FaceMoE: Mixture of Experts for Low-Resolution Face Recognition

Kartik Narayan
Johns Hopkins University
knaraya4@jhu.edu

Vishal M. Patel
Johns Hopkins University
vpatel36@jhu.edu

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**, a novel transformer-based architecture enhanced with a Mixture of Experts (MoE) design. 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 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, excelling in low-resolution face recognition. Code and models will be made public.

1 Introduction

Face recognition is one of the foundational tasks in computer vision and biometrics, involving the recognition and verification of individuals from images or videos. It plays a vital role in real-world applications such as authentication [1], banking [2], and border control [3]. Recently, there has been a growing focus on low-resolution face recognition (LR-FR) [4, 5, 6], due to its widespread applicability in surveillance [7]. 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.

Preprint. Under review.

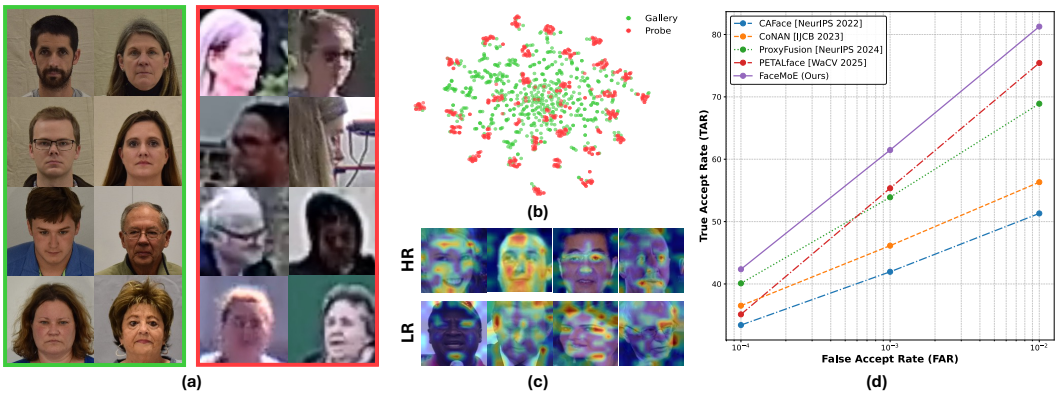

Figure 1: (a) BRIAR gallery and probe. (b) Domain difference between gallery and probe. (c) Activation maps corresponding to LR and HR images. (d) SOTA results on BRIAR Protocol 3.1.

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 [8], 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 [9], CoNAN [6], and ProxyFusion [10], focus on addressing *Challenge 1* by selecting relevant frames for fusion after the feature extraction. Specifically, CAFace [9] utilizes an intermediate style map; CoNAN [6] learns a context vector conditioned on distributional information to weigh features based on their estimated informativeness; and ProxyFusion [10] 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* [11] 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 framework 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* [12]. The modular design allows experts to function as semi-independent blocks; during fine-tuning, this structure induces *selective drift* [13], with some experts adapting to LR data while others retain their pretrained knowledge. 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 [14, 15, 16, 17, 18, 19] to improve the performance on high-resolution benchmarks [4, 7, 20]. In contrast, much less attention has been given to low-resolution unconstrained face recognition (LR-FR) datasets [4, 7, 20], 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 [21, 22] used super-resolution (SR) models to restore images prior to recognition, but later works [23, 24, 25] suggest that this approach can cause identity hallucination. Many studies [26, 27, 28, 29] 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, 33] use knowledge distillation to transfer information from the HR domain to the LR domain. For instance,[34, 35] adopt a teacher-student framework, while[36] proposes a distribution distillation loss. Additionally, [5] focuses on optimizing the embedding space to boost performance. In the area of feature fusion, CAFace [9] proposes a two-stage approach that leverages style information. CoNAN [6] learns a context vector conditioned on the distribution and weighs features based on their estimated informativeness. ProxyFusion [10] employs learnable queries to select a sparse set of expert networks for feature aggregation. Recent architecture-based methods include PETALface [11], 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 low-resolution face recognition 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. [37] introduced sparsely-gated MoEs, demonstrating their effectiveness in large language models. [38] 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 [39, 40, 41, 42], object detection [43, 44], semantic segmentation [45, 46], and image generation [47, 48, 49]. Building on these

advances, recent efforts have also explored MoE architectures for face-related applications. MoE-FFD [50] proposes a parameter-efficient ViT-based approach for face forgery detection by integrating MoE modules with LoRA and adapter layers. [51] 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 propose FaceMoE, an MoE-enhanced 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 for stable convergence.

3.1 Preliminaries: Mixture of Experts

The MoE framework [52, 37] 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 [37] 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 propose 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 [53]. 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 low-resolution 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 [54], $W_{i,1} \in \mathbb{R}^{d\times h}$, $W_{i,2} \in \mathbb{R}^{h\times d}$, $b_{i,1} \in \mathbb{R}^h$, 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 [15] 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.

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:

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

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. $\mathbb{1}[i \in \mathrm{TopK}(z_{b,t})]$ 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:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{face}} + \lambda \cdot (\mathcal{L}_z + \mathcal{L}_{\text{balance}}),$$

where λ is a weighting factor to balance the main recognition objective with the auxiliary routing regularizations. 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 architecutre is shown in Figure 2 and the training procedure is shown in Algorithm 1.

