# MoDE: Mixture of Diffusion Experts for Any Occluded Face Recognition

No Author Given

No Institute Given

In the supplementary material, we present more discussions about our MoDE, including expert ensemble manners, gating network architecture, model robustness, performance of MoDE on the occluded face verification, and more details of ablation studies and real occluded datasets. We compare two expert integration manners: in the feature space and in the decision space, demonstrating that the latter typically results in higher accuracy. Additionally, we conduct experiments on two gating network architectures, Softmax Gating and Noisy Top-K Gating, and find that when the number of experts is limited, Softmax Gating yields superior results. Moreover, we present additional quantitative results for face recognition under various types of occlusions and occluded face verification, further highlighting the robustness and generalizability of MoDE.

## 1 Effect of Experts Integration Manner

The differences between data integration and data fusion are hard to describe. In fact, some researchers acknowledge that these two concepts are almost the same under most circumstances [1]. Data fusion can be categorized into three levels: pixel-level fusion, feature-level fusion, and decision-level fusion [2]. Similarly, in this paper, the integration of experts can be also divided into these three categories. However, due to image noise, resolution difference between images and computational complexity, there are still numerous challenges in pixel-level fusion [3]. Therefore, we did not adopted it in our study. To draw comparisons between the integration at the feature level and decision level (as illustrated in Fig. 1) respectively, we evaluate the respective performance based on five baselines: ArcFace, FaceNet, CosFace, FFR, and Deepface-EMD on three datasets: Occluded CelebA, Occluded MS1M, and Occluded LFW.

The results presented in Table 1 demonstrate that both feature-level integration and decision-level integration outperform the baseline across all datasets. For instance, while feature-level integration fails to improve when combined with DeepFace-EMD (30.1% on Occ CelebA), decision-level integration leads to significant improvement (from 31.6% to 37.4% on Occ CelebA). This suggests that decision-level integration possesses superior error-correcting capabilities, enabling more effective elimination of errors caused by specific experts through adaptive selection. Additionally, decision-level integration embodies a more straightforward and intuitive approach, aligned with cognitive logic.

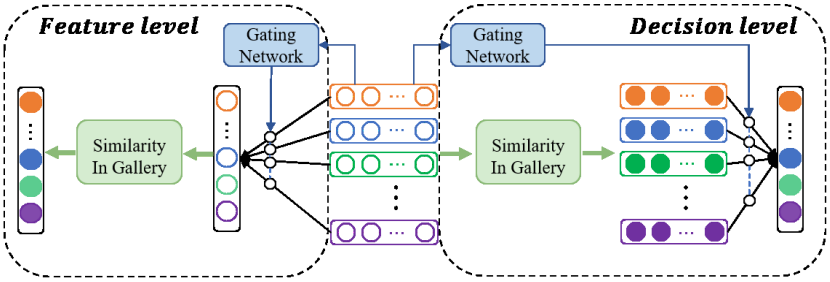


**Fig. 1.** Experts Integration Techniques Comparison.

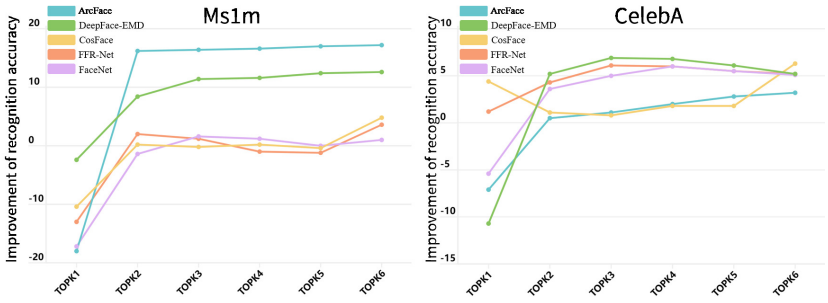


**Fig. 2.** Improvement of recognition accuracy on occluded MS1M and CelebA dataset.

## 2 Analysis of Gating Network

To demonstrate the rationality of the ID-Gate network, we conduct experiments on two advanced architectures: Softmax Gating [4] and Noisy Top-k Gating [5], and compare the corresponding performance. Softmax Gating involves performing a linear transformation on the input, followed by applying the Softmax function.

