FaceMoE: Mixture of Experts for Low-Resolution Face Recognition

Kartik Narayan and Vishal M. Patel

Johns Hopkins University
https://github.com/Kartik-3004/FaceMoE

Abstract. Low-resolution face recognition (LR-FR) remains a challenging task due to poor feature extraction and aggregation, as probe images often contain limited identity information resulting from extreme degradations such as blur, occlusion, and low contrast. Additionally, the domain gap between high-resolution (HR) gallery images and low-resolution (LR) probe images poses a significant challenge. A single feature encoder struggles to generalize effectively across both domains when fine-tuned on an LR dataset, and this issue is further magnified by catastrophic forgetting. To address these challenges, we propose **FaceMoE**, an effective adaptation of Mixture of Experts (MoE) transfomer architecture for low-resolution face-recognition . Specifically, we introduce multiple specialized feed-forward network (FFN) experts and incorporate a top-k router, which dynamically assigns tokens to appropriate experts. This design emergently promotes specialization across experts for different semantic regions of the face, which enables FaceMoE to perform *resolution-aware feature extraction*. Moreover, the top-k router facilitates sparse expert activation, enabling the model to preserve pretrained knowledge when finetuned on a LR dataset, while increasing model capacity without proportional computational overhead. FaceMoE is trained with a combined face recognition loss, router z-loss, and load balancing loss to ensure expert specialization and stable training. To the best of our knowledge, this is the first work leveraging MoE for LR-FR. Extensive experiments across eleven datasets, spanning HR, mixed-quality, and LR benchmarks, demonstrate that FaceMoE significantly outperforms state-of-the-art methods.

Keywords: Low-resolution Face Recognition · Biometrics

1 Introduction

The human face serves as a primary visual modality for a diverse range of computer vision tasks, including deepfake generation/detection [36–38, 55], face anti-spoofing [40], segmentation [43], and face analysis [41, 42, 56]. Among these, face recognition is one of the foundational tasks in computer vision and biometrics, involving the identification and verification of individuals from images or videos. It plays a vital role in real-world applications such as authentication [49], banking [58], and border control [10]. Recently, there has been a growing focus on low-resolution face recognition (LR-FR) [1, 2, 17, 35], due to its widespread

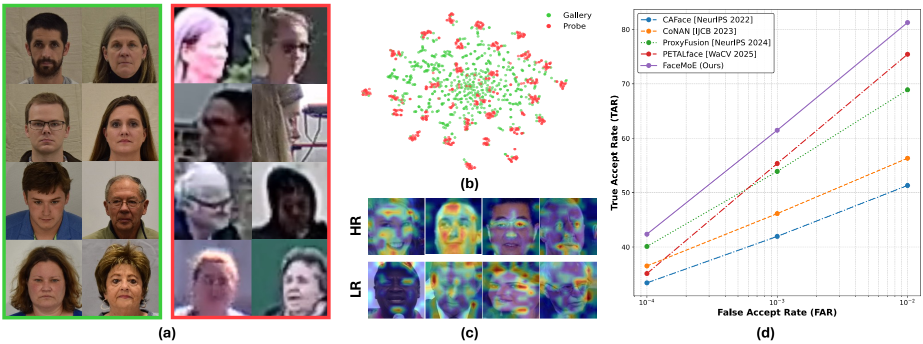

Fig. 1: (a) BRIAR gallery and probe. (b) Domain difference between gallery and probe. (c) Activation maps corresponding to LR and HR images. (d) SOTA results on BRIAR Protocol 3.1.

applicability in surveillance [21]. However, this task is particularly challenging because the input images or videos are often of surveillance quality and severely degraded by factors such as atmospheric turbulence, occlusion, overexposure, and motion blur. These degradations significantly reduce the discriminative features necessary for reliable identification, making conventional recognition techniques less effective. Additionally, variations in pose, illumination, and expression become more pronounced and harder to manage in low-resolution settings, often resulting in poor generalization and reduced performance. Therefore, LR-FR remains a challenging yet crucial problem to address.

To improve the effectiveness of low-resolution face-recognition, it is essential to address several key challenges: ***Challenge 1 - Effective face feature aggregation*:** Probe videos in low-resolution datasets often suffer from significant degradation, which makes face feature aggregation particularly difficult. Since only a limited subset of frames typically contains discriminative identity information, effective feature extraction, followed by aggregation is crucial to build robust face templates. ***Challenge 2 - HR gallery and LR probe domain difference*:** In LR-FR, gallery images are typically high-resolution (HR), while probe images are low-resolution (LR) and come from distinct domains, as validated in Figures 1(a) and 1(b). Models tend to rely on different semantic regions depending on the input resolution to achieve accurate recognition. For HR images, they focus on skin texture, landmarks regions and other fine details that provide sufficient identity information. In contrast, for LR inputs, the face region can be severely degraded to extract any identity information. In such cases, the focus shifts towards broader shapes and coarse facial structures. These resolution-dependent patterns are clearly illustrated by the activation maps in Figure 1(c). This gallery and probe domain gap poses a significant challenge for effective feature extraction. ***Challenge 3 - Catastrophic forgetting when adapting to LR dataset:*** LR-FR models are generally trained in two stages: large-scale pretraining on HR datasets, followed by finetuning on the target LR domain. The second-stage adaptation process makes the model prone to *catastrophic forgetting*, due to unstable gradient updates in the initial epochs of finetuning [27], caused by the significant domain difference between HR and LR datasets. As a result, the model not only loses its pretrained performance but also fails to effectively adapt to the low-resolution data. We validate this effect empirically and show the resulting performance drop in Figure 3 (Finetuning (CosFace)) and Figure 4 (Finetuned CosFace).

Existing works aimed at improving low-resolution face recognition, such as CAFace [23], CoNAN [17], and ProxyFusion [18], focus on addressing *Challenge 1* by selecting relevant frames for fusion after the feature extraction. Specifically, CAFace [23] utilizes an intermediate style map; CoNAN [17] learns a context vector conditioned on distributional information to weigh features based on their estimated informativeness; and ProxyFusion [18] employs learnable queries to identify the most relevant frames. However, the effectiveness of these methods is constrained by the quality of the trained feature encoders, which ultimately limits their overall performance. In contrast, PETAL*face* [39] introduces quality-adaptive dual low-rank modules aimed at developing a more generalized feature encoder across both high-resolution and low-resolution domains, thereby catering to all the key challenges in low-resolution face recognition. Nevertheless, its performance on the low-resolution domain remains subpar compared to the other state-of-the-art methods.

