APPENDIX

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

A RELATED WORK: MOE IN FACE ANALYSIS TASKS

AMEL Chen et al. (2025) combines a shared expert with low-rank adapted attack-specific experts and dynamically aggregates them to improve robustness to post-processing distortions in audio spoofing detection. MoE-FFD Kong et al. (2025) fuses transformer's global features with CNN-style local priors and uses a gating scheme to dynamically select the most relevant forgery expert, boosting generalization across face forgery types. However, the proposed FaceMoE differes from AMEL Chen et al. (2025) and MoE-FFD Kong et al. (2025) in several aspects, along with the fact that they target different tasks, such as:

- **Expert Integration:** *FaceMoE* incorporates multiple FFN-based experts within the MLP layers of a transformer encoder, enabling semantic specialization for different facial regions. *AMEL* introduces Domain-Specific Experts (DSEs), which are lightweight residual blocks appended to a shared CNN backbone, each modeling features from a specific source domain. *MoE-FFD* adopts parameter-efficient experts using LoRA and Adapter modules, allowing expert injection without modifying the backbone, and is optimized for face forgery detection.
- **Routing Strategy:** *FaceMoE* employs a learned top-k router at the token level, directing face patches to a subset of specialized experts based on resolution and semantic cues. *AMEL* uses Dynamic Expert Aggregation (DEA) at the sample level, computing soft aggregation weights across domain experts based on domain similarity. *MoE-FFD* utilizes a top-1 gating mechanism to assign each input to the most relevant forgery expert, with routing performed at the sample level for efficiency.
- **Training Paradigm:** *FaceMoE* enables full fine-tuning of the transformer model, complemented by auxiliary losses (router z-loss and load-balancing) to promote expert specialization and training stability. *AMEL* is trained using a meta-learning strategy that simulates domain shifts across source domains; it combines standard classification and depth supervision with a feature consistency loss. *MoE-FFD* adopts a parameter-efficient fine-tuning (PEFT) strategy, keeping the backbone frozen while training only the inserted expert modules, thereby reducing training overhead.

B Additional Results, Analysis and Ablation Studies

B.1 Expert Specialization

Facial Region	Expert 0	Expert 1	Expert 2	Dominant Expert
Eyes	**76.2**	12.4	11.4	Expert 0
Nose	**68.5**	15.7	15.8	Expert 0
Mouth	**61.0**	22.1	16.9	Expert 0
Forehead	13.2	**71.3**	15.5	Expert 1
Cheeks	12.5	**69.8**	17.7	Expert 1
Hair / Background	20.3	14.5	**65.2**	Expert 2
Chin / Jawline	25.6	18.4	**56.0**	Expert 2

Table B.1: Routing frequency (%) of each expert across different semantic facial regions. Results averaged over 10,000 samples.

Facial Region	Expert 0	Expert 1	Expert 2	Dominant Expert
Eyes	**81.7**	10.6	7.7	Expert 0
Nose	**78.1**	12.5	9.4	Expert 0
Mouth	**71.6**	16.0	12.4	Expert 0
Forehead	49.3	**31.1**	19.6	Expert 1
Cheeks	52.5	**27.8**	19.7	Expert 1
Hair / Background	40.2	17.3	**42.5**	Expert 2
Chin / Jawline	38.9	18.7	**42.4**	Expert 2

Table B.2: Pre-finetune Routing frequency (%) of each expert across different semantic facial regions. Results averaged over 10,000 samples.

To strengthen our claim regarding expert specialization, we conducted an additional analysis quantifying the consistency and evolution of expert assignments across facial regions. We computed the routing frequency of each expert before and after finetuning on a sample of 10,000 face images. Table B.2 reports the routing behavior of the pre-finetuned model, while Table B.1 shows the corresponding results after finetuning. The comparison reveals two important trends:

Finetuning strengthens and sharpens spatial specialization. Before finetuning, the router exhibits only weak spatial bias: Expert 0 is mildly preferred for central regions (eyes, nose, mouth), Expert 1 receives a moderate share of tokens from the forehead and cheeks, and Expert 2 shows a slight preference for peripheral or high-frequency regions such as hair and the jawline. However, these tendencies remain relatively diffuse, as reflected by the more evenly distributed routing frequencies across experts.