$$G(x) = Softmax\left(W_g \cdot x + b_g\right). \tag{1}$$

Noisy Top-k Gating is mainly designed for situations with a large number of experts [5]. Before applying the Softmax function, tunable Gaussian noise is added and only the top k weights are retained.

$$G(x) = Softmax(KeepTopK(H(x), k)), \tag{2}$$

$$H(x)_i = (x \cdot W_g)_i + StandardNormal() \cdot Softplus\left((W_{noise} \cdot x)_i\right) \tag{3}$$

$$KeepTopK(v, k)_i = \begin{cases} v_i & \text{if } v_i \text{ is in the top } k \text{ of } v. \\ -\infty & \text{otherwise.} \end{cases} \tag{4}$$

We traversed the hyperparameter k from 1 to 6. When k equals the total number of experts, 6, Softmax Gating is utilized.

In Table 2 and Fig. 2, we present the recognition accuracy and the improvement compared to different baselines. It is evident from Fig. 2 that regardless

**Table 1.** Comparisons between feature-level and decision-level integration. Bold red indicates the best, Bold blue indicates the second best, and Bold cyan indicates the third best.

| Model | Method | Occ CelebA | | Occ MS1M | | Occ LFW | |
|---|---|---|---|---|---|---|---|
| | | Top1 | Top5 | Top1 | Top5 | Top1 | Top5 |
| ArcFace | baseline | 23.2 | 38.4 | 87.8 | 93.6 | 64.6 | 79.3 |
| | feature | 32.1 | 45.6 | 87.8 | 94.4 | 68.6 | 80.3 |
| | decision | 31.9 | 46 | 88.8 | 95 | 71.4 | 82 |
| FaceNet | baseline | 16.4 | 31 | 30.6 | 47.4 | 27.6 | 51.4 |
| | feature | 21.3 | 37.5 | 30.6 | 50.2 | 33 | 55.2 |
| | decision | 21.5 | 37.4 | 31.6 | 49 | 34.2 | 56.7 |
| CosFace | baseline | 3.4 | 9.7 | 42.4 | 59 | 10.4 | 18.4 |
| | feature | 6.4 | 13.8 | 46.2 | 60.8 | 13 | 23.5 |
| | decision | 9.6 | 17 | 47.2 | 63.2 | 14.4 | 24 |
| FFR | baseline | 15 | 26 | 70.6 | 83.4 | 46.7 | 62.6 |
| | feature | 20.4 | 32.4 | 73.4 | 83.8 | 51.3 | 67.5 |
| | decision | 20.2 | 32.5 | 74.2 | 84.2 | 52.5 | 67.6 |
| Deepface-EMD | baseline | 24.6 | 42 | 18.6 | 32.6 | 22.2 | 40.2 |
| | feature | 30.1 | 48.6 | 31 | 49.4 | 35.9 | 55.1 |
| | decision | 31.6 | 50.2 | 31.2 | 48.8 | 35.9 | 55.2 |
| | EMD +feature | 30.1 | 48.4 | 31 | 49.4 | 35.7 | 54.8 |
| | EMD +decision | 37.4 | 54.3 | 38.2 | 54.3 | 47.5 | 63.9 |

of the value of k ranging from 1 to 5, certain methods exhibit performance reduction as compared to baselines. For instance, when combined with ArcFace on Occ MS1M, T-1 decreased by 34.2% compared to the baseline (from 87.8% to 53.6%). Top-k gating presented only a slight advantage on a few specific datasets and methods. Conversely, when Softmax Gating is employed (k=6), all methods present consistent improvement. This indicates that the more complex Noisy Top-k Gating method may not deliver the anticipated improvements for MoDE with only a few experts, which may be attributed to excessive identity information being filtered out by the sparsification process,

## 3 Discussion of occlusion in various situations

We produced two datasets consisting of diverse occlusion scenarios, including fixed occlusions such as Leaves and Slate, random occlusions such as Random loss, and occlusions that vary according to facial landmarks such as Glasses. We tested various methods on these datasets, and the results are presented in Table 5.

