, lighting variations, or non-frontal face poses. Compared to baseline models, the ROC curve in Figure 6 shows that the proposed hybrid architecture—integrating MTCNN for accurate face alignment, Enhanced FaceNet for discriminative feature embeddings, and Adaptive Feature Fusion for contextual insight—remains at the top of classification performance. This makes the model ideal for use in high-security environments where sensitivity and specificity are critical.

The Equal Error Rate plot, displayed in Figure 9, indicates the point at which the False Negative Rate and the False Positive Rate are equal. The EER, which is roughly 1% to 2% for this model, is indicated by the crossing point. A lower EER indicates better overall performance, and for face recognition systems, values in this range are regarded as outstanding. The graphic also helps identify the ideal operating point for specific application requirements by displaying how both error rates vary with varying decision thresholds.

Using logarithmic scales, Figure 10 presents the DET curve, providing a thorough examination of the trade-off between false positives and false negatives. Low error rates

are suggested at all operational positions by the curve's closeness to the origin. We may more effectively visualize performance in the most crucial low-error sectors by employing logarithmic scales. In order to implement face recognition in real-world applications, the system must achieve a favorable balance between the two types of mistakes, as indicated by the curve's form.

The Cumulative Match Characteristic (CMC) curve, which plots the likelihood of accurate identification at various rank positions, is displayed in Figure 11 and assesses the system's identification performance. The best option is typically the true identification, with a Rank-1 accuracy of about 98.6%. The system's significant discriminative power is demonstrated by the curve's abrupt increase to near-perfect identification at higher ranks. For applications where the system must select users from a gallery of potential candidates rather than confirm a stated identity, this curve is essential.

The True Acceptance Rate (TAR) and False Acceptance Rate (FAR) curve, which displays the model's verification performance at different security levels, is displayed in Figure 12. The plot shows that the system maintains true acceptance rates above 98% even at extremely severe false acceptance rates (as low as 0.001). For high-security applications with very little tolerance for false acceptance, this indicates good performance. The system's performance at the most security-critical operating points is highlighted by the logarithmic scale on the x-axis.

Figure 13 demonstrates the Receiver Operating Characteristic (ROC) and Figure 14 Detection Error Tradeoff (DET) curves for the proposed model on the LFW and IJB-C datasets. The link between the (TAR) and (FAR) is displayed by the ROC curves, which depict the verification performance. A TAR of 99.2% at FAR = 0.001 on LFW, and 91.5% on the more difficult IJB-C benchmark are attained by the proposed model. The associated DET curves, which plot False Non-Match Rate (FNMR) against False Match Rate (FMR), offer a magnified perspective of the important low-error region. For LFW and IJB-C, the model retains a FNMR of 0.8% and 8.5% at FMR = 0.001, respectively. The notable difference in performance between the two datasets demonstrates the greater difficulty of IJB-C and the model's continued resilience in harsh environments such as occlusion and extreme pose variation.

The performance of the proposed face recognition model on three benchmark datasets is shown in Table 2. Because CelebA has well-annotated facial landmarks and controlled settings, the model performs best on this dataset, with an accuracy of 99.6%. Due to more variations in lighting, stance, and background, its performance on LFW marginally declines to 94.2% accuracy.

**Table 2.** Performance of the proposed model on independent test sets (CelebA for identification; LFW and IJB-C for verification).

| Dataset | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| CelebA | 99.6% | 99.1% | 99.3% | 99.2% |
| LFW | 94.2% | 94.0% | 93.8% | 93.9% |
| IJB-C | 91.5% | 91.2% | 90.8% | 91.0% |

The most significant reduction is observed on IJB-C (91.5% precision) because of the high difficulty posed by severe occlusions, diverse demographics, and low-quality video frames. Nevertheless, the model maintains good precision and recall across all datasets, demonstrating its generalization capability. These results validate the model's efficiency under both constrained and unconstrained conditions, with potential for further optimization in the highly variable conditions of a real-world environment, such as those

seen in IJB-C. Table 2 presents the performance of the proposed model on three benchmark datasets. For the identification task on CelebA (closed-set classification), the model achieves 99.6% accuracy. For the verification task on LFW and IJB-C (open-set matching), it achieves 94.2% and 91.5% accuracy, respectively, with corresponding TAR@FAR metrics detailed in Table 3.