After finetuning, these patterns become markedly more pronounced. Expert 0 becomes the dominant processor for identity-rich central regions (eyes, nose, mouth), consistently exceeding 60-76% routing frequency. Expert 1 specializes in smoother, low-frequency regions such as the forehead and cheeks, each receiving over 69-71% assignments. Expert 2 emerges as the primary expert for high-frequency or peripheral structures, including hair, background, and the jawline, with routing frequencies between 56-65%. These results indicate that finetuning drives the router toward strong and interpretable spatial specialization.

Experts become more complementary and disentangled. A direct comparison of the pre- and post-finetune distributions shows that finetuning reduces expert overlap and increases region-specific dominance. Whereas the pre-finetuned model routes many regions across experts in a relatively mixed manner (e.g., cheeks: 52.5/27.8/19.7), the finetuned model exhibits clear expert preferences (e.g., cheeks: 12.5/69.8/17.7). This shift demonstrates that the router learns to assign experts based on semantic and frequency characteristics, yielding complementary specialization across landmark, low-frequency, and high-frequency regions.

B.2 Expert Specialization Mechanism in FaceMoE

In this section, we provide deeper insight into the mechanism through which the experts in FaceMoE specialize in distinct facial regions. While the main paper introduces the motivation behind incorporating sparse FFN experts, here we examine how this specialization implicitly emerges during training, how it is reinforced by the top-k router, and how it manifests both quantitatively and qualitatively.

FaceMoE integrates a sparse mixture-of-experts (MoE) layer within each MLP block of the transformer. Unlike a dense FFN, which applies the same transformation to all tokens, the MoE layer enables *conditional computation*, allowing each token to be processed by only a subset of experts. This design naturally promotes expert specialization. Facial tokens exhibit diverse semantic and frequency characteristics, such as high-frequency regions (edges, contours, hair boundaries), smooth low-frequency regions (cheeks, forehead), and structured landmark regions (eyes, nose, mouth). Since the router computes routing logits through a linear projection of token embeddings, tokens from these different regions generate *distinct* routing patterns even early in training. This asymmetry initiates the specialization process.

A positive feedback loop then emerges: tokens from a specific region (e.g., the eyes) initially receive slightly higher routing logits for a particular expert, leading them to be repeatedly routed to that expert. As a result, the expert's parameters gradually specialize to model the statistical patterns characteristic of those tokens. In parallel, the router learns to reinforce these region–expert correspondences. This dynamic ultimately produces experts that specialize in semantically coherent facial subsets.

Conditional Routing Behavior For the default configuration with 3 experts and top-2 routing, we observe that routing probabilities converge to region-consistent patterns. Empirically:

$$P(E_i \mid R_t = r) \gg P(E_i \mid R_t \neq r), \qquad r \in \{\text{high-frequency}, \text{low-frequency}, \text{landmarks}\}.$$

indicating that each expert becomes the preferred destination for a specific category of tokens.

Quantitative Evidence of Specialization Appendix B.1 includes statistics of token-assignment distribution showing that:

- routing variance decreases over training,
- each expert receives consistent token subsets,
- spatial patterns on the face correspond to stable expert clusters.

Qualitative Evidence via Activation Maps Activation maps provided in Appendix D demonstrate:

- Spatial consistency: each expert highlights distinct facial zones;
- Semantic coherence: landmark-oriented experts focus on eyes/nose/mouth;
- Resolution-aware specialization: LR images trigger higher reliance on experts specializing in coarse structural cues.

B.3 Resolution Ablation

We conduct a resolution ablation study by varying the resolution of the test images and observe only a minimal drop in performance across resolutions as shown in Table B.3. This result reinforces FaceMoE's capability to effectively handle inputs of varying resolutions.

B.4 Bias Analysis

We quantify the bias implications of our mixture-of-experts based FaceMoE architecture compared to the baseline to showcase that FaceMoE exhibits minimal bias. We conducted further experiments on LFW, CFP-FF, and AgeDB datasets. We used FaceXFormer Narayan et al. (2024) to infer age (0–19, 20–39, 40–59, 60+), gender (male, female), and race (Black, Latino/Hispanic, Middle Eastern, Asian, White) labels. To evaluate fairness, we adopted the Selective Ratio (SeR) and Degree of Bias (DoB) as our metrics.