To address the aforementioned challenges, we propose **FaceMoE**, a novel adaptation designed to tackle the core issues in LR-FR. We introduce an architectural modification to the transformer block by incorporating a mixture of feed-forward network (FFN) experts in place of the standard single FFN. Existing transformer-based face recognition encoders typically employ a single FFN following the self-attention operation. However, we argue that a single FFN is insufficient for the complex task of low-resolution face recognition, as it struggles to effectively handle both the HR gallery and LR probe domains. Moreover, it lacks the *resolution-aware feature extraction* necessary for robust identity representation. Our modified transformer block addresses these limitations by using multiple FFN experts and a top-k router that directs each input patch to a subset of k out of n experts based on the input resolution. This design enables different FFN experts to specialize in distinct facial regions, with the top-k router dynamically assigning the subset of experts based on resolution, thereby achieving *resolution-aware feature extraction*. This enables the model to extract strong identity representations by routing input tokens from regions that retain identity cues in degraded images to specialized experts tailored to those regions. This improves feature extraction from LR probes and enhances overall face feature aggregation. Furthermore, the presence of multiple FFN experts facilitates effective adaptation to LR datasets with a minimal drop in pretrained performance. This is achieved through the modular and sparsely activated nature of MoE, which restricts weight updates to only a subset of experts during fine-tuning, thereby reducing *catastrophic forgetting* [5]. The modular design allows experts to function as semi-independent blocks; during fine-tuning, this structure induces *selective drift* [50], with some experts adapting to LR data while others retain their pretrained knowledge (evidence in supplementary material). This retained knowledge enables the model to perform effective feature extraction for both the HR gallery and LR probe domains, further enhanced by its resolution-aware feature extraction capabilities. FaceMoE is trained using a combination of router z-loss and load-balancing loss, which promotes both expert specialization and balanced utilization, thereby preventing training collapse. The top-k routing ensures sparse expert utilization, with a increase in model capacity without a proportional rise in computational cost achieving 2.17$\times$ more capacity with only 1.66$\times$ more FLOPs.

To summarize our contributions are as follows:

1. We propose **FaceMoE**, a modified transformer encoder with sparsely activated FFN experts. It enables efficient adaptation to low-resolution datasets while minimizing *catastrophic forgetting*, effectively addressing the domain gap between gallery and probe images.
2. We introduce a top-k router that assigns each input token to a subset of FFN experts, each specializing in distinct semantic facial regions. This enables *resolution-aware feature extraction.* The router directs tokens containing discriminative identity cues to the most relevant experts, thereby enhancing feature representation and improve LR-FR performance.
3. We demonstrate the effectiveness of FaceMoE by outperforming state-of-the-art models on low-resolution face recognition (see Figure 1(c)). We showcase its capabilities through evaluations on eleven datasets, covering HR, mixed-quality, and LR scenarios.

2 Related Work

Low Resolution Face-Recognition. Face recognition research has largely focused on developing variants of margin-based loss functions [4, 14, 22, 29, 59, 61] to improve the performance on high-resolution benchmarks [2, 3, 21]. In contrast, much less attention has been given to low-resolution unconstrained face recognition (LR-FR) datasets [2, 3, 21], which contain heavily degraded face images that are unidentifiable by humans. Efforts to improve LR-FR can be broadly categorized into four areas based on their focus: data, training methodology, feature fusion, and architectural design. Early works [54, 68] used super-resolution (SR) models to restore images prior to recognition, but later works [19, 26, 70] suggest that this approach can cause identity hallucination. Many studies [11, 53, 66, 67] relate recognition to visual quality. However, this is infeasible as it requires paired HR and LR images of the same subject, which are mostly unavailable in LR datasets. [30, 31] introduce augmentations to mitigate the performance gap between HR and LR samples. In terms of training methods, some works [32, 74] use knowledge distillation to transfer information from the HR domain to the LR domain. For instance, [6, 7] adopt a teacher-student framework, while [13] proposes a distribution distillation loss. Additionally, [1] focuses on optimizing the embedding space to boost performance. In the area of feature fusion, CAFace [23] proposes a two-stage approach that leverages style information. CoNAN [17] learns a context vector conditioned on the distribution and weighs features based on their estimated informativeness. ProxyFusion [18] employs learnable queries to select a sparse set of expert networks for feature aggregation. Recent architecture-based methods include PETALface [39], which introduces two image quality-adaptive LoRA modules. Our work, FaceMoE, also falls within the architecture category. We introduce multiple FFN experts, each specialized in different face regions for enhanced feature encoding. This design achieves state-of-the-art performance on multiple LR-FR benchmarks.

Mixture of Experts. Mixture of Experts (MoE) architectures have emerged as a powerful approach to scale model capacity efficiently by activating only a subset of specialized experts per input. [52] introduced sparsely-gated MoEs, demonstrating their effectiveness in large language models. [25] employed conditional computation and automatic sharding to scale transformer-based models to the trillion-parameter range through efficient model and data parallelism. Several works have adopted the MoE design for vision applications such as image classification [8,46,47,69], object detection [16,44], semantic segmentation [48,60], and image generation [20, 45, 64]. Building on these advances, recent efforts have also explored MoE architectures for face-related applications. MoE-FFD [24] proposes a parameter-efficient ViT-based approach for face forgery detection by integrating MoE modules with LoRA and adapter layers. [73] presents a MoE-injected architecture with a dynamic expert aggregation network for generalizable face anti-spoofing. In our work, we aim to use an MoE-enhanced transformer architecture to boost the performance of LR-FR.

3 Method

In this work, we aim to enhance the generalization capability of face recognition models, with a particular focus on improving LR-FR performance. We first introduce preliminary concepts regarding Mixture of Experts. We then introduce FaceMoE, an MoE-transformer that facilitates robust feature extraction across both HR and LR domains, while mitigating catastrophic forgetting when fine-tuned on LR datasets. Finally, we outline our training framework.

3.1 Preliminaries: Mixture of Experts

The MoE framework [15, 52] is a modular neural architecture that leverages multiple specialized sub-models (experts) to model complex data distributions. Formally, let $x \in \mathbb{R}^d$ be an input vector. The MoE model consists of N experts $\{f_i(x; \theta_i)\}_{i=1}^N$, where each $f_i : \mathbb{R}^d \to \mathbb{R}^m$ is parameterized by θ_i, and a gating network $G(x; \phi) = [w_1(x), \ldots, w_N(x)]$, parameterized by ϕ, which outputs a probability distribution over the experts such that $\sum_{i=1}^N w_i(x) = 1$. The gating weights are commonly obtained using a softmax, defined as $w_i(x) = \frac{\exp(g_i(x))}{\sum_{j=1}^N \exp(g_j(x))}$, where $g_i(x)$ denotes the score of the i-th expert. The final output of the MoE is a convex combination of the expert outputs, given by

$$y = \sum_{i=1}^{N} w_i(x) f_i(x; \theta_i).$$

The training objective minimizes a loss function

$$\mathcal{L} = \frac{1}{K} \sum_{k=1}^{K} \ell \left(y^{(k)}, \sum_{i=1}^{N} w_i(x^{(k)}) f_i(x^{(k)}; \theta_i) \right),$$

where $\ell(\cdot, \cdot)$ is a task-specific loss (such as mean squared error or cross-entropy). Sparse MoE variants [52] further improve computational efficiency by restricting active experts to a subset $S \subset \{1, \ldots, N\}$, yielding $y = \sum_{i \in S} w_i(x) f_i(x; \theta_i)$. In this work, we introduce FaceMoE, which adopts the MoE paradigm within a transformer-based FR model to enable dynamic routing and specialization across experts, thereby enhancing feature extraction and improving LR-FR performance.