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
      $(i_1, \ldots, i_k) = \text{TopK}(z_{b,t})$ ▷ 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
    end
    $\mathcal{L}_{\text{face}} = \textit{CosFace}(y_{b,t})$ ▷ face recognition loss
    $\mathcal{L}_z = \lambda_z \cdot \frac{1}{BT} \sum_{b,t} \|z_{b,t}\|_2^2$ ▷ router z-loss
    $\mathcal{L}_{\text{balance}} = \lambda_b N \frac{1}{(BT)^2} \sum_i \left(\sum_{b,t} p_{b,t,i} \right)$
    $\cdot \left(\sum_{b,t} \mathbb{1}[i \in (i_1, \ldots, i_k)] \right)$ ▷ 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
  end
end

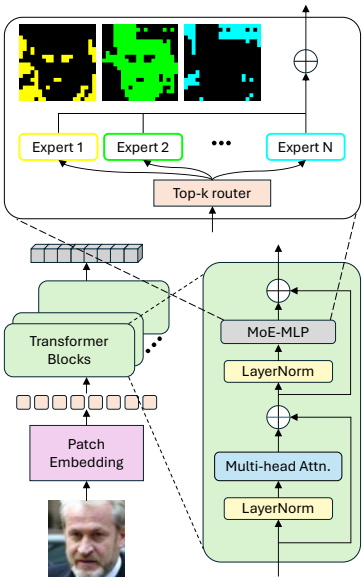

Figure 2: FaceMoE Architecture.

4 Experimental Setup

Datasets. We use WebFace4M [55] 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 [56], CFP-FP [57], CPLFW [58], AgeDB [59], CALFW [60], and CFP-FF [57]. The mixed-quality datasets are IJB-B [61] and IJB-C [62]. The low-resolution datasets include TinyFace [4], IJB-S [7], and BRIAR 3.1 [20]. The TinyFace [4] 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 [7] 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 [20] 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 [6, 10]. 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.

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 [55], finetune it on the challenging low-resolution BRIAR [20] 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 [7] which is another challenging video-surveillance dataset to show its out-of-distribution performance. In **Protocol-2**, we finetune

Method		TAR@FAR		
		0.01%	0.1%	1%
Pretrained				
CosFace [15] (R50)		22.55	35.43	52.20
CosFace [15] (ViT-B)		34.29	47.41	62.81
CosFace [15] (Swin-B)		33.77	45.93	61.17
Finetuned on BRIAR train set				
GAP [63]	[ICLR 2014]	31.70	40.81	50.76
NAN [64]	[CVPR 2017]	34.86	44.96	54.44
CosFace [15]	[CVPR 2018]	11.62	29.68	58.66
MCN [65]	[BMVC 2018]	34.84	45.01	54.25
CAFace [9]	[NeurIPS 2022]	33.41	41.95	51.31
CoNAN [6]	[IJCB 2023]	36.52	46.14	56.32
ProxyFusion [10]	[NeurIPS 2024]	40.10	53.90	68.90
PETAL*face* [11]	[WaCV 2025]	35.12	55.35	75.43
FaceMoE		**42.36**	**61.47**	**81.27**

Table 1: Results on BRIAR Protocol 3.1.

Method		TPIR@FPIR	Rank Retrieval	
		1%	Rank-1	Rank-5
Pretrained				
CosFace [15] (R50)		3.67	33.62	49.40
CosFace [15] (ViT-B)		2.58	25.76	40.69
CosFace [15] (Swin-B)		2.11	22.52	37.97
Finetuned on BRIAR train set				
CosFace [15]	[CVPR 2018]	1.72	16.44	31.58
PFE [66]	[CVPR 2019]	0.84	9.20	20.82
RSA [67]	[ICCV 2019]	0.75	16.82	31.80
MARN [68]	[ICCVW 2019]	0.19	22.25	34.16
ArcFace [14]	[CVPR 2019]	5.32	32.13	46.67
CFAN [69]	[IJCB 2019]	5.79	31.66	45.59
CurricularFace [19]	[CVPR 2020]	2.53	19.54	32.80
AdaFace [18]	[CVPR 2022]	4.96	35.05	48.22
CAFace [9]	[NeurIPS 2022]	8.78	36.51	49.59
PETAL*face* [11]	[WaCV 2025]	12.25	38.32	51.50
FaceMoE		**14.85**	**44.81**	**56.12**

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

our model on TinyFace [4] 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.

Implementation Details. We train FaceMoE on the WebFace4M dataset with a batch size of 128 per GPU for 26 epochs. We employ the AdamW optimizer with a weight decay of $5e^{-2}$. A Polynomial learning rate scheduler is used, with 1 warmup epoch and an initial learning rate of 10^{-3}. We fine-tune FaceMoE on the TinyFace and BRIAR datasets in two stages: linear probing followed by full fine-tuning. For TinyFace, during linear probing, we train for 10 epochs with an additional 2 warmup epochs. For full fine-tuning, we train for 40 epochs with an additional 4 warmup epochs. The learning rates and batch sizes for the two stages are 10^{-3}, 10^{-4} and 16, 8, respectively. For BRIAR, both linear probing and full fine-tuning are conducted for 20 epochs with an additional 2 warmup epochs. The learning rates and batch sizes for the two stages are 10^{-3}, 5×10^{-6} and 64, 8, respectively. During training, we employ a combination of face recognition loss, router z-loss, and load balancing loss. The corresponding hyperparameters λ, λ_z, and λ_b are set to 10, 1, and 1, respectively. We obtain the best results with 3 experts ($N = 3$) and 2 active experts per token ($k = 2$). All code is implemented in PyTorch, and experiments are conducted on eight NVIDIA A6000 GPUs, each with 48 GB of memory. Additional details are provided in the appendix.

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 [63], NAN [64], MCN [65], CAFace [9], and CoNAN [6] 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 [10] and PETAL*face* [11], achieve a TAR@FAR of 40.10, 53.90, 68.90 and 35.12, 55.35, 75.43 at thresholds 0.01%, 0.1% and 1%, respectively.