Dataset	8x8	10x10	12x12	16x16	32x32	48x48	64x64	96x96
LFW	80.75	86.98	91.71	96.06	99.61	99.68	99.68	99.73
CFP-FP	62.38	68.45	72.94	80.87	94.75	96.22	96.57	96.72
CPLFW	65.20	71.33	77.18	82.26	92.35	93.01	93.23	93.28
AgeDB-30	56.23	59.13	63.48	70.51	92.53	95.48	96.06	96.25
CALFW	63.18	68.58	74.66	80.31	93.75	94.76	95.10	95.43
CFP-FF	73.24	78.71	83.44	90.32	99.02	99.68	99.67	99.65

Table B.3: Accuracy (%) across different image resolutions on various datasets.

The results, summarized in Table B.4, demonstrate that FaceMoE not only achieves superior performance but also results in a fairer model with reduced bias across age, gender, and racial attributes compared to the baseline.

Dataset	Model	Age		Gender		Race	
		SeR	DoB	SeR	DoB	SeR	DoB
LFW	Swin-B	0.95	2.16	0.99	0.25	0.80	8.07
	FaceMoE	0.95	2.20	0.99	0.07	0.84	6.30
CFP FF	Swin-B	0.93	3.29	0.99	0.27	0.86	5.52
	FaceMoE	0.93	3.28	0.99	0.25	0.86	5.54
AgeDB	Swin-B	0.99	0.34	0.99	0.15	0.77	9.26
	FaceMoE	0.99	0.28	0.99	0.12	0.77	9.87

Table B.4: Performance comparison of Swin-B and FaceMoE across different datasets for Age, Gender, and Race attributes.

The intrinsic reason behind FaceMoE's improved fairness lies in its mixture-of-experts design, which encourages different experts to specialize in complementary facial regions and frequency patterns. This specialization allows the router to dynamically select the most informative experts for each input, particularly beneficial when demographic groups differ in blur level, pose variation, skin texture, or age-related changes. Consequently, the model avoids over-reliance on any single facial attribute that may be demographically sensitive, leading to more stable SeR/DoB scores across groups. In practice, FaceMoE appears fairer because (1) sparse expert activation mitigates biased drift during fine-tuning, and (2) expert diversity distributes representational responsibility across multiple specialized pathways rather than amplifying group-specific biases.

B.5 Comparison with other MoE variants

We conducted ablations on multiple MoE configurations to evaluate their effectiveness for low-resolution face recognition as shown in Table B.5:

- **Shared MoE:** A single shared MLP expert activated for all tokens. This limits specialization and makes the model overly rigid when adapting to low-resolution data, leading to degraded performance.
- **LoRA FFN:** Uses LoRA experts instead of dense FFNs. While lightweight, it lacks sufficient representational capacity and is prone to catastrophic forgetting.
- **LoRA FFN + Attn:** Extends LoRA to Q, K, V projections. Although slightly better, it still lacks expressiveness for resolution-aware specialization.
- **FaceMoE (Ours):** Incorporates multiple full-capacity FFN experts in transformer MLP layers, combined with a token-wise top-k router. This enables spatially-aware, resolution-sensitive expert activation, leading to significantly better performance.

Method	Rank-1	Rank-5	Rank-10
Shared MoE	62.71	69.39	73.92
LoRA FFN	43.61	52.62	59.81
LoRA FFN + Attn	44.98	53.37	60.48
FaceMoE (Ours)	**76.18**	**79.69**	**81.75**

Table B.5: Performance comparison of different MoE variants on TinyFace.

B.6 Impact of each component

We provide an ablation study in Table B.6 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 B.6: Ablation study showing the impact of top-k router, MoE, and auxiliary loss on FaceMoE performance. Results are shown on TinyFace dataset

B.7 Large N and random expert assignment

To evaluate whether the observed performance gains stem from meaningful expert specialization or simply increased model capacity, we conducted controlled experiments with (a) random expert assignment and (b) increased number of experts (N = 8). The results are shown in Table B.7.

As evident, using random expert assignment (i.e., bypassing learned routing) results in a performance drop across all metrics (e.g., Rank-1: 75.40 vs. 76.18), suggesting that the learned routing mechanism does contribute meaningfully to the model's discriminative ability. More notably, increasing the number of experts to N = 8 leads to training collapse (Rank-1: 2.31), highlighting that larger expert sets can destabilize training without proper balancing, as discussed in Figure 5(a)(b)). This instability stems from expert under-utilization and routing noise.

These results collectively support our claim that expert specialization, when properly routed and regularized, is fundamental to the model's performance, not merely a byproduct of added parameters.