3.2 FaceMoE

To address the challenge of feature extraction in LR-FR, we introduce FaceMoE, a novel transformer architecture enhanced with an MoE mechanism. The primary motivation behind integrating MoE within the transformer blocks is to encourage dynamic specialization of sub-networks (experts) to different patterns present in facial data. FaceMoE inserts the experts into the feed-forward (MLP) layers. We select linear projections as experts due to their proven capacity to introduce additional non-linearity when composed with transformer self-attention, enhancing the model's ability to capture complex patterns in data [57]. The linear layer experts serve to extract complementary information from the attended tokens generated by the multi-head self-attention operation. This design choice balances expressiveness and computational efficiency, as the MLP layers constitute a significant portion of transformer model capacity. This modular approach allows the model to adapt better to LR face images, while preserving pretrained knowledge.

Mixture of Experts MLP Layer: In FaceMoE, the MoE is incorporated inside the MLP layers of the transformer block. Let $x \in \mathbb{R}^{T \times d}$ represent a sequence of T tokens, each of dimensionality d, output by the self-attention block. The expert layer comprises N expert MLPs, $\{f_i(x; \theta_i)\}_{i=1}^{N}$, each parameterized by weights θ_i. The experts operate independently but in parallel to process the input tokens. An individual expert is a two-layer fully connected network with weights $\{W_{i,1}, W_{i,2}\}$ and biases $\{b_{i,1}, b_{i,2}\}$, defined as:

$$f_i(x_t) = W_{i,2} \cdot \sigma(W_{i,1} x_t + b_{i,1}) + b_{i,2}, \quad \forall t \in \{1, \ldots, T\},$$

where $\sigma(\cdot)$ is an activation function, in this case GELU [9], $W_{i,1} \in \mathbb{R}^{d \times h}$, $W_{i,2} \in \mathbb{R}^{h \times d}$, $b_{i,1} \in \mathbb{R}^{h}$, and $b_{i,2} \in \mathbb{R}^{d}$, with h being the hidden dimension. This formulation enables each expert to non-linearly transform and project each token representation.

Top-k Router: The top-k router is a core component of FaceMoE, responsible for dynamically assigning input tokens to a subset of experts. Given token embeddings $x \in \mathbb{R}^{T\times d}$, the router computes expert selection logits for each token x_t using a linear projection: $z_t = x_t W_r$, where $W_r \in \mathbb{R}^{d\times N}$ are learnable routing weights, and $z_t \in \mathbb{R}^N$ contains the routing scores for the N experts. For each token t, the router selects the indices of the top-k experts with the highest activations: $(i_1, i_2, \ldots, i_k) = \mathrm{TopK}(z_t)$, where $i_j \in \{1, \ldots, N\}$. The logits of the selected experts are normalized by a softmax over the top-k values to produce the routing probabilities: $w_{i_j}(x_t) = \frac{\exp(z_{t,i_j})}{\sum_{j=1}^{k}\exp(z_{t,i_j})}$. The final output of the MoE layer for token x_t is a convex combination of the outputs of the selected experts: $y_t = \sum_{j=1}^{k} w_{i_j}(x_t) f_{i_j}(x_t)$. This sparse routing strategy leads to significant computational savings, as only $k < N$ experts are active per token. Importantly, it enables efficient adaptation to low-resolution datasets. In our experiments, we empirically found that setting $N = 3$ and $k = 2$ yielded the best trade-off between model performance and efficiency. Under this configuration, we observed that the router exhibits conditional routing behavior, where each expert is implicitly specialized for certain semantic regions of the face, as shown in Figure 2. This behavior can be expressed by the conditional routing probability:

$$\mathbb{P}(i_j \mid R_t = r) > \mathbb{P}(i_j \mid R_t \neq r), \quad \forall r \in \{\text{high-freq}, \text{low-freq}, \text{landmarks}\},$$

where R_t denotes the semantic or frequency region of token x_t. Specifically, tokens corresponding to high-frequency regions (e.g., edges, contours, hair textures, background) are primarily routed to one expert; tokens from low-frequency smooth regions (e.g., cheeks, forehead) are directed to a second expert; and tokens corresponding to landmark regions (e.g., eyes, nose) are routed to the third expert.

3.3 Training Framework

To train FaceMoE, we optimize a composite objective combining a primary face recognition loss with auxiliary regularization terms designed to stabilize the MoE routing process. The primary loss is based on the well-established *CosFace* margin-based softmax loss [59] denoted as $\mathcal{L}_{\text{face}}$, which encourages inter-class separability and intra-class compactness in the learned embedding space. In addition, we introduce two auxiliary losses applied to the router network:

1. Router z-loss: This regularization term penalizes the magnitude of the routing logits to mitigate over-confident expert assignments and support stable gradient flow throughout training. For a batch size B, where each sample contains T tokens, the router z-loss is formulated as:

$$\mathcal{L}_z = \lambda_z \cdot \frac{1}{B \cdot T} \sum_{b=1}^{B} \sum_{t=1}^{T} \|z_{b,t}\|_2^2,$$

where $z_{b,t} \in \mathbb{R}^N$ is the vector of raw routing logits for token t in sample b, $\|\cdot\|_2$ denotes the ℓ_2-norm, and λ_z is a regularization coefficient controlling the penalty strength. This quadratic penalty, distributed over the entire batch, encourages the router to generate smoothly varying logits with lower variance, enhancing routing stability and mitigating expert collapse.

Algorithm 1 FaceMoE Training Framework

Input: Training samples $\{x^{(k)}, y^{(k)}\}_{k=1}^{K}$, FaceMoE weights $\theta = \{\theta_1, \ldots, \theta_N, W_r\}$, Experts $\{f_i(\cdot;\theta_i)\}_{i=1}^{N}$, Router Weights W_r
Hyperparameters: $\lambda, \lambda_z, \lambda_b$
Output: Trained FaceMoE weights θ

for *each training epoch* **do**
  for *each batch* $\{x_b, y_b\}_{b=1}^{B}$ **do**
    for *each token* $x_{b,t}$ *in* x_b **do**
      $z_{b,t} = x_{b,t} W_r$ ▷ compute routing logits
      $(i_1, \ldots, i_k) = \mathrm{TopK}(z_{b,t})$ ▷ select top-k experts
      $w_{i_j}(x_{b,t}) = \frac{\exp(z_{b,t,i_j})}{\sum_{l=1}^{k} \exp(z_{b,t,i_l})}$ ▷ routing weights
      $p_{b,t,i} = \frac{\exp(z_{b,t,i})}{\sum_{j=1}^{N} \exp(z_{b,t,j})}$ ▷ softmax prob. for f_i
      $y_{b,t} = \sum_{j=1}^{k} w_{i_j}(x_{b,t}) f_{i_j}(x_{b,t})$ ▷ MoE output
    $\mathcal{L}_{\text{face}} = CosFace(y_{b,t})$ ▷ face recognition loss
     $\mathcal{L}_z = \lambda_z \cdot \frac{1}{BT} \sum_{b,t} \|z_{b,t}\|_2^2$ ▷ router z-loss
     $\mathcal{L}_{\text{balance}} = \lambda_b N \frac{1}{(BT)^2} \sum_i \left(\sum_{b,t} p_{b,t,i}\right) \cdot \left(\sum_{b,t} \nVdash[i \in (i_1, \ldots, i_k)]\right)$ ▷ load balancing loss
     $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{face}} + \lambda(\mathcal{L}_z + \mathcal{L}_{\text{balance}})$ ▷ total loss
     $\theta \leftarrow \texttt{Optimizer}(\theta, \nabla_\theta \mathcal{L}_{\text{total}})$ ▷ parameter update