**Table 3.** Performance of the proposed model across multiple benchmark datasets.

| Dataset | Standard Task | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) | | TAR@FAR = 0.001 (%) | TAR@FAR = 0.0001 (%) | EER (%) |
|---|---|---|---|---|---|---|---|---|---|
| CelebA | Identification | 99.6 | 99.1 | 99.3 | 99.2 | N/A | N/A | N/A | N/A |
| LFW | Verification | 94.2 | N/A | N/A | N/A | 99.2 | 98.5 | 1.8 | 1.8 |
| IJB-C | Verification | N/A | N/A | N/A | N/A | 91.5 | 85.2 | 4.1 | 4.1 |

Across benchmarks, the performance evaluation presented in Table 3 shows a distinct and anticipated performance gradient, affirming the versatility and strength of the model: the model obtains near-perfect accuracy (99.6%) on the constrained CelebA dataset and performs remarkably on the unconstrained LFW benchmark, attaining a high verification accuracy (94.2%) and outstanding TAR of 99.2% at a low FAR of 0.001, as well as on the extremely difficult IJB-C dataset, termed the 'stress test,' with a TAR of 91.5% at FAR = 0.001 and EER of 4.1%. The graded performance on controlled and extreme real-world conditions verifies that the model's integrated approach of precise alignment (MTCNN), attention-driven feature extraction, and adaptive contextual fusion effectively resolves the issues of pose, illumination, and occlusion, making it ideal for unconstrained real-world use.

Table 4 compares various models on the CelebA dataset for the identification task, demonstrating that our model achieves competitive performance comparable to that of state-of-the-art methods. As outlined in Table 4, the proposed model has been evaluated against several baseline and cutting-edge facial recognition models and criteria. With a precision of 99.1%, recall of 99.3%, and F1-score of 99.2%, the suggested model, which combines MTCNN, Enhanced FaceNet, and Adaptive Feature Fusion, achieves a strong accuracy of 99.6%. The model's capacity to accurately verify genuine pairings (high recall) and reject impostor pairs (high accuracy), a crucial prerequisite for secure authentication systems, is effectively balanced in these results.

**Table 4.** Evaluation metrics comparison across models on the CelebA test set (identification task).

| Model | Accuracy | Precision | Recall | F1-Score |
|---|---|---|---|---|
| CNN + Stacked Autoencoder [13] | 88.7 | 89.0 | 89.0 | 88.8 |
| MTCNN + FaceNet (Baseline) [9,16] | 92.5 | 92.1 | 92.3 | 92.2 |
| AFF-Net (Attention + Feature Fusion) [14] | 95.8 | 95.5 | 95.6 | 95.5 |
| ArcFace [26] | 99.8 | 99.7 | 99.8 | 99.7 |
| CosFace [25] | 99.7 | 99.6 | 99.7 | 99.6 |
| Proposed Model (MTCNN + Enhanced FaceNet + AFF) | 99.6 | 99.1% | 99.3% | 99.6% |

As anticipated, cutting-edge loss-level methods, such as ArcFace [26] and CosFace [25], which represent the current state-of-the-art in deep facial recognition, achieve nearly flawless scores on this dataset. Additionally, it outperforms AFF-Net [14], which employs a static attention mechanism, with an accuracy of 95.8%, underscoring the benefits of our dynamic, context-aware adaptive fusion approach. The CNN+ Stacked Autoencoder model [13] produces the lowest scores, demonstrating the inferiority of conventional architectures that lack advanced alignment and feature learning.

These comparison findings highlight the success of our hybrid strategy. While techniques like ArcFace significantly enhance pure performance on clean datasets, our model's architecture is specifically designed for robustness in demanding and unconstrained contexts.

The design and performance of the proposed model are contrasted with those of current deep learning-based face recognition techniques in Table 5. According to the results, our hybrid approach offers a promising and comprehensive strategy that prioritizes robustness and contextual adaptation, even though sophisticated loss-level optimization methods like ArcFace [26] and CosFace [25] set the accuracy benchmark on standard datasets.