Our MoDE demonstrates significant enhancements across all five occlusion conditions, especially in cases of fragmented or random occlusions (Leaves, Line, Random loss), displaying even more astounding improvements than continu-

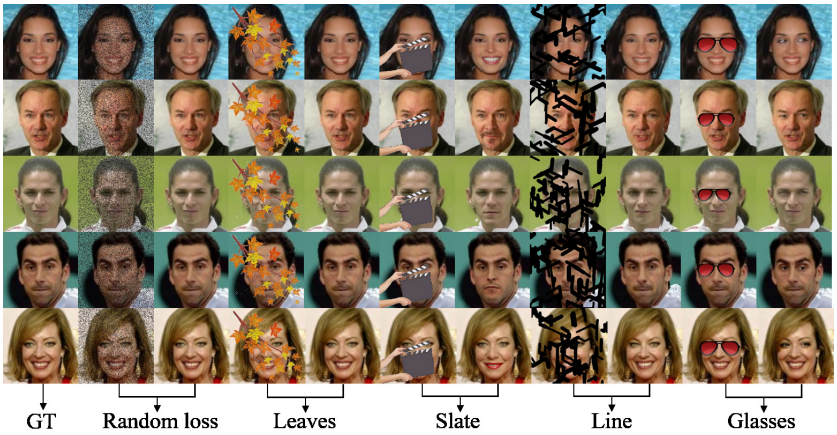


**Fig. 3.** Repainting faces with different occlusions

**Table 2.** Recognition accuracy of Softmax Gating and Noisy Top-k Gating. T-k represents retaining top k weights. Bold red indicates the best, Bold blue indicates the second best, and Bold cyan indicates the third best.

| Dateset | Method | baseline | T-1 | T-2 | T-3 | T-4 | T-5 | T-6 |
|---|---|---|---|---|---|---|---|---|
| Occ MS1M | ArcFace | 87.8 | 53.6 | 87.8 | 88 | 88.2 | 88.6 | 88.8 |
| | CosFace | 42.4 | 32 | 42.6 | 42.2 | 42.6 | 42 | 47.2 |
| | FFR-Net | 70.6 | 57.6 | 72.6 | 71.8 | 69.6 | 69.4 | 74.2 |
| | FaceNet | 30.6 | 13.4 | 29.2 | 32.2 | 31.8 | 30.6 | 31.6 |
| | DeepFace-EMD | 25.6 | 23.2 | 34 | 37 | 37.2 | 38 | 38.2 |
| Occ CelebA | ArcFace | 23.2 | 21.6 | 29.2 | 29.8 | 30.7 | 31.5 | 31.9 |
| | CosFace | 3.4 | 7.7 | 4.4 | 4.1 | 5.1 | 5.1 | 9.6 |
| | FFR-Net | 15 | 16.2 | 19.3 | 21.1 | 21 | 20.5 | 20.2 |
| | FaceNet | 16.4 | 11 | 20 | 21.4 | 22.4 | 21.9 | 21.5 |
| | DeepFace-EMD | 32.2 | 21.5 | 37.4 | 39.1 | 39 | 38.3 | 37.3 |

**Table 3.** Comparison between five state-of-art models and MoDE the in face verification task on five datasets. T@F represents TPR@FPR=0.1. Bold red indicates the best, Bold blue indicates the second best.