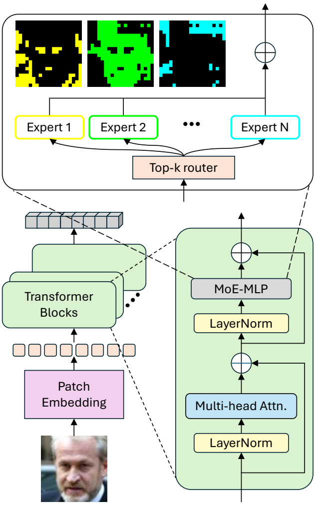

Fig. 2: FaceMoE architecture: a transformer backbone augmented with a top-k MoE-MLP routing module, where tokens are assigned to specialized experts for robust LR-FR.

2. Load balancing loss: This loss promotes uniform utilization of experts across all tokens and samples, mitigating the risk of expert under-utilization or collapse. For a batch size B, the load balancing loss is defined as:

$$\mathcal{L}_{\text{balance}} = \lambda_b \cdot N \cdot \frac{1}{(B \cdot T)^2} \sum_{i=1}^{N} \left(\sum_{b=1}^{B} \sum_{t=1}^{T} p_{b,t,i} \right) \cdot \left(\sum_{b=1}^{B} \sum_{t=1}^{T} \nVdash [i \in \mathrm{TopK}(z_{b,t})] \right),$$

where $p_{b,t,i} = \frac{\exp(z_{b,t,i})}{\sum_{j=1}^{N} \exp(z_{b,t,j})}$ is the softmax probability of assigning token t in sample b to expert i. $\nVdash[i \in \mathrm{TopK}(z_{b,t})]$ is an indicator function that equals 1 if expert i is among the top-k selected experts for token t in sample b, and 0 otherwise. The hyperparameter λ_b controls the strength of this regularization term. This formulation jointly considers the *importance* of expert i (measured by the sum of routing probabilities across all tokens) and the *load* (the count of tokens routed to expert i). The inclusion of $\mathcal{L}_{\text{balance}}$ in the final objective promotes balanced expert selection and prevents bottlenecks in expert utilization.

The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{face}} + \lambda_1 \mathcal{L}_z + \lambda_2 \mathcal{L}_{\text{balance}},$$

where λ_1 and λ_2 are the weighting factor of router-z loss and load-balancing loss, respectively. This joint optimization framework allows FaceMoE to efficiently scale model capacity while dynamically specializing experts to different facial regions, thereby enhancing low-resolution face recognition performance. The FaceMoE architecture is shown in Figure 2 and the training procedure is shown in Algorithm 1.

4 Experimental Setup

4.1 Datasets

We use WebFace4M [75] as our pre-training dataset, which consist of approximately 4M images, with $205,990$ identities. To demonstrate the effectiveness of the proposed FaceMoE for low-resolution face recognition, we evaluate it on 3 low-resolution datasets. Further, to validate the minimal drop in pretrained performance, we also evaluate its performance on 6 high-quality datasets and 2 mixed-quality datasets. The high-quality datasets include LFW [12], CFP-FP [51], CPLFW [71], AgeDB [34], CALFW [72], and CFP-FF [51]. The mixed-quality datasets are IJB-B [62] and IJB-C [33]. The low-resolution datasets include TinyFace [2], IJB-S [21], and BRIAR 3.1 [3]. The TinyFace [2] dataset contains $169,403$ low-resolution images spanning $5,139$ identities, with a designated training subset of $7,804$ images covering $2,570$ identities. The IJB-S [21] dataset, designed for surveillance video-based face recognition, comprises 398 videos and 202 identities. We evaluate it under *Surveillance-to-Surveillance* protocol, where "*Surveillance*" refers to footage from surveillance cameras. The BRIAR [3] training set includes $550,000$ images from 577 distinct identities. For the BRIAR evaluation, we follow Protocol 3.1 (face-included treatment), in line with prior works [17, 18]. This evaluation protocol features a gallery of $86,958$ controlled images representing 615 identities and a probe set comprising $5,435$ clips from 260 identities.

4.2 Evaluation Setup and Metrics

We organize our experiments into two protocols to comprehensively evaluate FaceMoE across a variety of scenarios. In **Protocol-1**, we pre-train FaceMoE on the WebFace4M [75], finetune it on the challenging low-resolution BRIAR [3] dataset, and evaluate its performance using BRIAR Protocol 3.1, demonstrating the effectiveness of FaceMoE for low-resolution face recognition. We also test the model on IJB-S [21] which is another challenging video-surveillance dataset to show its out-of-distribution performance. In **Protocol-2**, we finetune our model on TinyFace [2] and evaluate it on its test set. With this protocol, we aim to highlight the capability of FaceMoE to adapt to low-resolution datasets while maintaining performance on high-resolution and mixed-quality datasets.

We evaluate the models on high-resolution and mixed-quality datasets using 1:1 verification accuracy and TAR@FAR across various thresholds. For TinyFace, we apply rank retrieval metrics at Rank-1, Rank-5, and Rank-10. On the BRIAR dataset, we report both TAR@FAR at different thresholds and closed-set rank retrieval at Rank-1, Rank-5, and Rank-20. For IJB-S, we evaluate open-set performance using TPIR@FPIR = 1% and 10%, along with closed-set rank retrieval at Rank-1, Rank-5, and Rank-10.