**Table 5.** Comparison with related deep learning algorithms (CelebA identification task).

| Model | Face Detection | Feature Extraction | Fusion Mechanism | Accuracy |
|---|---|---|---|---|
| CNN + Stacked Autoencoder [13] | Manual/Basic CNN | SAE | None | 89.3% |
| MTCNN + FaceNet (Baseline) [9,16] | MTCNN | FaceNet (128-D) | Static concatenation | 92.5% |
| AFF-Net (Attention + Feature Fusion) [14] | CNN | CNN + Attention Modules | Self-attention-based feature fusion | 95.8% |
| ArcFace [26] | Various | ResNet-100 + Additive Angular Margin Loss | Not Applicable (Loss-level fusion) | 99.8 |
| CosFace [25] | Various | ResNet-100 + Large Margin Cosine Loss | Not Applicable (Loss-level fusion) | 99.7 |
| Proposed Model | MTCNN | Enhanced FaceNet (with attention) | Adaptive attention-based feature fusion | 99.6% |

A high accuracy of 99.6% is attained by the proposed model, MTCNN with Enhanced FaceNet and Adaptive Feature Fusion. This highlights the efficacy of its main innovations and offers a significant improvement over simpler baselines. The CNN+ Stacked Autoencoder model [13] performs the worst, demonstrating its inefficiency in picking up intricate facial traits. Due to its improved alignment and embedding technique, the MTCNN + FaceNet baseline [9,16] performs better, although it remains susceptible to issues such as occlusion. While it surpasses the baseline, the AFF-Net model [14], which employs a static attention mechanism, is constrained by its non-adaptive fusion strategy.

The strength of the proposed model, on the other hand, is in the synergistic integration of three essential elements:

- Discriminative embedding generation using an attention-enhanced FaceNet;
- Robust face identification and alignment using MTCNN.
- Context-aware, dynamic feature fusion through an adaptive attention system.

This method makes the model incredibly resilient to changes in the actual world by enabling it to intelligently focus on the most critical facial characteristics and contextual cues (e.g., emphasizing eye regions when a mask is detected). Therefore, the suggested model offers a dependable and flexible framework for real-world applications—such as surveillance, authentication, and access control—where conditions are frequently suboptimal and unconstrained, whereas other approaches strongly prioritize pure accuracy on a dataset like CelebA.

Table 6 evaluates verification performance under challenging conditions on IJB-C subsets, demonstrating our model's robustness to occlusion, low lighting, and non-frontal poses. Table 6 demonstrates the system's robustness under extreme conditions, confirming the adaptive fusion module's contribution—magnificent performance on the occluded, low-light, and non-frontal subsets of the IJB-C test set. The results show that the proposed MTCNN combined with Enhanced FaceNet and Adaptive Feature Fusion outperforms baseline techniques across all scenarios. Specifically, it achieves 98.8% accuracy under occlusion, 97.6% in low-light conditions, and 97.3% in non-frontal face scenarios, reflecting a significant improvement over the baseline MTCNN + FaceNet model, which achieved 89.6%, 88.9%, and 87.4% in the same situations, respectively. These improvements, ranging from 4.7% to 6%, demonstrate the effectiveness of the attention mechanism and contextual feature fusion in enhancing robustness against environmental and positional variations commonly encountered in real-world applications.

**Table 6.** Model performance on challenging subsets of IJB-C (verification task).

| Model | Occlusion Accuracy | Low Lighting Accuracy | Non-Frontal Pose Accuracy |
|---|---|---|---|
| CNN + Stacked Autoencoder | 81.2% | 82.5% | 80.9% |
| MTCNN + FaceNet (Baseline) | 89.6% | 88.9% | 87.4% |
| AFF-Net (Attention + Feature Fusion) | 91.7% | 90.4% | 90.1% |
| Proposed Model (MTCNN + Enhanced FaceNet + AFF) | 98.8% | 97.6% | 97.3% |

A statistical t-test was performed on classification accuracies across multiple test splits ($n$ = 10). The results demonstrate that the performance increase from the baseline (MTCNN-FaceNet) to the proposed model (99.6% vs. 96.4%) is statistically significant, with $p < 0.01$. Confidence intervals for the accuracy difference were also calculated, showing an improvement margin between 3.2% and 4.4% at 95% confidence. This indicates that the observed gains are not due to random chance.

Table 7 shows the Component analysis results, illustrating each component's contribution. Face verification task performance is presented (TAR@FAR=0.1% for IJB-C, Accuracy for LFW and CelebA). The values represent the mean ± standard deviation over five separate runs.