| Model | Method | Occ MS1M | | | | Occ CelebA | | | | Occ LFW | | | | OVF | | | | WWCF | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc | T@F | EER | AUC | Acc | T@F | EER | AUC | Acc | T@F | EER | AUC | Acc | T@F | EER | AUC | Acc | T@F | EER | AUC |
| ArcFace | baseline | 97 | 99 | 0.03 | 0.994 | 95.2 | 96.9 | 0.048 | 0.988 | 83.5 | 77.1 | 0.165 | 0.907 | 97.7 | 98.9 | 0.023 | 0.995 | 77.2 | 64.7 | 0.228 | 0.852 |
| | MoDE | 97.3 | 99.2 | 0.027 | 0.994 | 95.5 | 97.2 | 0.045 | 0.99 | 85.4 | 81.8 | 0.146 | 0.925 | 99.2 | 98.9 | 0.0057 | 0.997 | 77.9 | 68.3 | 0.22 | 0.862 |
| FaceNet | baseline | 86.6 | 80.6 | 0.134 | 0.94 | 93 | 95.8 | 0.07 | 0.981 | 83.9 | 78.5 | 0.161 | 0.917 | 83.2 | 78.4 | 0.167 | 0.925 | 86.2 | 64.5 | 0.218 | 0.862 |
| | MoDE | 86 | 80.8 | 0.14 | 0.942 | 93.6 | 96.4 | 0.064 | 0.982 | 85.7 | 82.5 | 0.143 | 0.937 | 84.7 | 93 | 0.155 | 0.929 | 78.6 | 64.1 | 0.214 | 0.871 |
| CosFace | baseline | 86.9 | 83.8 | 0.131 | 0.945 | 79.1 | 67.2 | 0.209 | 0.862 | 70 | 48.4 | 0.3 | 0.768 | 87.8 | 87.5 | 0.121 | 0.961 | 66 | 29.6 | 0.34 | 0.714 |
| | MoDE | 87.8 | 86.2 | 0.122 | 0.949 | 82 | 71.7 | 0.18 | 0.883 | 70.8 | 45.6 | 0.292 | 0.78 | 90.1 | 92 | 0.104 | 0.983 | 67.2 | 33.1 | 0.328 | 0.725 |
| FFR-Net | baseline | 93.2 | 95.4 | 0.068 | 0.983 | 89.8 | 90.5 | 0.102 | 0.965 | 80 | 67.4 | 0.2 | 0.882 | 97.7 | 100 | 0.023 | 0.997 | 75 | 55 | 0.25 | 0.824 |
| | MoDE | 94.1 | 96 | 0.059 | 0.986 | 90.7 | 90.8 | 0.093 | 0.97 | 82 | 70.8 | 0.181 | 0.899 | 99.2 | 98.9 | 0.006 | 0.998 | 77.4 | 62.7 | 0.226 | 0.846 |
| DeepFace-EMD | baseline | 82.8 | 73.6 | 0.172 | 0.904 | 87.2 | 85.4 | 0.129 | 0.945 | 85.6 | 80.6 | 0.144 | 0.924 | 89.3 | 86.4 | 0.109 | 0.959 | 79.6 | 68.4 | 0.204 | 0.877 |
| | MoDE | 86.5 | 80 | 0.135 | 0.932 | 91.9 | 93.3 | 0.08 | 0.975 | 88.1 | 87.5 | 0.118 | 0.947 | 98.5 | 97.7 | 0.011 | 0.992 | 80.6 | 73.4 | 0.193 | 0.89 |
| | EMD | 83 | 76.4 | 0.17 | 0.913 | 88.4 | 87.5 | 0.116 | 0.953 | 86.9 | 83.7 | 0.132 | 0.931 | 93.1 | 98.9 | 0.069 | 0.986 | 80.7 | 67.7 | 0.193 | 0.879 |
| | EMD+MoDE | 87.2 | 84.8 | 0.128 | 0.942 | 93.3 | 95.7 | 0.067 | 0.982 | 88.7 | 87.7 | 0.113 | 0.952 | 99.2 | 98.9 | 0.006 | 0.998 | 81.1 | 73.5 | 0.189 | 0.894 |