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. This capability is especially valuable in low-resolution scenarios, where identity information is limited and often

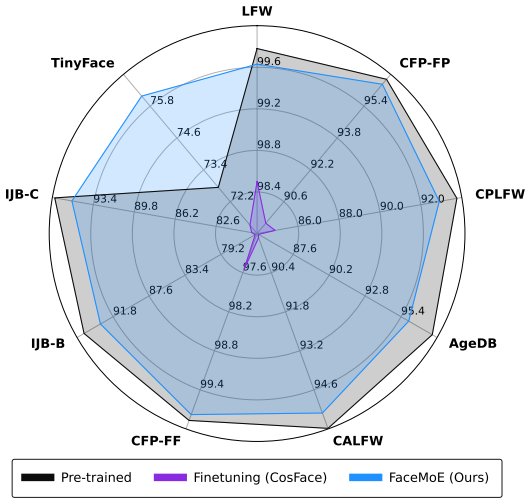

Figure 3: FaceMoE incurs minimal performance drop on HR and mixed-quality datasets, effectively extracting features from HR gallery and LR probe.

Method	Arch.	Data	Rank-1	Rank-5	Rank-10
Pretrained					
URL [70]	R-100	MS1MV2	63.89	68.67	–
CurricularFace [19]	R-100	MS1MV2	63.68	67.65	–
CosFace [15]	R-50	WF4M	72.71	76.36	78.99
ArcFace [14]	R-50	WF4M	73.04	76.85	79.45
AdaFace [18]	R-50	WF4M	73.49	76.60	79.07
CosFace [15]	ViT-B	WF4M	73.57	76.95	78.94
ArcFace [14]	ViT-B	WF4M	72.74	76.28	78.13
AdaFace [18]	ViT-B	WF4M	74.03	77.22	79.37
CosFace [15]	Swin-B	WF4M	72.74	76.79	79.18
ArcFace [14]	Swin-B	WF4M	73.31	76.68	79.23
AdaFace [18]	Swin-B	WF4M	74.40	77.62	79.51
KP-RPE [71]	ViT-B	WF4M	75.80	78.49	–
Finetuned on TinyFace					
CosFace [15]	Swin-B	WF4M	71.32	76.42	79.45
ArcFace [14]	Swin-B	WF4M	71.11	76.63	79.96
PETAL*face* [11]	Swin-B	WF4M	75.45	79.05	81.19
FaceMoE (Ours)	Swin-B	WF4M	**76.18**	**79.69**	**81.75**

Figure 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.

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 [15] and ArcFace [14] 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.

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 [13], 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

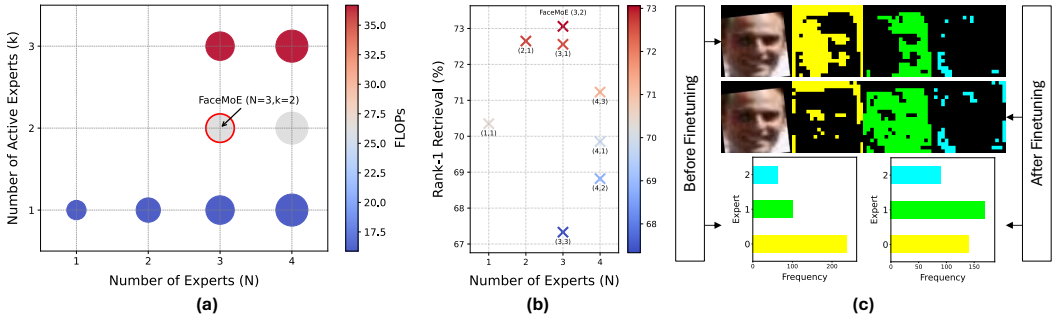

Figure 5: (a) Computational trade-off analysis across different MoE configurations (Bubble size $\propto$ #Parameters). (b) Impact of N and k on performance, evaluated on the BRIAR dataset. (c) FaceMoE expert activation maps and token assignment histograms before and after fine-tuning on a low-resolution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

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 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.

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.

Acknowledgments and Disclosure of Funding

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] Monideepa Roy, Sujoy Datta, Muhit Khan, Methu Paroi, and MD Mehedi Hasan. Ai-powered face authentication system for web and native apps. In *2025 International Conference on Machine Learning and Autonomous Systems (ICMLAS)*, pages 1406–1412. IEEE, 2025.

[2] Gopireddy Vishnuvardhan and Vadlamani Ravi. 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*, pages 301–309. Springer, 2021.

[3] Fadhil Hidayat, Ulva Elviani, George Bryan Gabriel Situmorang, Muhammad Zaky Ramadhan, Figo Agil Alunjati, and Reza Fauzi Sucipto. Face recognition for automatic border control: a systematic literature review. *IEEE Access*, 12:37288–37309, 2024.

[4] Zhiyi Cheng, Xiatian Zhu, and Shaogang Gong. 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*, pages 605–621. Springer, 2019.

[5] Jacky Chen Long Chai, Tiong-Sik Ng, Cheng-Yaw Low, Jaewoo Park, and Andrew Beng Jin Teoh. 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*, pages 9957–9967, 2023.

[6] Bhavin Jawade, Deen Dayal Mohan, Prajwal Shetty, Dennis Fedorishin, Srirangaraj Setlur, and Venu Govindaraju. Conan: Conditional neural aggregation network for unconstrained long range biometric feature fusion. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[7] Nathan D Kalka, Brianna Maze, James A Duncan, Kevin O'Connor, Stephen Elliott, Kaleb Hebert, Julia Bryan, and Anil K Jain. Ijb–s: Iarpa janus surveillance video benchmark. In *2018 IEEE 9th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–9. IEEE, 2018.