4.3 Implementation Details

We train FaceMoE on WebFace4M with a batch size of 128 per GPU for 26 epochs, using AdamW (weight decay 5×10^{-2}) and a Polynomial LR scheduler with 1 warmup epoch and initial LR 10^{-3}. Fine-tuning is done on TinyFace and BRIAR in two stages: linear probing and full fine-tuning. For TinyFace, linear probing runs 10 epochs (2 warmup) at LR 10^{-3}, batch size 16; full fine-tuning runs 40 epochs (4 warmup) at LR 10^{-4}, batch size 8. For BRIAR, both stages run 20 epochs (2 warmup), with LRs 10^{-3} and 5×10^{-6}, batch sizes 64 and 8. Training uses face recognition loss, router z-loss, and load balancing loss with $\lambda_1 = 10$, $\lambda_2 = 10$, $\lambda_z = 1$, $\lambda_b = 1$. The λ_1 and λ_2 values scale the auxiliary loss terms so that their magnitudes are comparable to the main face recognition loss, ensuring stable optimization without either term dominating training. The best results are obtained with 3 experts ($N = 3$) and 2 active experts per token ($k = 2$). All experiments use PyTorch on eight NVIDIA A6000 GPUs (48GB). Additional details are provided in the supplementary material.

5 Results and Analysis

Results on Protocol 1: The results for Protocol 1 are summarized in Table 1 and Table 2. The pretrained transformer backbones ViT-B and Swin-B show superior performance than ResNet-50, however these models are not finetuned on low-resolution datasets and perform poorly compared to finetuned methods. Traditional feature aggregation methods such as GAP [28], NAN [65], MCN [63], CAFace [23], and CoNAN [17] yield incremental improvements, but remain limited in their ability to extract discriminative identity features from degraded probe images, as they use a feature encoder with single FFN and focus on selecting relevant frames with sufficient identity information. However, our method aims to improve the identity extraction of all the frames by improving the feature extractor itself. Recent methods, ProxyFusion and PETAL*face*, achieve a TAR@FAR of 40.10, 53.90, 68.90 & 35.12, 55.35, 75.43 at thresholds 0.01%, 0.1% & 1%, resp.

Our proposed FaceMoE achieves the highest performance across all thresholds with 42.36%, 61.47%, and 81.27% TAR at 0.01%, 0.1%, and 1% FAR, respectively. The superior performance of FaceMoE can be attributed to its *resolution-aware feature extraction* enabled by specialized experts. Each expert is implicitly trained to focus on distinct semantic regions of the face, such as edges, contours, or landmark regions, enabling dynamic adaptation to severely degraded probe images.

Method		TAR@FAR		
		0.01%	0.1%	1%
Pretrained				
CosFace (R50)		22.55	35.43	52.20
CosFace (ViT-B)		34.29	47.41	62.81
CosFace (Swin-B)		33.77	45.93	61.17
Finetuned on BRIAR train set				
GAP	[ICLR 2014]	31.70	40.81	50.76
NAN	[CVPR 2017]	34.86	44.96	54.44
CosFace	[CVPR 2018]	11.62	29.68	58.66
	[BMVC 2018]	34.84	45.01	54.25
CAFace	[NeurIPS 2022]	33.41	41.95	51.31
CoNAN	[IJCB 2023]	36.52	46.14	56.32
ProxyFusion	[NeurIPS 2024]	40.10	53.90	68.90
PETAL*face*	[WaCV 2025]	35.12	55.35	75.43
FaceMoE		**42.36**	**61.47**	**81.27**

Table 1: BRIAR Protocol 3.1 results

Method		TPIR@FPIR	Rank Retrieval	
		1%	Rank-1	Rank-5
Pretrained				
CosFace (R50)		3.67	33.62	49.40
CosFace (ViT-B)		2.58	25.76	40.69
CosFace (Swin-B)		2.11	22.52	37.97
Finetuned on BRIAR train set				
CosFace	[CVPR 2018]	1.72	16.44	31.58
PFE	[CVPR 2019]	0.84	9.20	20.82
RSA	[ICCV 2019]	0.75	16.82	31.80
MARN	[ICCVW 2019]	0.19	22.25	34.16
ArcFace	[CVPR 2019]	5.32	32.13	46.67
CFAN	[IJCB 2019]	5.79	31.66	45.59
CurricularFace	[CVPR 2020]	2.53	19.54	32.80
AdaFace	[CVPR 2022]	4.96	35.05	48.22
CAFace	[NeurIPS 2022]	8.78	36.51	49.59
PETAL*face*	[WaCV 2025]	12.25	38.32	51.50
FaceMoE		**14.85**	**44.81**	**56.12**

Table 2: Results on IJB-S (Surv. to Surv.).

This capability is especially valuable in low-resolution scenarios, where identity information is limited and often confined to localized regions. In such cases, key identity discriminative features, such as the eyes, nose, or mouth may be occluded, blurred, or affected by extreme lighting conditions. FaceMoE addresses this by allotting specialized semantic experts to other informative regions, enabling a more robust and comprehensive identity representation. This enhanced feature extraction from low-resolution probes directly contributes to superior feature aggregation, resulting in state-of-the-art performance for low-resolution face recognition on the BRIAR dataset. Table 2 reports the generalization performance on the IJB-S dataset under *Surveillance-to-Surveillance* protocol. We observe similar trends, with FaceMoE outperforming all prior methods by a significant margin. FaceMoE achieves 14.85% TPIR at 1% FPIR, along with 44.81% and 56.12% Rank-1 and Rank-5 retrieval accuracies, respectively. The *resolution-aware feature extraction* and expert specialization effectively handle the extreme variability and degradation inherent in surveillance footage, extracting identity features from limited and inconsistent information across frames. This enhanced feature extraction leads to robust identity recognition under the most challenging low-resolution conditions.

Results on Protocol 2: The results for Protocol 2 are shown in Figure 4 and 3. Finetuning pretrained models such as CosFace [59] and ArcFace [4] on TinyFace leads to a drop in performance not only on the LR dataset but also on the mixed-quality and HR datasets. This degradation is primarily due to catastrophic forgetting, as these models lack mechanisms to effectively adapt to low-resolution data while retaining the discriminative features learned during pretraining. This effect can also be observed in Table 1 and Table 2, where finetuned CosFace shows a significant performance drop on BRIAR Protocol 3.1 and IJB-S compared to pretrained CosFace. In contrast, FaceMoE establishes a new state-of-the-art on TinyFace with 76.18%, 79.69%, and 81.75% Rank-1, Rank-5, and Rank-10 retrieval accuracy, respectively, with a minimal drop in performance on the HR and mixed quality datasets as illustrated in Figure 3.

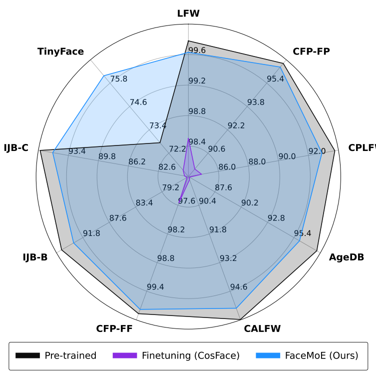

Fig. 3: FaceMoE incurs minimal performance drop on HR and mixed-quality datasets, effectively extracting features from HR gallery and LR probe.