	Rank-1	Rank-5	Rank-10
FaceMoE	76.18	79.69	81.75
Random Expert Assignment	75.40	78.46	80.63
Large N (N=8)	2.31	3.82	6.04

Table B.7: Performance comparison of Random Experts and Large N on TinyFace dataset.

B.8 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 B.8 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 B.8: 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.9 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 B.9: 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.9, 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.

B.10 Performance under Synthetic Degradations

Method	LFW	CFP-FF	AgeDB-30	Expert 0	Expert 1	Expert 2
FaceMoE	99.75	99.86	97.45	33.4	33.6	33.0
Gaussian std = 1	99.71	99.81	97.30	33.2	35.2	31.6
Gaussian std = 5	76.53	69.77	55.90	32.7	37.0	30.3
JPEG 30%	**99.6**	**99.65**	**96.93**	33.5	34.6	31.9

Table B.10: Performance under synthetic degradations

We perform a stress test on FaceMoE under synthetic degradations. We synthetically apply Gaussian blur and JPEG compression to the probe images while keeping the gallery fixed, and we report the verification accuracy on LFW/CFP-FP/AgeDB-30 and retrieval performance on TinyFace for occlusion robustness. As shown in Table B.10, mild degradations (Gaussian $\sigma = 1$, JPEG 30%) lead to only marginal changes, indicating that the routing mechanism remains stable and continues to select suitable experts even when image quality is moderately reduced. Under severe blur ($\sigma = 5$), performance drops substantially particularly on AgeDB-30 consistent with the fact that heavy low-pass filtering removes discriminative identity cues that no expert can fully compensate for. The routing statistics adapt and show increased activation of experts ($\approx 37\%$) specialized for low-frequency representations, confirming the hypothesized adaptive behavior. For occlusion, we evaluate on the TinyFace benchmark, which includes masks over the eyes, mouth, and nose. These landmark occlusions severely reduce the available identity information, as they block key facial regions used for recognition, which in turn leads to a significant drop in absolute performance. Although absolute performance is lower due to the extreme occlusions, the model maintains stable rank-1/5/20 trends.

These results demonstrate that FaceMoE adapts its routing under synthetic degradations, and performance only degrades significantly when identity information becomes intrinsically unrecoverable.

B.11 Inference Computation Analysis

We perform an inference computation analysis and 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. All experiments were conducted on the same dataset (161,599 images) with a batch size of 800, and include the full routing overhead of our

Metric	RTX A5000	RTX A6000
Total Images	161,599	161,599
Batch Size	800	800
Average Batch Latency (ms)	8109.92	6170.72
Per-Image Latency (ms)	10.27	7.80
Throughput (fps)	97.32	128.19
Peak GPU Memory Usage (GB)	10.40	10.40

Table B.11: Inference latency, throughput, and memory metrics (including router overhead) on RTX A5000 and RTX A6000.

method. As shown in the Table B.11, the A6000 provides a substantial improvement in throughput (128.19 fps vs. 97.32 fps) and reduced average batch latency, while peak GPU memory usage remains identical across GPUs (10.40 GB). These results demonstrate that our method scales efficiently across hardware tiers and maintains practical latency/memory characteristics even on a constrained GPU such as the A5000.

B.12 Quantitative evidence of selective drift

Layer	CKA Similarity
patch_embed	0.999867
layers.0.blocks.0	0.998419
layers.0.blocks.1	0.996523
layers.1.blocks.0	0.995274
layers.1.blocks.1	0.995154
layers.1.blocks.2	0.994973
layers.1.blocks.3	0.995113
layers.1.blocks.4	0.994977
layers.1.blocks.5	0.995119
layers.1.blocks.6	0.995238
layers.1.blocks.7	0.995003
layers.1.blocks.8	0.994763
layers.1.blocks.9	0.995177
layers.1.blocks.10	0.994537
layers.1.blocks.11	0.994277
layers.1.blocks.12	0.993579
layers.1.blocks.13	0.993211
layers.1.blocks.14	0.991946
layers.1.blocks.15	0.990840
layers.1.blocks.16	0.989383
layers.1.blocks.17	0.983919
layers.2.blocks.0	0.977410
layers.2.blocks.1	0.808624
norm	0.877010
feature_layer	0.867713

Table B.12: CKA similarity for each layer

In this subsection, we provide quantitative evidence of reduced forgetting and selective drift which makes our paper stronger. To address this, we performed additional analyses comparing the HR-pretrained model with the LR-finetuned FaceMoE model.