[8] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[9] Minchul Kim, Feng Liu, Anil K Jain, and Xiaoming Liu. Cluster and aggregate: Face recognition with large probe set. *Advances in Neural Information Processing Systems*, 35:36054–36066, 2022.

[10] Bhavin Jawade, Alexander Stone, Deen Dayal Mohan, Xiao Wang, Srirangaraj Setlur, and Venu Govindaraju. Proxyfusion: Face feature aggregation through sparse experts. *Advances in Neural Information Processing Systems*, 37:70130–70147, 2024.

[11] Kartik Narayan, Nithin Gopalakrishnan Nair, Jennifer Xu, Rama Chellappa, and Vishal M Patel. Petalface: Parameter efficient transfer learning for low-resolution face recognition. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 804–814. IEEE, 2025.

[12] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[13] Grzegorz Rypeść, Sebastian Cygert, Valeriya Khan, Tomasz Trzciński, Bartosz Zieliński, and Bartłomiej Twardowski. Divide and not forget: Ensemble of selectively trained experts in continual learning. *arXiv preprint arXiv:2401.10191*, 2024.

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[15] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[16] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[17] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 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*, pages 499–515. Springer, 2016.

[18] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.

[19] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5901–5910, 2020.

[20] David Cornett, Joel Brogan, Nell Barber, Deniz Aykac, Seth Baird, Nicholas Burchfield, Carl Dukes, Andrew Duncan, Regina Ferrell, Jim Goddard, 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*, pages 593–602, 2023.

[21] Maneet Singh, Shruti Nagpal, Mayank Vatsa, Richa Singh, and Angshul Majumdar. Identity aware synthesis for cross resolution face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 479–488, 2018.

[22] Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal processing*, 128:389–408, 2016.

[23] Pei Li, Loreto Prieto, Domingo Mery, and Patrick J Flynn. On low-resolution face recognition in the wild: Comparisons and new techniques. *IEEE Transactions on Information Forensics and Security*, 14(8):2000–2012, 2019.

[24] Kaipeng Zhang, Zhanpeng Zhang, Chia-Wen Cheng, Winston H Hsu, Yu Qiao, Wei Liu, and Tong Zhang. Super-identity convolutional neural network for face hallucination. In *Proceedings of the European conference on computer vision (ECCV)*, pages 183–198, 2018.

[25] Junjun Jiang, Yi Yu, Jinhui Hu, Suhua Tang, and Jiayi Ma. Deep cnn denoiser and multi-layer neighbor component embedding for face hallucination. *arXiv preprint arXiv:1806.10726*, 2018.

[26] Chih-Chung Hsu, Chia-Wen Lin, Weng-Tai Su, and Gene Cheung. Sigan: Siamese generative adversarial network for identity-preserving face hallucination. *IEEE Transactions on Image Processing*, 28(12):6225–6236, 2019.

[27] Xi Yin, Ying Tai, Yuge Huang, and Xiaoming Liu. Fan: Feature adaptation network for surveillance face recognition and normalization. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[28] Xin Yu, Basura Fernando, Richard Hartley, and Fatih Porikli. Super-resolving very low-resolution face images with supplementary attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 908–917, 2018.

[29] Maneet Singh, Shruti Nagpal, Richa Singh, and Mayank Vatsa. Derivenet for (very) low resolution image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6569–6577, 2021.

[30] Cheng-Yaw Low and Andrew Beng-Jin Teoh. An implicit identity-extended data augmentation for low-resolution face representation learning. *IEEE Transactions on Information Forensics and Security*, 17:3062–3076, 2022.

[31] Cheng-Yaw Low, Andrew Beng-Jin Teoh, and Jaewoo Park. Mind-net: A deep mutual information distillation network for realistic low-resolution face recognition. *IEEE Signal Processing Letters*, 28:354–358, 2021.

[32] Fabio Valerio Massoli, Giuseppe Amato, and Fabrizio Falchi. Cross-resolution learning for face recognition. *Image and Vision Computing*, 99:103927, 2020.

[33] Mingjian Zhu, Kai Han, Chao Zhang, Jinlong Lin, and Yunhe Wang. Low-resolution visual recognition via deep feature distillation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3762–3766. IEEE, 2019.

[34] Shiming Ge, Shengwei Zhao, Chenyu Li, and Jia Li. Low-resolution face recognition in the wild via selective knowledge distillation. *IEEE Transactions on Image Processing*, 28(4):2051–2062, 2018.

[35] Shiming Ge, Kangkai Zhang, Haolin Liu, Yingying Hua, Shengwei Zhao, Xin Jin, and Hao Wen. Look one and more: Distilling hybrid order relational knowledge for cross-resolution image recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 10845–10852, 2020.

[36] Yuge Huang, Pengcheng Shen, Ying Tai, Shaoxin Li, Xiaoming Liu, Jilin Li, Feiyue Huang, and Rongrong Ji. 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*, pages 138–154. Springer, 2020.

[37] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[38] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

[39] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

[40] Xumeng Han, Longhui Wei, Zhiyang Dou, Zipeng Wang, Chenhui Qiang, Xin He, Yingfei Sun, Zhenjun Han, and Qi Tian. Vimoe: An empirical study of designing vision mixture-of-experts. *arXiv preprint arXiv:2410.15732*, 2024.

[41] Jihai Zhang, Xiaoye Qu, Tong Zhu, and Yu Cheng. Clip-moe: Towards building mixture of experts for clip with diversified multiplet upcycling. *arXiv preprint arXiv:2409.19291*, 2024.