Method	Arch.	Data	Rank-1	Rank-5	Rank-10
Pretrained					
URL	R-100	MS1MV2	63.89	68.67	–
CurricularFace	R-100	MS1MV2	63.68	67.65	–
CosFace	R-50	WF4M	72.71	76.36	78.99
ArcFace	R-50	WF4M	73.04	76.85	79.45
AdaFace	R-50	WF4M	73.49	76.60	79.07
CosFace	ViT-B	WF4M	73.57	76.95	78.94
ArcFace	ViT-B	WF4M	72.74	76.28	78.13
AdaFace	ViT-B	WF4M	74.03	77.22	79.37
CosFace	Swin-B	WF4M	72.74	76.79	79.18
ArcFace	Swin-B	WF4M	73.31	76.68	79.23
AdaFace	Swin-B	WF4M	74.40	77.62	79.51
KP-RPE	ViT-B	WF4M	75.80	78.49	–
Finetuned on TinyFace					
CosFace	Swin-B	WF4M	71.32	76.42	79.45
ArcFace	Swin-B	WF4M	71.11	76.63	79.96
PETAL*face*	Swin-B	WF4M	75.45	79.05	81.19
FaceMoE (Ours)	Swin-B	WF4M	**76.18**	**79.69**	**81.75**

Fig. 4: Results on TinyFace. Pre-trained models when finetuned on TinyFace dataset results in performance drop. FaceMoE achieves SOTA performance and is capable of adapting to low-resolution dataset with minimal performance drop in HQ and mixed-quality dataset.

The superior performance of FaceMoE can be attributed to its unique architectural design, which leverages multiple sparse FFN experts to facilitate effective adaptation to low-resolution datasets, while incurring minimal performance drop on high-resolution and mixed-quality datasets. The top-k router renders the network modular and sparsely activated, restricting weight updates during finetuning to only a subset of experts. As a result, the model avoids *catastrophic forgetting* as observed in traditional models. During finetuning of FaceMoE, the model exhibits a phenomenon known as selective drift [50], where certain experts adapt specifically to the low-resolution dataset, while others retain the pretrained knowledge. As shown in Figure 5(c), expert 2's focus remains largely consistent before and after finetuning, focusing on broader facial shapes, indicating the preservation of pretrained semantic knowledge. However, token assignment changes significantly during finetuning: before finetuning, expert 0 was predominantly utilized, whereas after finetuning, expert 1 becomes more active. This shift highlights FaceMoE's resolution-aware capability and its dynamic utilization of experts based on input resolution. The expert activation maps after finetuning display more semantically coherent and well-defined regions, showcasing the efficacy of employing multiple FFN experts in conjunction with a top-k router for stable adaptation to low-resolution data. FaceMoE's ability to adapt to low-resolution data while preserving pretrained knowledge enables effective feature extraction across high-resolution gallery and low-resolution probe domains.

Impact of N and k on Performance: We perform an ablation study to investigate the effect of the number of experts (N) and the number of active experts per token (k) on model performance. Figure 5(b) shows the Rank-1

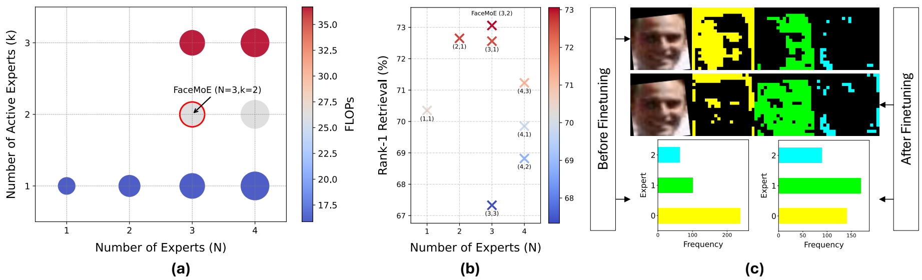

Fig. 5: (a) Computational trade-off analysis across different MoE configurations (Bubble size $\propto$ #Parameters). (b) Impact of N and k on performance, evaluated on the BRIAR dataset. (c) FaceMoE expert activation maps and token assignment histograms before and after fine-tuning on a low-resolution dataset. The updated token assignments indicate *resolution-aware feature extraction*, while the semantically coherent expert activation maps demonstrate stable convergence.

retrieval accuracy on the BRIAR dataset for different (N, k) configurations. We observe that both under-parameterization and over-parameterization can adversely impact performance. A low number of experts ($N = 1$) limits the model's capacity to specialize across facial regions, resulting in sub-optimal performance (70.2%). On the other hand, increasing the number of experts excessively ($N = 4$) introduces routing instability and model fragmentation, leading to degraded performance across multiple k settings. Our best performance is achieved with $N = 3$ experts and $k = 2$ active experts per token, corresponding to the FaceMoE configuration, which achieves 73.1% Rank-1 retrieval. This setting strikes an effective balance between model capacity and routing stability, providing sufficient expert diversity to allow specialization across semantic regions (e.g., hair, landmarks, textures), while avoiding excessive fragmentation of the feature space.

Computational Analysis: We study the computational cost of different (N, k) configurations. Figure 5(a) shows the FLOPs for various combinations of number of experts N and active experts per token k. As expected, computational cost scales with k, since more experts are evaluated per token. Importantly, for fixed k, the parameter count remains constant regardless of N, as only k experts contribute to the forward pass. For example, with $k = 2$, both $(N = 3, k = 2)$ and $(N = 4, k = 2)$ have the same number of active parameters with 26.29 GFLOPs, despite differing in total experts. The optimal configuration for FaceMoE is $(N = 3, k = 2)$, achieving a favorable trade-off between model capacity and computational cost. This results in a moderate 26.29 GFLOPs, offering a 2.17$\times$ increase in capacity over the standard Swin-B backbone (15.88 GFLOPs) with only a 1.66$\times$ increase in FLOPs. This validates the efficiency of sparsely activated experts, enabling the model to significantly boost its representation power while maintaining practical inference cost. We include an inference computation analysis in the supplementary material where we measure inference latency, peak memory consumption, and throughput on two GPU configurations: an NVIDIA RTX A5000 (representing a resource-constrained setting) and an NVIDIA A6000.

Impact of each component We provide an ablation study in Table 3 to assess the impact of each component in FaceMoE. We see a drop in Rank-1 accuracy from 76.18% to 75.40%, if we remove the top-k router, highlighting the importance of dynamic token routing. We see a further reduction in performance to (75.10%), if we exclude the MoE module, confirming the benefit of expert specialization. Finally, omitting the auxiliary losses results in the lowest accuracy of (74.94%), underscoring their role in stabilizing routing and balancing expert utilization. These results demonstrate that all components contribute meaningfully to the overall performance. Please note that we cannot report a configuration with MoE, Aux Loss but without a top-k router, as the auxiliary loss depends on the logits produced by the top-k router and cannot function without them.