**Table 4.** Face identification and verification comparisons of three variants on four datasets

| Method | Occ MS1M | | | | Occ LFW | | | | OVF | | | | WWCF | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Top1 | Top5 | Acc | EER | Top1 | Top5 | Acc | EER | Top1 | Top5 | Acc | EER | Top1 | Top5 | Acc | EER |
| Baseline | 87.8 | 93.6 | 97 | 0.03 | 64.6 | 79.3 | 95.2 | 0.048 | 90.9 | 100 | 97.7 | 0.023 | 22 | 36.7 | 77.2 | 0.228 |
| Baseline + RF | 71.6 | 82.8 | 95.5 | 0.045 | 54.4 | 71.4 | 92.5 | 0.075 | 88.6 | 100 | 95.4 | 0.046 | 13.2 | 23.9 | 72.7 | 0.273 |
| MoDE | 88.8 | 95 | 97.3 | 0.027 | 71.4 | 82 | 95.5 | 0.045 | 93.2 | 100 | 98.5 | 0.011 | 22.5 | 37 | 77.9 | 0.22 |

**Table 5.** The improvement of MoDE for various occluded face recognition. RL represents Random loss.

| Dataset | Method | | Slate | Glasses | Leaves | Line | RL |
|---|---|---|---|---|---|---|---|
| MS1M | ArcFace | Base | 84 | 96 | 64 | 89 | 44 |
| | | MoDE | 86(+2.0) | 96.5(+0.5) | 90(+26.0) | 98(+9.0) | 95(+51.0) |
| | FaceNet | Base | 10 | 36.5 | 13 | 4 | 0 |
| | | MoDE | 18(+8.0) | 47.5(+11.0) | 42(+29.0) | 58(+54.0) | 35(+35.0) |
| | CosFace | Base | 37 | 55.5 | 16 | 18 | 0 |
| | | MoDE | 43(+6.0) | 66(+10.5) | 52(+36.0) | 66(+48.0) | 57(+57.0) |
| | FFR | Base | 60 | 62.5 | 24 | 37 | 22 |
| | | MoDE | 71(+11.0) | 73(+10.5) | 68(+44.0) | 81(+44.0) | 78(+56.0) |
| LFW | ArcFace | Base | 56 | 88 | 54 | 43 | 38 |
| | | MoDE | 67(+11.0) | 95(+7.0) | 93(+39.0) | 79(+36.0) | 90(+52.0) |
| | FaceNet | Base | 11 | 57 | 14 | 16 | 3 |
| | | MoDE | 28(+17.0) | 76(+19.0) | 75(+61.0) | 59(+43.0) | 72(+69.0) |
| | CosFace | Base | 12 | 31 | 5 | 6 | 0 |
| | | MoDE | 25(+13.0) | 42(+11.0) | 42(+37.0) | 33(+27.0) | 38(+38.0) |
| | FFR | Base | 40 | 55 | 10 | 17 | 17 |
| | | MoDE | 54(+14.0) | 64(+9.0) | 63(+53.0) | 53(+36.0) | 64(+47.0) |

ous or large occlusions (Slate, Glasses). We can also observe from Fig. 3 that repainting images with fragmented occlusions (Leaves, Line) exhibits a closer resemblance to the original image. In contrast, the repainting of images with complete occlusion blocks (Slate, Glasses) may result in slight deviations in facial expression or overall structure, making the improvement in recognizing occluded faces less obvious.

## 4 Results of the verification task

The MoDE model was trained exclusively for the face recognition task. To evaluate its performance in the face verification task, we conducted experiments in comparison with five state-of-art models on five datasets. In addition to the metrics mentioned in the main text, including EER and Acc, our evaluation also considers TPR@FPR=0.1 and AUC. Specifically, TPR@FPR=0.1 indicates the True Positive Rate (TPR) when the False Positive Rate (FPR) equals 0.1,

**Fig. 4.** This part visualizes the WWCF-Dataset we collected from the Internet. Each identity consists of one normal face and several different occluded faces.

while AUC represents the area enclosed by the Receiver Operating Characteristic (ROC) curve and the coordinate axis.