[42] Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023.

[43] Kemal Oksuz, Selim Kuzucu, Tom Joy, and Puneet K Dokania. Mocae: Mixture of calibrated experts significantly improves object detection. *arXiv preprint arXiv:2309.14976*, 2023.

[44] Yash Jain, Harkirat Behl, Zsolt Kira, and Vibhav Vineet. Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. *Advances in Neural Information Processing Systems*, 36:69625–69637, 2023.

[45] Xin Wang, Fisher Yu, Lisa Dunlap, Yi-An Ma, Ruth Wang, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. Deep mixture of experts via shallow embedding. In *Uncertainty in artificial intelligence*, pages 552–562. PMLR, 2020.

[46] Leonardo Rossi, Vittorio Bernuzzi, Tomaso Fontanini, Massimo Bertozzi, and Andrea Prati. Swin2-mose: A new single image supersolution model for remote sensing. *IET Image Processing*, 19(1):e13303, 2025.

[47] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. *Advances in Neural Information Processing Systems*, 36:41693–41706, 2023.

[48] David Keetae Park, Seungjoo Yoo, Hyojin Bahng, Jaegul Choo, and Noseong Park. Megan: Mixture of experts of generative adversarial networks for multimodal image generation. *arXiv preprint arXiv:1805.02481*, 2018.

[49] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2human: Text-driven controllable human image generation. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022.

[50] Chenqi Kong, Anwei Luo, Peijun Bao, Yi Yu, Haoliang Li, Zengwei Zheng, Shiqi Wang, and Alex C Kot. Moe-ffd: Mixture of experts for generalized and parameter-efficient face forgery detection. *arXiv preprint arXiv:2404.08452*, 2024.

[51] Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma. Adaptive mixture of experts learning for generalizable face anti-spoofing. In *Proceedings of the 30th ACM international conference on multimedia*, pages 6009–6018, 2022.

[52] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[54] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[55] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, 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*, pages 10492–10502, 2021.

[56] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. 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.

[57] S. Sengupta, J.C. Cheng, C.D. Castillo, V.M. Patel, R. Chellappa, and D.W. Jacobs. Frontal to profile face verification in the wild. In *IEEE Conference on Applications of Computer Vision*, February 2016.

[58] Tianyue Zheng and Weihong Deng. 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.

[59] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.

[60] Tianyue Zheng, Weihong Deng, and Jiani Hu. Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments. *arXiv preprint arXiv:1708.08197*, 2017.

[61] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, Jocelyn Adams, Tim Miller, Nathan Kalka, Anil K Jain, James A Duncan, Kristen Allen, et al. Iarpa janus benchmark-b face dataset. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 90–98, 2017.

[62] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.

[63] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[64] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4362–4371, 2017.

[65] Weidi Xie and Andrew Zisserman. Multicolumn networks for face recognition. *arXiv preprint arXiv:1807.09192*, 2018.

[66] Yichun Shi and Anil K Jain. Probabilistic face embeddings. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6902–6911, 2019.

[67] X Liu, Z Guo, S Li, L Kong, P Jia, J You, and B Kumar. Permutation-invariant feature restructuring for correlation-aware image set-based recognition. 2019 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 4985–4995, 2019.

[68] Sixue Gong, Yichun Shi, and Anil Jain. Low quality video face recognition: Multi-mode aggregation recurrent network (marn). In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 1027–1035, 2019.

[69] Sixue Gong, Yichu Shi, Nathan D Kalka, and Anil K Jain. Video face recognition: Component-wise feature aggregation network (c-fan). In *2019 International Conference on Biometrics (ICB)*, pages 1–8. IEEE, 2019.

[70] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6817–6826, 2020.

[71] Minchul Kim, Yiyang Su, Feng Liu, Anil Jain, and Xiaoming Liu. Keypoint relative position encoding for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 244–255, 2024.

Appendix

As part of the appendix, we present the following as an extension to the ones shown in the paper:

- Backbone Ablation (Section A)
- Performance with Data Scaling (Section B)
- Additional Implementation Details (Section C)
- Expert Activation Maps (Section D)
- Failure Case Analysis (Section E)
- Limitations and Future Work (Section F)
- Social Impact Statement (Section G)
- Ethical Impact Statement (Section H)

A Backbone Ablation

To evaluate the backbone-agnostic nature of FaceMoE, we conduct experiments using both the standard Vision Transformer (ViT-B) and the hierarchical Swin Transformer (Swin-B). Table A presents performance results across four challenging benchmarks: IJB-B and IJB-C (TAR at FAR = 10^{-4}), TinyFace (Rank-1), and BRIAR Protocol 3.1 (Rank-1/5/20). The results lead to four key observations. First, FaceMoE integrates seamlessly with both ViT-B and Swin-B architectures without requiring any architecture-specific modifications, highlighting its generality. Second, FaceMoE-equipped models retain performance on IJB-B and IJB-C that is comparable to the ViT-B baseline, demonstrating that the Mixture-of-Experts routing mechanism preserves the generalizable features learned during pretraining. Third, FaceMoE consistently improves performance on difficult benchmarks, including an approximately 2.3% absolute increase in Rank-1 accuracy on TinyFace and a notable 15.8% gain on BRIAR Protocol 3.1 (Rank-1). Finally, combining FaceMoE with the hierarchical Swin-B backbone yields further performance improvements, particularly under stringent evaluation settings, such as a 1.72% increase in Rank-1 accuracy on BRIAR. These findings collectively confirm that FaceMoE is inherently backbone-agnostic, maintains pretrained discriminative capacity, and significantly enhances robustness in low-FAR and low-resolution face recognition scenarios.