(1) CKA-based representational drift. We compute CKA similarity layer-by-layer to measure representational changes. As shown in Table B.12, most layers maintain extremely high similarity (0.99+), including patch embedding and early/mid transformer blocks, indicating negligible forgetting. Drift gradually increases only in deeper layers (e.g., layers.1.blocks.17: 0.984; layers.2.blocks.0: 0.977),

Layer	CKA Similarity
layers.2.blocks.1	0.808624
feature_layer	0.867713
norm	0.877010
layers.2.blocks.0	0.977410
layers.1.blocks.17	0.983919

Table B.13: Top 5 Most Changed Layers (Lowest CKA).

Expert	L2 Shift (Magnitude)	Expert Shift (%)
Expert 0	45.9942	7.9835%
Expert 1	46.0058	8.0334%
Expert 2	45.6348	8.0775%

Table B.14: L2 shift and approximate expert shift percentage.

with the largest adaptation occurring in layers.2.blocks.1 (0.8086) and the final feature layer (0.8677). These are precisely the layers responsible for high-level identity semantics, supporting our claim that FaceMoE preserves foundational HR features while adapting selectively for LR data.

(2) Localizing drift (lowest-CKA layers). The five most changed layers (Table B.13) are exclusively in the deepest stage and output head. This indicates targeted high-level adaptation rather than global drift, supporting our claim that forgetting is minimized and adaptation is concentrated on identity-semantic layers.

(3) Expert parameter update We also measure the L2 parameter shift for each expert. As reported in Table B.14, all experts undergo only a small relative shift of 8%, despite being trainable during LR finetuning. This modest drift indicates that FaceMoE adapts sufficiently to the LR domain while preserving the majority of the HR-pretrained structure. The small magnitude of change across experts quantitatively supports our claim that MoE reduces forgetting by enabling controlled, localized adaptation rather than wholesale parameter updates.

B.13 Hyperparameter Sensitivity

λ	TinyFace			BRIAR		
	Rank-1	Rank-5	Rank-20	0.01%	0.10%	1%
1	76.09	79.82	81.66	42.27	61.53	81.22
5	76.48	0.06	0.06	42.56	61.61	81.52
9	76.31	79.83	82.24	42.30	61.68	81.43
10	76.18	79.69	81.75	42.36	61.47	81.27
11	76.42	79.90	82.00	42.56	61.52	81.62
15	76.18	79.77	82.10	42.46	61.38	81.21
100	76.31	79.66	82.05	42.34	61.52	81.41

Table B.15: Performance metrics for TinyFace and BRIAR across different λ values.

We perform a sensitivity analysis for the loss-weighting hyperparameter λ, which jointly scales the router z-loss and load-balancing loss in the total objective:

$$L_{\text{total}} = L_{\text{face}} + \lambda \left(L_z + L_{\text{balance}} \right).$$

The table above reports performance obtained by sweeping

$$\lambda \in \{1, 5, 9, 10, 11, 15, 100\}$$

on two datasets (TinyFace and BRIAR):

- **TinyFace:** Rank-1 accuracy ranges from 76.09% to 76.48% (a spread < 0.4 percentage points). Rank-5 and Rank-20 accuracies vary by only about 0.3 percentage points across the entire sweep.
- **BRIAR (Protocol 3.1):** TAR@FAR=0.01% ranges from 42.27% to 42.56%, TAR@0.10% from 61.38% to 61.68%, and TAR@1% from 81.21% to 81.62%. All metrics vary by roughly 0.4 percentage points or less.

These results show that performance is *not* brittle with respect to λ, even when varied over more than two orders of magnitude. In particular, there is a clear performance plateau for

$$\lambda \in [5, 15],$$

within which both TinyFace and BRIAR metrics remain effectively unchanged.

B.14 Routing Stability and Causality

Attack	Rank-1	Rank-5	Rank-20	Expert 0	Expert 1	Expert 2
FGSM	74.88	79.09	81.63	33.2	33.5	33.3
PGD	73.41	77.54	80.06	32.6	34.6	32.8
MIM	74.14	78.32	80.87	33.1	32.7	34.2

Table B.16: Performance comparison across attacks and experts.

	TinyFace		
	Rank-1	**Rank-5**	**Rank-20**
FaceMoE	76.18	79.69	81.75
Switch Expert	69.68	75.40	79.07
Drop Expert			
0	75.05	78.99	81.65
1	75.10	78.88	81.62
2	75.08	78.91	81.94
0, 1	69.68	74.83	78.75
1, 2	69.98	75.80	80.12
0, 2	68.56	74.38	78.13

Table B.17: Performance on TinyFace when dropping individual or pairs of experts, and switching experts.