The quantitative results were presented in Table 3. Obviously, MoDE outperforms other methods (ArcFace [6], FaceNet [7], CosFace [8], FFR-Net [9] and DeepFace-EMD [10]) across almost all datasets and metrics. Notably, combining MoDE with certain methods (DeepFace-EMD [10]), yields exceptional results with an accuracy of 99.2%, resulting in a significant boost of 9.9% (from 89.3% to 99.2%) on real occluded datasets such as OVF. These findings suggest that MoDE has the potential to be applied in reality. MoDE employs diffusion experts for the retrieval of obscured information and ID-Gate for identity evaluation, thus offering a near-accurate approximation of the original facial image at the pixel and decision level. These results demonstrate the effectiveness of MoDE as a *plug-and-play* approach, scalable and robust across both face recognition and verification.

## 5 Ablation Studies on multiple datasets

The results of the ablation studies on the Occ CelebA dataset have been presented in the main text, and in this section, we present the results of ablation studies on four datasets: Occ MS1M, Occ LFW, OVF, and WWCF. Compared to the two variants, MoDE achieved the best performance across all datasets and metrics, demonstrating a 6.8% increase (from 64.6% to 71.4%) specifically on the Occ LFW dataset. In contrast to the experiments on Occ CelebA, the "Baseline + RF" approach did not yield the expected performance enhancement, showing rather a certain degree of decline, especially on the real occluded dataset WWCF. This could be attributed to the inherent inaccuracy of facial landmark detection on occluded faces, resulting in a negative impact on facial alignment and reconstruction. On the other hand, the use of average fusion in the decision space lacks differentiation and specificity, leading to the amplification of identity noises. Therefore, the diverse reconstruction experts and dynamically selected ID-Gate based on identity credibility are indispensable, jointly using them produces maximal improvement. This further demonstrates the effectiveness of our MoDE framework.

## 6 Examples of OVF and WWCF dataset

In this work, we collected two real occluded datasets, namely Occlded Volunteer Face (OVF) and Web Wlid Occluded Celebrity Face (WWCF), as shown in Fig. 4. The OVF consists of 352 facial images from 44 identities. Each identity contains three normal faces, two regular masked faces, and three deep blue masked faces. The WWCF consists of 873 facial images from 233 identities. Each identity contains one normal face and up to 31 occluded faces. All the images have been preprocessed for face alignment to maintain consistency. These images demonstrate multiple occlusion possibilities, with different shapes, sizes, and colors, providing a realistic depiction of the complexities and variations present in real-world scenarios.

## References

1. J. Zhou, X. Hong, and P. Jin, "Information fusion for multi-source material data: Progress and challenges," *Applied Sciences*, vol. 9, no. 17, p. 3473, 2019.
2. P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, pp. 345–379, 2010.
3. S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *information Fusion*, vol. 33, pp. 100–112, 2017.
4. M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the em algorithm," *Neural computation*, vol. 6, no. 2, pp. 181–214, 1994.
5. N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *arXiv preprint arXiv:1701.06538*, vol. n.d., no. n.d., p. n.d., 2017.
6. J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. n.d.: n.d., 2019, pp. 4690–4699.
7. F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. n.d.: n.d., 2015, pp. 815–823.
8. H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. n.d.: n.d., 2018, pp. 5265–5274.
9. S. Hao, C. Chen, Z. Chen, and K.-Y. K. Wong, "A unified framework for masked and mask-free face recognition via feature rectification," in *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE. n.d.: n.d., 2022, pp. 726–730.
10. H. Phan and A. Nguyen, "Deepface-emd: Re-ranking using patch-wise earth mover's distance improves out-of-distribution face identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. n.d.: n.d., 2022, pp. 20 259–20 269.