Backbone	**IJBB**	**IJBC**	**TinyFace**	**BRIAR Protocol 3.1**		
	e-4	**e-4**	**Rank-1**	**Rank-1**	**Rank-5**	**Rank-20**
ViT-B	95.18	96.87	73.57	55.59	63.44	72.76
ViT-B (FaceMoE)	89.75	92.08	75.85	71.34	80.24	89.20
Swin-B (FaceMoE)	93.27	95.28	76.18	73.06	82.18	89.03

Table 3: Results of FaceMoE with ViT-B backbone on IJBB, IJBC, TinyFace, and BRIAR Protocol 3.1. FaceMoE works for all kind of transformer backbones.

B Performance with Data Scaling

Pretraining Dataset	**IJBB**	**IJBC**	**TinyFace**	**BRIAR Protocol 3.1**		
	e-4	**e-4**	**Rank-1**	**Rank-1**	**Rank-5**	**Rank-20**
WebFace4M	93.27	95.28	76.18	73.06	82.18	89.03
WebFace12M	**93.77**	**95.66**	**76.42**	**74.77**	**83.36**	**90.56**

Table 4: Performance of FaceMoE improves with increase in pre-training dataset size.

When we increase the size of the pre-training dataset from WebFace4M to WebFace12M, FaceMoE's performance consistently improves across a spectrum of face recognition benchmarks. On the IJBB protocol at a FAR of $1e_{-4}$ (after fine-tuning on TinyFace), we observe a gain from 93.27% to 93.77%.

A similar trend holds on IJBC (also after TinyFace fine-tuning), where accuracy at the same operating point increases by 0.38, from 95.28% to 95.66%. Even on the challenging TinyFace dataset—where both pre-trained models are further fine-tuned on TinyFace—the Rank-1 accuracy climbs from 76.18% to 76.42%, demonstrating that additional data yields measurable benefits under difficult, low-resolution conditions. The gains are most pronounced on the BRIAR Protocol 3.1 benchmarks (after BRIAR fine-tuning), with Rank-1 accuracy improving by 1.71 (from 73.06% to 74.77%), Rank-5 by 1.18 (from 82.18% to 83.36%), and Rank-20 by 1.53 (from 89.03% to 90.56%). These results not only confirm that FaceMoE continues to harness extra data to push its recognition capabilities forward, but also illustrate strong preservation of pre-trained knowledge through successive fine-tuning stages.

All data scaling results are shown in Table B, where IJBB and IJBC results are reported after fine-tuning on TinyFace; the TinyFace results likewise follow TinyFace fine-tuning; and the BRIAR Protocol 3.1 results are after BRIAR fine-tuning. When the pre-training dataset is increased from WebFace4M to WebFace12M, FaceMoE's performance improves uniformly across all benchmarks. On IJBB at a FAR of 1×10^{-4}, the TAR rises from 93.27% to 93.77% ($+0.50$). Similarly, on IJBC under the same operating point, TAR increases by 0.38, from 95.28% to 95.66%. On TinyFace, Rank-1 accuracy climbs from 76.18% to 76.42% ($+0.24$), demonstrating benefits even under low-resolution conditions. The most substantial gains appear on BRIAR Protocol 3.1: Rank-1 improves by 1.71 (from 73.06% to 74.77%), Rank-5 by 1.18 (from 82.18% to 83.36%), and Rank-20 by 1.53 (from 89.03% to 90.56%). These results confirm that scaling the pre-training data both enhances FaceMoE's recognition accuracy and preserves its learned representations after fine-tuning on low-resolution face recognition dataset.

Several architectural and training factors contribute to the successful scaling of data. First, the mixture-of-experts design enables conditional computation. Although the overall model capacity increases with the addition of more experts, each input activates only a small subset of them. This means that tripling the dataset size does not significantly increase the computational cost for each example. At the same time, the larger pool of experts allows the model to capture more subtle variations in the data, such as differences in pose, lighting, and demographic diversity present in the WebFace12M dataset. As a result, FaceMoE learns a richer set of feature subspaces, which enhances its robustness on both standard and challenging benchmarks, even after fine-tuning on downstream datasets.

Moreover, sparse routing serves as an implicit regularizer. FaceMoE updates only a fraction of the model parameters in each mini-batch, which helps reduce co-adaptation among experts and protects against overfitting, even as the dataset continues to grow. This built-in regularization becomes increasingly valuable when training on tens of millions of images, as it ensures that each expert develops a distinct specialization rather than converging into redundant representations. In addition, the computational efficiency of mixture-of-experts models allows for high model capacity while keeping the floating point operations per example manageable. This efficiency enables longer and more thorough training within a fixed compute budget, allowing FaceMoE to fully leverage the extensive data available in WebFace12M. Together, these factors explain why increasing the size of the pre-training dataset leads to consistent and cost-effective improvements in FaceMoE's recognition performance during both pre-training and downstream fine-tuning.

C Additional Implementation Details

These are the additional details provided in addition to the ones mentioned in the main paper. Our base architecture for all experiments is the Swin-B (Swin Transformer - Base), which serves as the backbone for the FaceMoE model. To provide a rough estimate of computational requirements, we report training times for various configurations of the number of experts (N) and the number of active experts per token (k). These estimates are not intended for comparison, as the experiments were conducted on both NVIDIA A6000 (48GB) and A5000 (24GB) GPUs, leading to variability in runtime. Specifically, training times (in hours) are approximately: 49 for (N=2, k=1), 57 for (N=3, k=1), 81 for (N=3, k=2), 120 for (N=3, k=3), 49 for (N=4, k=1), 50 for (N=4, k=2), and 88 for (N=4, k=3). To ensure a consistent and fair evaluation, we retrained the CosFace, ArcFace, and AdaFace baselines. For other baselines, we report results as presented in their respective original publications. All models and experiments are implemented in PyTorch and run across eight GPUs.