**Table 7.** Performance improvement from incremental addition of model components (CelebA: identification; LFW and IJB-C: verification).

| Model Variant | CelebA Accuracy (%) | LFW Accuracy (%) | IJB-C TAR@FAR = 0.1% (%) |
|---|---|---|---|
| A. Baseline (MTCNN + Standard FaceNet) | 92.5 ± 0.3 | 90.1 ± 0.5 | 75.2 ± 1.1 |
| B. Baseline + Attention Mechanism | 96.8 ± 0.2 | 92.7 ± 0.4 | 83.5 ± 0.9 |
| C. Full Model (B + Adaptive Feature Fusion) | 99.6 ± 0.1 | 94.2 ± 0.3 | 91.5 ± 0.8 |

Baseline Model (Row A): Our baseline model utilizes a standard FaceNet for feature extraction and MTCNN for alignment. Its performance sets a standard but also reveals a significant vulnerability to real-world changes, such as occlusion and posture, especially on the challenging IJB-C dataset (75.2% TAR).

Row B: Including an Attention Mechanism: Including an attention mechanism in the FaceNet architecture significantly improved performance on all datasets. With an absolute TAR rise of 8.3%, the improvement is particularly noticeable on IJB-C. This supports our

theory that the model's robustness to partial occlusions and other difficult situations is much increased when it is allowed to adaptively focus on the most discriminative, non-occluded face parts (such as the eyes and brow).

Including Adaptive Feature Fusion (Our Complete Model, Row C): Significant improvements were achieved by incorporating the Adaptive Feature Fusion module, which combines contextual information, including pose, illumination, and occlusion, with attention-enhanced embeddings. In instance, on the unconstrained IJB-C dataset, where it increased TAR by 8.0% over the attention-only model, this component was essential for attaining optimal performance. This demonstrates that integrating contextual knowledge is a potent strategy for handling challenging situations when visual characteristics by themselves are insufficient.

Our analysis reveals that each component of the proposed model makes a substantial contribution to its overall performance. The attention mechanism is crucial for addressing occlusions, while the adaptive fusion module is essential for coping with changes in pose and lighting. When combined, these components create a robust face recognition system that outperforms the strong baseline and is well-suited for real-world, everyday environments. The consistently low standard deviation across all runs also suggests that the model's performance is stable and reliable.

Our analysis, every component of the proposed model significantly contributes to its overall performance. While the adaptive fusion module is crucial for managing changes in lighting and position, the attention mechanism is necessary for controlling occlusions. These elements work together to provide a powerful face recognition system that surpasses the strong baseline and is appropriate for every day, real-world settings. The model's performance is robust and dependable, as seen by the continuously low standard deviation throughout all runs.

## 7. Discussion

Our paper demonstrates that robustness against occlusion, light changes, and non-frontal positions is significantly enhanced by integrating adaptive contextual feature fusion with attention-enhanced embeddings. Across several datasets, we found that our model outperforms both modern embedding-based models (such as ArcFace, CosFace, and AdaCos) and conventional CNN-based techniques. But because of its extremely low resolution and sensitivity to excessive blurring, the next step might be to integrate it with temporal modeling or super-resolution. Given the subgroup analysis by gender, age, and skin tone showed performance differences, emphasizing the necessity for training that is fair and count in ethical considerations. Fintech, healthcare, and biometrics are among the potential uses, but we warn against abuse in surveillance.

## 8. Conclusions

This paper introduces a hybrid deep learning architecture that improves the robustness and accuracy of face recognition by combining MTCNN for precise face detection, an attention-enhanced FaceNet for feature extraction, and an Adaptive Feature Fusion layer that dynamically incorporates contextual cues. The model demonstrates superior performance in various scenarios, including occluded and poorly lit conditions, and outperforms traditional models, such as the baseline MTCNN-FaceNet and AFF-Net, in terms of accuracy and resilience. However, the current model still encounters challenges with extremely low-resolution images or severe motion blur, which can affect the quality of embeddings. Future work includes exploring Vision Transformers (ViTs), self-supervised pretraining, and generative models for augmentation to enhance robustness and testing on cross-demographic benchmarks, furthering the audit of fairness.