MoE	**Top-k Router**	**Aux Loss**	**Rank-1**	**Rank-5**	**Rank-10**
×	×	×	75.10	78.16	80.20
✓	×	×	75.40	78.46	80.63
✓	✓	×	74.94	77.92	79.90
✓	✓	✓	76.18	79.69	81.75

Table 3: Ablation study showing the impact of top-k router, MoE, and auxiliary loss on FaceMoE performance. Results are shown on TinyFace dataset

6 Conclusion

In this work, we present FaceMoE, a novel transformer-based architecture enhanced with a Mixture of Experts mechanism to address persistent challenges in low-resolution face recognition. We incorporate multiple FFN experts and a top-k router, enabling the experts to specialize in different semantic regions of the face. The proposed framework enhances the discriminative power of feature extraction under severe image degradations, and the presence of multiple FFN experts ensures stable finetuning with minimal performance loss on high-resolution and mixed-quality datasets. Extensive evaluations across eleven diverse benchmarks, including challenging low-resolution datasets such as TinyFace, IJB-S, and BRIAR, demonstrate that FaceMoE consistently outperforms existing methods, establishing new SOTA performance in low-resolution face recognition. We believe FaceMoE offers a promising foundation for future research in adaptive, resolution-aware face recognition models and provides a scalable solution for real-world applications in surveillance and security systems.

Acknowledgment

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via [2022-21102100005]. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The US Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

1. Chai, J.C.L., Ng, T.S., Low, C.Y., Park, J., Teoh, A.B.J.: Recognizability embedding enhancement for very low-resolution face recognition and quality estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9957–9967 (2023)
2. Cheng, Z., Zhu, X., Gong, S.: Low-resolution face recognition. In: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14. pp. 605–621. Springer (2019)
3. Cornett, D., Brogan, J., Barber, N., Aykac, D., Baird, S., Burchfield, N., Dukes, C., Duncan, A., Ferrell, R., Goddard, J., et al.: Expanding accurate person recognition to new altitudes and ranges: The briar dataset. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 593–602 (2023)
4. 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. pp. 4690–4699 (2019)
5. Fedus, W., Zoph, B., Shazeer, N.: Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research **23**(120), 1–39 (2022)
6. Ge, S., Zhang, K., Liu, H., Hua, Y., Zhao, S., Jin, X., Wen, H.: Look one and more: Distilling hybrid order relational knowledge for cross-resolution image recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 10845–10852 (2020)
7. Ge, S., Zhao, S., Li, C., Li, J.: Low-resolution face recognition in the wild via selective knowledge distillation. IEEE Transactions on Image Processing **28**(4), 2051–2062 (2018)
8. Han, X., Wei, L., Dou, Z., Wang, Z., Qiang, C., He, X., Sun, Y., Han, Z., Tian, Q.: Vimoe: An empirical study of designing vision mixture-of-experts. arXiv preprint arXiv:2410.15732 (2024)
9. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
10. Hidayat, F., Elviani, U., Situmorang, G.B.G., Ramadhan, M.Z., Alunjati, F.A., Sucipto, R.F.: Face recognition for automatic border control: a systematic literature review. IEEE Access **12**, 37288–37309 (2024)
11. Hsu, C.C., Lin, C.W., Su, W.T., Cheung, G.: Sigan: Siamese generative adversarial network for identity-preserving face hallucination. IEEE Transactions on Image Processing **28**(12), 6225–6236 (2019)

12. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
13. Huang, Y., Shen, P., Tai, Y., Li, S., Liu, X., Li, J., Huang, F., Ji, R.: Improving face recognition from hard samples via distribution distillation loss. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 138–154. Springer (2020)
14. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: adaptive curriculum learning loss for deep face recognition. In: proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5901–5910 (2020)
15. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
16. Jain, Y., Behl, H., Kira, Z., Vineet, V.: Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. Advances in Neural Information Processing Systems **36**, 69625–69637 (2023)
17. Jawade, B., Mohan, D.D., Shetty, P., Fedorishin, D., Setlur, S., Govindaraju, V.: Conan: Conditional neural aggregation network for unconstrained long range biometric feature fusion. IEEE Transactions on Biometrics, Behavior, and Identity Science (2024)
18. Jawade, B., Stone, A., Mohan, D.D., Wang, X., Setlur, S., Govindaraju, V.: Proxyfusion: Face feature aggregation through sparse experts. Advances in Neural Information Processing Systems **37**, 70130–70147 (2024)
19. Jiang, J., Yu, Y., Hu, J., Tang, S., Ma, J.: Deep cnn denoiser and multi-layer neighbor component embedding for face hallucination. arXiv preprint arXiv:1806.10726 (2018)
20. Jiang, Y., Yang, S., Qiu, H., Wu, W., Loy, C.C., Liu, Z.: Text2human: Text-driven controllable human image generation. ACM Transactions on Graphics (TOG) **41**(4), 1–11 (2022)
21. Kalka, N.D., Maze, B., Duncan, J.A., O'Connor, K., Elliott, S., Hebert, K., Bryan, J., Jain, A.K.: Ijb–s: Iarpa janus surveillance video benchmark. In: 2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS). pp. 1–9. IEEE (2018)
22. Kim, M., Jain, A.K., Liu, X.: Adaface: Quality adaptive margin for face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18750–18759 (2022)
23. Kim, M., Liu, F., Jain, A.K., Liu, X.: Cluster and aggregate: Face recognition with large probe set. Advances in Neural Information Processing Systems **35**, 36054–36066 (2022)
24. Kong, C., Luo, A., Bao, P., Yu, Y., Li, H., Zheng, Z., Wang, S., Kot, A.C.: Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. arXiv preprint arXiv:2404.08452 (2024)
25. Lepikhin, D., Lee, H., Xu, Y., Chen, D., Firat, O., Huang, Y., Krikun, M., Shazeer, N., Chen, Z.: Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint arXiv:2006.16668 (2020)
26. Li, P., Prieto, L., Mery, D., Flynn, P.J.: On low-resolution face recognition in the wild: Comparisons and new techniques. IEEE Transactions on Information Forensics and Security **14**(8), 2000–2012 (2019)
27. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
28. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)

29. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
30. Low, C.Y., Teoh, A.B.J.: An implicit identity-extended data augmentation for low-resolution face representation learning. IEEE Transactions on Information Forensics and Security **17**, 3062–3076 (2022)
31. Low, C.Y., Teoh, A.B.J., Park, J.: Mind-net: A deep mutual information distillation network for realistic low-resolution face recognition. IEEE Signal Processing Letters **28**, 354–358 (2021)
32. Massoli, F.V., Amato, G., Falchi, F.: Cross-resolution learning for face recognition. Image and Vision Computing **99**, 103927 (2020)
33. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 international conference on biometrics (ICB). pp. 158–165. IEEE (2018)
34. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 51–59 (2017)
35. Nair, N.G., Narayan, K., Suin, M., Kathirvel, R.P., Xu, J., Stevens, S., Gleason, J., Shnidman, N., Chellappa, R., Patel, V.: Improved representation learning for unconstrained face recognition. In: 2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10. IEEE (2025)
36. Narayan, K., Agarwal, H., Mittal, S., Thakral, K., Kundu, S., Vatsa, M., Singh, R.: Desi: Deepfake source identifier for social media. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2858–2867 (2022)
37. Narayan, K., Agarwal, H., Thakral, K., Mittal, S., Vatsa, M., Singh, R.: Deephy: On deepfake phylogeny. In: 2022 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–10. IEEE (2022)
38. Narayan, K., Agarwal, H., Thakral, K., Mittal, S., Vatsa, M., Singh, R.: Df-platter: Multi-face heterogeneous deepfake dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9739–9748 (2023)
39. Narayan, K., Nair, N.G., Xu, J., Chellappa, R., Patel, V.M.: Petalface: Parameter efficient transfer learning for low-resolution face recognition. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 804–814. IEEE (2025)
40. Narayan, K., Patel, V.M.: Hyp-oc: Hyperbolic one class classification for face anti-spoofing. In: 2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–10. IEEE (2024)
41. Narayan, K., Vibashan, V., Patel, V.M.: Facexbench: Evaluating multimodal llms on face understanding. IEEE Transactions on Biometrics, Behavior, and Identity Science (2026)
42. Narayan, K., VS, V., Chellappa, R., Patel, V.M.: Facexformer: A unified transformer for facial analysis. arXiv preprint arXiv:2403.12960 (2024)
43. Narayan, K., Vs, V., Patel, V.M.: Segface: Face segmentation of long-tail classes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 6182–6190 (2025)
44. Oksuz, K., Kuzucu, S., Joy, T., Dokania, P.K.: Mocae: Mixture of calibrated experts significantly improves object detection. arXiv preprint arXiv:2309.14976 (2023)

45. Park, D.K., Yoo, S., Bahng, H., Choo, J., Park, N.: Megan: Mixture of experts of generative adversarial networks for multimodal image generation. arXiv preprint arXiv:1805.02481 (2018)
46. Puigcerver, J., Riquelme, C., Mustafa, B., Houlsby, N.: From sparse to soft mixtures of experts. arXiv preprint arXiv:2308.00951 (2023)
47. Riquelme, C., Puigcerver, J., Mustafa, B., Neumann, M., Jenatton, R., Susano Pinto, A., Keysers, D., Houlsby, N.: Scaling vision with sparse mixture of experts. Advances in Neural Information Processing Systems **34**, 8583–8595 (2021)
48. Rossi, L., Bernuzzi, V., Fontanini, T., Bertozzi, M., Prati, A.: Swin2-mose: A new single image supersolution model for remote sensing. IET Image Processing **19**(1), e13303 (2025)
49. Roy, M., Datta, S., Khan, M., Paroi, M., Hasan, M.M.: Ai-powered face authentication system for web and native apps. In: 2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS). pp. 1406–1412. IEEE (2025)
50. Rypeść, G., Cygert, S., Khan, V., Trzciński, T., Zieliński, B., Twardowski, B.: Divide and not forget: Ensemble of selectively trained experts in continual learning. arXiv preprint arXiv:2401.10191 (2024)
51. Sengupta, S., Cheng, J., Castillo, C., Patel, V., Chellappa, R., Jacobs, D.: Frontal to profile face verification in the wild. In: IEEE Conference on Applications of Computer Vision (February 2016)
52. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. arXiv preprint arXiv:1701.06538 (2017)
53. Singh, M., Nagpal, S., Singh, R., Vatsa, M.: Derivenet for (very) low resolution image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6569–6577 (2021)
54. Singh, M., Nagpal, S., Vatsa, M., Singh, R., Majumdar, A.: Identity aware synthesis for cross resolution face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 479–488 (2018)
55. Thakral, K., Agarwal, H., Narayan, K., Mittal, S., Vatsa, M., Singh, R.: Deephynet: Toward detecting phylogeny in deepfakes. IEEE Transactions on Biometrics, Behavior, and Identity Science **7**(1), 132–145 (2024)
56. Tu, A., Narayan, K., Gleason, J., Xu, J., Meyn, M., Goldstein, T., Patel, V.M.: Transfira: Transfer learning for face image recognizability assessment. In: 2026 IEEE 20th International Conference on Automatic Face and Gesture Recognition (FG). pp. 1–9. IEEE (2026)
57. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
58. Vishnuvardhan, G., Ravi, V.: Face recognition using transfer learning on facenet: application to banking operations. In: Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough: Latest Trends in AI, Volume 2, pp. 301–309. Springer (2021)
59. 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. pp. 5265–5274 (2018)
60. Wang, X., Yu, F., Dunlap, L., Ma, Y.A., Wang, R., Mirhoseini, A., Darrell, T., Gonzalez, J.E.: Deep mixture of experts via shallow embedding. In: Uncertainty in artificial intelligence. pp. 552–562. PMLR (2020)

61. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14. pp. 499–515. Springer (2016)
62. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., et al.: Iarpa janus benchmark-b face dataset. In: proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 90–98 (2017)
63. Xie, W., Zisserman, A.: Multicolumn networks for face recognition. arXiv preprint arXiv:1807.09192 (2018)
64. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. Advances in Neural Information Processing Systems **36**, 41693–41706 (2023)
65. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G.: Neural aggregation network for video face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4362–4371 (2017)
66. Yin, X., Tai, Y., Huang, Y., Liu, X.: Fan: Feature adaptation network for surveillance face recognition and normalization. In: Proceedings of the Asian Conference on Computer Vision (2020)
67. Yu, X., Fernando, B., Hartley, R., Porikli, F.: Super-resolving very low-resolution face images with supplementary attributes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 908–917 (2018)
68. Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., Zhang, L.: Image super-resolution: The techniques, applications, and future. Signal processing **128**, 389–408 (2016)
69. Zhang, J., Qu, X., Zhu, T., Cheng, Y.: Clip-moe: Towards building mixture of experts for clip with diversified multiplet upcycling. arXiv preprint arXiv:2409.19291 (2024)
70. Zhang, K., Zhang, Z., Cheng, C.W., Hsu, W.H., Qiao, Y., Liu, W., Zhang, T.: Super-identity convolutional neural network for face hallucination. In: Proceedings of the European conference on computer vision (ECCV). pp. 183–198 (2018)
71. Zheng, T., Deng, W.: Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Beijing University of Posts and Telecommunications, Tech. Rep **5**(7), 5 (2018)
72. Zheng, T., Deng, W., Hu, J.: Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. arXiv preprint arXiv:1708.08197 (2017)
73. Zhou, Q., Zhang, K.Y., Yao, T., Yi, R., Ding, S., Ma, L.: Adaptive mixture of experts learning for generalizable face anti-spoofing. In: Proceedings of the 30th ACM international conference on multimedia. pp. 6009–6018 (2022)
74. Zhu, M., Han, K., Zhang, C., Lin, J., Wang, Y.: Low-resolution visual recognition via deep feature distillation. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3762–3766. IEEE (2019)
75. Zhu, Z., Huang, G., Deng, J., Ye, Y., Huang, J., Chen, X., Zhu, J., Yang, T., Lu, J., Du, D., et al.: Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10492–10502 (2021)