We conducted experiments to show the performance across perturbations, and further performed experiments by dropping and switching experts, to strengthen our claim that the performance is achieved by the proposed design and not by increased capacity.

1. Routing is Stable Across Perturbations To assess stability, we measure token-to-expert assignments under several common perturbations (FGSM, PGD, MIM). As shown in Table B.16, the routing distribution across the three experts remains highly stable:

- Under all perturbations, expert usage stays nearly uniform (≈33% per expert), with $< 2\%$ deviation across attacks.
- Even stronger iterative attacks (PGD, MIM) do not cause expert collapse or oscillation. This consistency shows that the router is not sensitive to small input perturbations such as adversarial noise, and that token assignments converge to stable semantic regions, not noise-driven fluctuations. Therefore, routing behavior is robust and not an unstable byproduct of MoE capacity.

2. Controlled Expert Ablations Reveal Causal Contribution of Specialization To evaluate whether FaceMoE's performance stems from learned specialization rather than increased parameters, we run two sets of interventions as shown in Table B.17:

(a) Dropping Individual Experts

Removing any one expert while keeping model capacity nearly unchanged produces only a small drop ($\sim 1.0\%$) in Rank-1 relative to full FaceMoE:

- Expert 0 dropped: 75.05
- Expert 1 dropped: 75.10
- Expert 2 dropped: 75.08
- Full model: 76.18

This small but consistent degradation indicates that each expert contributes complementary information rather than redundant capacity. Since our router uses top-k routing with $k = 2$, every token is always processed by two experts, ensuring that even after removing one expert, at least one of the originally assigned specialists is still active. This limits the performance drop while still revealing the non-redundant contribution of each expert.

(b) Dropping Pairs of Experts (forcing single-expert routing)

When two experts are removed, routing collapses into a single FFN branch. This mirrors a standard transformer's feed-forward layer capacity but accuracy drops drastically:

- Experts {0,1}: 69.68
- Experts {1,2}: 69.98
- Experts {0,2}: 68.56

The $\sim 6 - 8\%$ absolute drop demonstrates that increased capacity alone cannot account for performance gains. If raw capacity were the cause, single-expert models (same depth, same FLOPs) would not collapse this sharply. Instead, this strongly supports specialization as the mechanism driving improvements.

Together, these interventions demonstrate a causal chain:

- Routing remains stable across perturbations (Table B.16).
- Experts are not interchangeable (Switch Expert $\rightarrow$ significant drop).
- Experts are not redundant (dropping experts reduces performance).

Thus, improvements are not attributable to mere extra parameters but arise from structured expert specialization and resolution-aware routing, as intended in the design.

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

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.

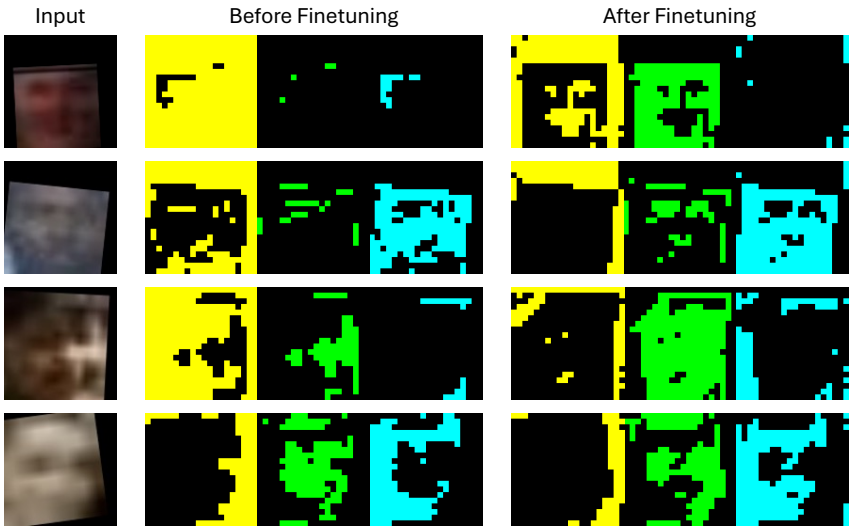

Figure 6: **Expert activation maps before and after TinyFace finetuning.** Each row shows the spatial activations of all k experts on a input face image, before and after TinyFace finetuning. (Left) After pretraining on WebFace4M, experts exhibit broadly overlapping activations focusing on general facial 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.