D Expert Activation Maps

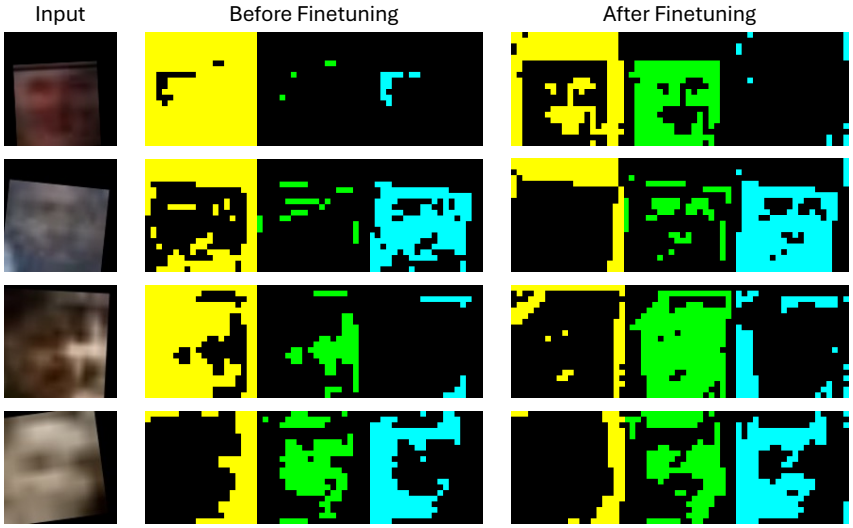

Figure 6: **Expert activation maps before and after TinyFace finetuning.** Each row shows the spatial activations of all k experts on a input face image, before and after TinyFace finetuning. (Left) After pretraining on WebFace4M, experts exhibit broadly overlapping activations focusing on general facial regions (eyes, nose bridge, mouth outline). (Right) Following TinyFace finetuning, experts specialize on distinct, localized cues (eye corners, nose shape, cheek textures, etc.), yielding complementary attention patterns better suited to low-resolution face recognition.

To gain insight into how each expert specializes before and after TinyFace finetuning, we visualize their spatial activation patterns on a few facial images, as shown in Figure 6. Each row presents the activations of all k experts for a single input image.

Pretraining on WebFace4M: Before undergoing any adaptation to the TinyFace dataset, the model is pretrained for face recognition using the large-scale WebFace4M dataset. During this phase, all experts learn from a diverse collection of face images that vary in quality and pose, ranging from frontal to non-frontal views. As a result, their activation maps tend to highlight broad, coarse-grained regions, such as the overall outline of the face, the contours of the eyes, and the mouth area. There is substantial overlap between the activation patterns of different experts, suggesting that in the absence of further specialization, the experts tend to redundantly focus on the most generally discriminative facial features, such as the eyes and the bridge of the nose. These features remain consistently informative across a wide range of identities and imaging conditions.

After TinyFace Finetuning: Following finetuning on the TinyFace dataset, which consists of low-resolution face crops extracted from unconstrained scenes, the experts begin to capture more localized and complementary features. The activation maps demonstrate that individual experts now respond to specific subregions or patterns. Some experts focus closely on areas such as the eye corners and eyelid textures, which are particularly important in low-resolution scenarios. Others concentrate on features such as the shape of the nose or the contours of the mouth. Additional experts respond to compound patterns, including shadows on the cheeks or the silhouettes of ears. This diversity in focus reflects the model's adaptation to the characteristics of the TinyFace dataset. By distributing representational capacity across multiple experts, the network learns that fine-grained, region-specific textural cues are essential for distinguishing identities when the global structural features of the face are degraded due to low resolution.

The transition from broadly overlapping activations in the WebFace4M pretraining phase to highly specialized and non-redundant activation maps after TinyFace finetuning highlights the effectiveness of the MoE architecture for domain adaptation. In low-resolution settings, relying on a single shared backbone imposes a trade-off between capturing global structures and preserving fine-grained local details. In contrast, the MoE framework enables different sub-networks to allocate their representational capacity to the most reliable cues for the target domain. First, the model demonstrates

robustness to resolution degradation. Experts that are tuned to textural patterns, such as the micro-structure of skin around the eyes, retain their discriminative ability even when the overall facial shape becomes indistinct. Second, the architecture facilitates the integration of complementary evidence. By aggregating signals from multiple specialized experts, the model can combine weak, localized features into a coherent and robust identity representation. Finally, the approach allows for efficient adaptation. Only a subset of experts needs to specialize deeply in the new domain, while others can maintain their generalist knowledge from pretraining. This division of labor ensures a balanced trade-off between plasticity and stability.

These activation patterns offer clear evidence that finetuning on low-resolution dataset induces functional specialization among experts, enabling the model to perform effectively in challenging, low-resolution face recognition tasks.