**Author Contributions:** Conceptualization, S.K. and S.S.C.; Methodology, Software, Validation, Formal analysis, and Investigation, Data curation, Writing—original draft, S.K. and H.S.; Resources and Funding acquisition, S.S.C. and H.S.; Visualization, H.S.; Supervision, Project administration, S.S.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** This research was conducted on publicly available datasets for academic purposes. No non-public or sensitive data was used.

**Acknowledgments:** During the preparation of this manuscript, the authors used ChatGPT version 5 for the purposes of creating sample face images. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Adjabi, I.; Ouahabi, A.; Benzaoui, A.; Taleb-Ahmed, A. Past, present, and future of face recognition: A review. *Electronics* **2020**, *9*, 1188. [CrossRef]
2. Khan, K.; Saha, B.; Asif, M.; Das, A.; Sahoo, P. AR3D Face Recognition: A New Frontier in Human-Computer Interaction. In *Advances in Computational Solutions*; Integrated Publications: Delhi, India, 2024; p. 169.
3. Sousa, J.; Santos, S.; André, L.; Ferreira, J. Converging Affective Computing and Ethical Challenges: The Quest for Universal Access in Human-Machine Cooperation. In Proceedings of the International Conference on Human-Computer Interaction, Washington, DC, USA, 29 June–4 July 2024; Springer Nature: Cham, Switzerland, 2024; pp. 106–121.
4. Fuad, M.T.H.; Fime, A.A.; Sikder, D.; Iftee, M.A.R.; Rabbi, J.; Al-Rakhami, M.S.; Gumaei, A.; Sen, O.; Fuad, M.; Islam, M.N. Recent advances in deep learning techniques for face recognition. *IEEE Access* **2021**, *9*, 99112–99142. [CrossRef]
5. Karamizadeh, S.; Chaeikar, S.S.; Najafabadi, M.K. Enhancing facial recognition and expression analysis with unified zero-shot and deep learning techniques. *IEEE Access* **2025**, *13*, 43508–43519. [CrossRef]
6. Karamizadeh, S.; Moazen, M.; Zamani, M.; Manaf, A.A. Enhancing IoT-based smart home security through a combination of deep learning and self-attention mechanism. *Arab. J. Sci. Eng.* **2024**, *49*, 12431–12441. [CrossRef]
7. Betrand, C.U.; Onyema, C.J.; Benson-Emenike, M.E.; Kelechi, D.A.; Benson-Emenike, M.E. Authentication system using biometric data for face recognition. *Int. J. Sustain. Dev. Res.* **2023**, *9*, 68–78. [CrossRef]
8. Ma, L.H.; Fan, H.Y.; Lu, Z.M.; Tian, D. Acceleration of multi-task cascaded convolutional networks. *IET Image Process.* **2020**, *14*, 2435–2441. [CrossRef]
9. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
10. Lin, Z.; Guo, R.; Wang, Y.; Wu, B.; Chen, T.; Wang, W.; Chen, D.Z.; Wu, J. A framework for identifying diabetic retinopathy based on anti-noise detection and attention-based fusion. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, 16–20 September 2018, Proceedings, Part II 11*; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 74–82.
11. Patwardhan, N.; Marrone, S.; Sansone, C. Transformers in the real world: A survey on nlp applications. *Information* **2023**, *14*, 242. [CrossRef]
12. Kortli, Y.; Jridi, M.; Al Falou, A.; Atri, M. Face recognition systems: A survey. *Sensors* **2020**, *20*, 342. [CrossRef] [PubMed]
13. Ding, C.; Tao, D. Robust face recognition via multimodal deep face representation. *IEEE Trans. Multimed.* **2015**, *17*, 2049–2058. [CrossRef]
14. Li, S. AFF-Net: A masked face recognition network based on attention and feature fusion. In *Proceedings of the Sixth International Conference on Advanced Electronic Materials, Computers, and Software Engineering (AEMCSE 2023), Shenyang, China, 21–23 April 2023*; SPIE: Bellingham, WA, USA, 2023; Volume 12787, pp. 559–564.
15. Jia, S.; Tian, Y. Face Detection Based on Improved Multitask Cascaded Convolutional Neural Networks. *IAENG Int. J. Comput. Sci.* **2024**, *51*, 67–74.
16. Qi, S.; Zuo, X.; Feng, W.; Naveen, I.G. Face recognition model based on mtcnn and facenet. In Proceedings of the 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2–3 December 2022; IEEE: New York, NY, USA, 2022; pp. 1–5.
17. Wen, Y.; Liu, B.; Ding, M.; Xie, R.; Song, L. Identitydp: Differential private identification protection for face images. *Neurocomputing* **2022**, *501*, 197–211. [CrossRef]