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

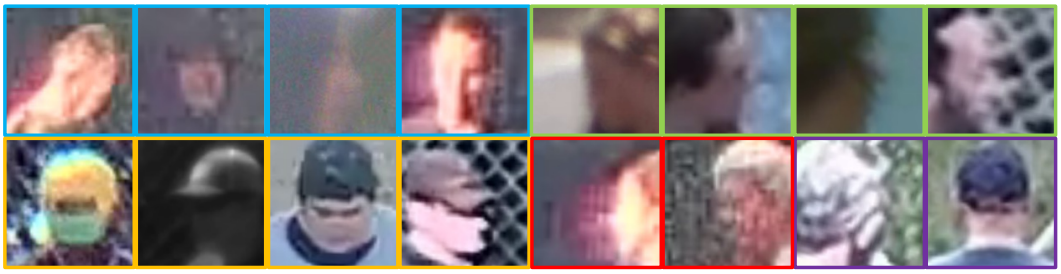

Figure 7: Failure Case Analysis of FaceMoE model on the BRIAR dataset.

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.

To evaluate the routing mechanism under extreme degradation, we visualize the activation maps for each expert in Figure 8. The figure contrasts success and failure cases by illustrating how activation patterns behave under challenging visual conditions. In successful examples, despite blur or moderate pose variations, the experts activate coherently around semantically meaningful facial regions, such as landmarks, contours, and stable low-frequency structures, allowing the network to extract sufficient identity cues. In failure cases, however, extreme pose, over/under-exposure, or severe occlusion disrupt this specialization: activation maps become diffuse, fragmented, or erroneously concentrated in non-informative regions. As shown in the failed inputs, experts often shift their focus to background patches or large smooth areas lacking discriminative detail, indicating that the model can no longer reliably localize or route tokens to the appropriate experts. This divergence between structured and unstable activation patterns highlights the sensitivity of low-resolution recognition models to severe degradations and explains why extreme angles, overexposure, and occlusion frequently lead to identity misclassification and degraded recognition performance.

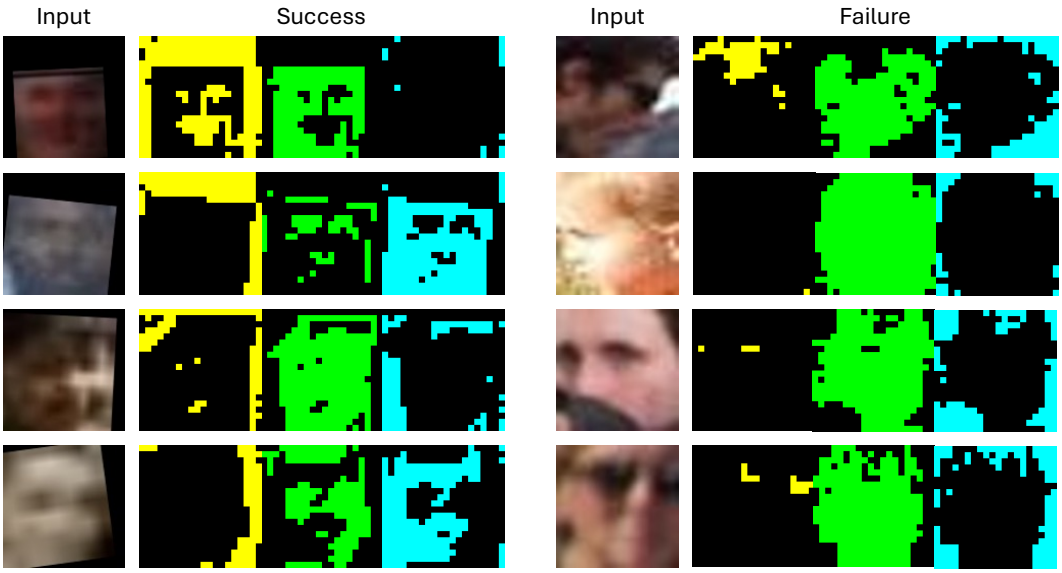

Figure 8: Comparison of activation maps for success and failure cases.

F Limitations and Future Work

Our training data, WebFace4M Zhu et al. (2021), 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.