E Failure Case Analysis

To diagnose the remaining weaknesses of our FaceMoE model, we conducted a detailed examination of representative failure cases on the BRIAR probe set as shown in Figure 7. We identified five dominant scenarios that consistently lead to recognition errors. First, **extremely low-resolution** face crops, typically below approximately 8×8 pixels, contain too little texture or shape information for reliable matching. This causes the expert ensemble's activations to become noisy and prone to errors. Second, **extreme head poses**, such as profiles or tilts greater than 60 degrees, often result in facial landmarks moving outside the visible region. In these situations, experts trained on frontal-view patterns perform poorly. Third, **heavy occlusion** caused by items like masks, caps, or scarves can obscure important facial regions. As a result, the experts struggle to extract meaningful unoccluded features, which increases confusion with other identities. Fourth, **atmospheric turbulence**, including visual distortions such as heat shimmer and motion blur that are common in long-range surveillance, disrupts the spatial consistency of facial features. These effects fragment the activation maps and reduce the model's ability to form coherent representations. Finally, **non-frontal views**, where subjects never present a clear frontal face during a sequence, prevent the model from obtaining a stable canonical reference. Consequently, even viewpoint-specialized experts are unable to generate consistent embeddings, leading to recognition failures. These failure modes illustrate that, while FaceMoE is effective in handling low-resolution images, it remains vulnerable to conditions that obscure or dynamically distort facial information.

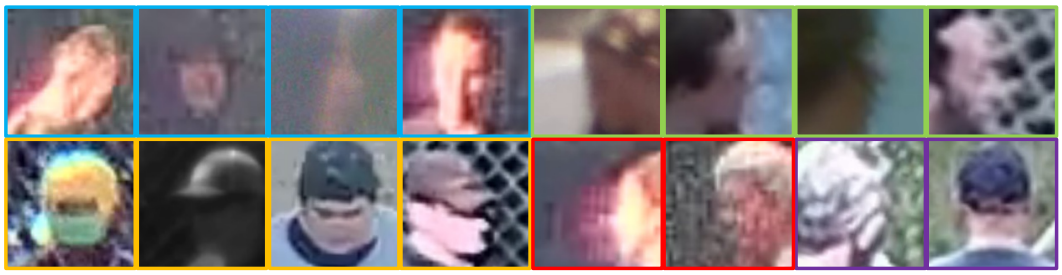

Figure 7: Failure Case Analysis of FaceMoE model on the BRIAR dataset.

F Limitations and Future Work

Our training data, WebFace4M [55], is predominantly composed of Western, young, and light-skinned subjects. We have not yet incorporated balanced sampling, debiasing loss functions, or demographic-specific experts, which means the model may amplify existing biases. While Mixture-of-Experts (MoE) architectures are typically used to scale model capacity efficiently, their application in face recognition introduces unique challenges. We observe that increasing the number of experts (N) can lead to over-fragmentation and routing instability, which may negatively affect performance. Addressing these issues remains an important area for future work.

G Social Impact Statement

The proposed work, FaceMoE, presents a transformer-based Mixture of Experts (MoE) architecture that significantly advances low-resolution face recognition (LR-FR). FaceMoE enhances recognition performance on degraded or surveillance-quality imagery, offering the potential to improve operational effectiveness in domains such as public safety, disaster response, border control, and missing persons investigations. These improvements enable faster and more accurate identification in scenarios where traditional face recognition systems often underperform, particularly in time-sensitive or resource-constrained environments.

Beyond technical improvements, the broader societal implications of these advancements merit careful consideration. As face recognition systems become increasingly capable of identifying individuals from poor-quality images, their deployment in everyday settings such as public transit, city surveillance, or consumer electronics is likely to accelerate. This trend could contribute to a societal shift in which continuous identity tracking becomes normalized, potentially eroding expectations of anonymity and reshaping perceptions of privacy in public spaces. The widespread presence of such systems may also influence individual behavior and social engagement, particularly in communities that are already subject to heightened surveillance.

Furthermore, access to advanced recognition systems like FaceMoE may not be distributed evenly. Organizations with greater financial and technical resources are more likely to benefit from such technologies, which could deepen existing disparities in areas such as law enforcement, national security, and institutional capacity. Public trust in face recognition systems depends not only on their technical performance but also on how transparently and equitably they are implemented. To ensure that FaceMoE contributes positively to society, its deployment in real-world applications must be supported by inclusive access, meaningful public dialogue, and policies that emphasize fairness, accountability, and the protection of civil liberties.

H Ethical Impact Statement

In this research, we have carefully addressed the ethical implications surrounding face recognition technology, particularly focusing on issues of privacy, surveillance, and potential biases. Our model was trained on publicly available datasets: WebFace4M and WebFace12M [55], acquired through signing the official license agreement. For benchmarking, we utilized IJB-B [61], IJB-C [62], IJB-S [7], BRIAR [20], and TinyFace [4], which contain diverse, mixed-quality, and low-resolution images from real-world settings. These datasets were obtained through official repositories and websites, ensuring adherence to ethical standards. Informed consent for publication was acquired for all subjects depicted in the paper, supporting ethical data use.

This research offers significant benefits within authorized security contexts, where accurate low-resolution face recognition enhances identification capabilities in challenging environments. When applied responsibly, these advancements contribute to security and enable legitimate monitoring efforts. Importantly, the model's design and training process adhere to standards that do not introduce risks beyond those inherent in traditional face recognition systems. However, we acknowledge the potential for misuse in unauthorized surveillance, profiling, or privacy infringements if deployed outside controlled, ethical frameworks. Our work aims to support face recognition for responsible use within authorized security settings, while recognizing that unintended applications or misinterpretations could lead to societal issues, such as privacy erosion or biased treatment of certain groups. By proactively addressing these considerations, we seek to mitigate risks associated with the model's deployment and advocate for ethical oversight to prevent misuse.

Ethical considerations for human subjects and data usage were fully respected. This research relies solely on existing datasets and no new consent was required. These datasets are approved for research use, ensuring adherence to ethical data standards. No individuals were recruited which eliminates the need for compensation. The datasets do not predominantly include vulnerable populations, such as minors, elderly individuals, or other at-risk groups, instead representing a standard demographic spectrum. Given our commitment to ethical standards, this research presents minimal risk to individuals while advancing low-resolution face recognition technology.