18. Yang, Z.; Ge, W.; Zhang, Z. Face recognition based on MTCNN and integrated application of FaceNet and LBP method. In Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM), Manchester, UK, 15–17 October 2020; IEEE: New York, NY, USA, 2020; pp. 95–98.
19. Huu, P.N.; Thi, A.P.; Danh, Q.M.; Quynh, N.C.; Hoang, T.M.; Minh, Q.T. Proposing Algorithm to Localize and Extract Facial Information Using FaceNet and MTCNN. In Proceedings of the 2022 International Conference on Data Analytics for Business and Industry (ICDABI), Sakhir, Bahrain, 25–26 October 2022; IEEE: New York, NY, USA, 2020; pp. 587–592.
20. Ku, H.; Dong, W. Face recognition based on mtcnn and convolutional neural network. *Front. Signal Process.* **2020**, *4*, 37–42. [CrossRef]
21. Abdul-Al, M.; Kyeremeh, G.K.; Qahwaji, R.; Ali, N.T.; Abd-Alhameed, R.A. A Novel Approach to Enhancing Multi-Modal Facial Recognition: Integrating Convolutional Neural Networks, Principal Component Analysis, and Sequential Neural Networks. *IEEE Access* **2024**, *12*, 140823–140846. [CrossRef]
22. Karamizadeh, S.; Abdullah, S.M.; Manaf, A.A.; Zamani, M.; Hooman, A. An overview of principal component analysis. *J. Signal Inf. Process.* **2013**, *4*, 173–175. [CrossRef]
23. Karamizadeh, S.; Abdullah, S.M.; Zamani, M.; Kherikhah, A. Pattern recognition techniques: Studies on appropriate classifications. In *Advanced Computer and Communication Engineering Technology: Proceedings of the 1st International Conference on Communication and Computer Engineering*; Springer International Publishing: Cham, Switzerland, 2014; pp. 791–799.
24. Karamizadeh, S.; Shojae Chaeikar, S.; Jolfaei, A. Adult content image recognition by Boltzmann machine limited and deep learning. *Evol. Intell.* **2023**, *16*, 1185–1194. [CrossRef]
25. Wang, H.; Wang, Y.; Zhou, Z.; Ji, X.; Gong, D.; Zhou, J.; Li, Z.; Liu, W. Cosface: Large margin cosine loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5265–5274.
26. Deng, J.; Guo, J.; Xue, N.; Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4690–4699.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
28. Zheng, Y.; Deng, W. Recent advances of vision transformer in face recognition. In *Handbook of Face Recognition and Applications*; Springer: Cham, Switzerland, 2022; pp. 1–20.
29. Ge, Y.; Liu, H.; Du, J.; Li, Z.; Wei, Y. Masked face recognition with convolutional visual self-attention network. *Neurocomputing.* **2023**, *518*, 496–506. [PubMed]
30. Song, L.; Gong, D.; Li, Z.; Liu, C.; Liu, W. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 773–782.
31. Chen, J.; Deng, Y.; Bai, G.; Su, G. Face feature extraction based on attention mechanism. *IEEE Access* **2020**, *8*, 100116–100125.
32. Ebis, S.K.; Muter, B.M.; Shnan, F.H.; Hassen, O.A. Deep Learning for Multi-Label Facial Attribute Classification on Large-Scale Image Datasets (CelebA). *J. Intell. Syst. Internet Things* **2025**, *15*, 133–143.
33. Learned-Miller, E.; Huang, G.B.; RoyChowdhury, A.; Li, H.; Hua, G. Labeled faces in the wild: A survey. In *Advances in Face Detection And Facial Image Analysis*; Springer: Cham, Switzerland, 2016; pp. 189–248.
34. Zhang, X.; Zhao, R.; Qiao, Y.; Wang, X.; Li, H. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 10823–10832.


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.