OFER: Occluded Face Expression Reconstruction

Supplementary Material

A. Preliminaries

Denoising Diffusion Probabilistic Model (DDPM). The underlying mechanism of DDPM is the transformation of an unknown data distribution, $p(X) \in \mathcal{R}^d$ into a simple known distribution. Here, X represents the data from the underlying distribution. This transformation is achieved by iteratively applying a transition kernel q via a Markov chain process with infinitesimal time steps, ensuring a stationary distribution at each time step t i.e $x_t \sim q(x_t|x_{t-1}), \forall t > 0$ where $x \in X$. DDPM models these transformations by parameterizing a neural network to capture complex dependencies in the data by modeling the sequential evolution of the data distribution. In the work of Sohl *et al.* [51], the known distribution is set as Gaussian with a decaying variance schedule $\beta_t \in \mathcal{R}$ such that $q(x_t|x_{t-1} = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I); q(x_T) = \mathcal{N}(x_T; 0, I)$ due to its simplicity and traceability. Denoising in DDPM involves reversing this process by starting with unit Gaussian noise to model the underlying data distribution $p'(X)$, which models the training process.

Learning-based Ranking. Ranking is an important technique used in the information retrieval domain to retrieve relevant documents given a query. There are three primary learning-based ranking techniques [35]: pointwise, pairwise, and listwise. The pointwise model takes in a single query(q)-document(d_i) pair and gives the relevance score (s_i) of each in isolation, which requires the ground-truth score: $\bar{r}(q, d_i) = s_i$. Pairwise ranking methods compare two documents against each other based on importance and relevance to the query: $\bar{r}(q, d_i, d_j) = P(d_i \triangleright d_j)$. Finally, listwise ranking [43] generates an optimal order of a list of documents by calculating the importance score of each: $\bar{r}(q, d_i, \ldots, d_n) = (r_1, \ldots, r_n)$, which is the method we adopt in our network.

B. Ablation Study

We conducted several ablation studies on the choice of conditioning signal, embedding, and loss functions to optimize our model networks, which are detailed in the following sections.

B.1. Identity Generative Network (IdGen)

Choice of image embedding. The representation used for the conditioning image provided to the diffusion network plays an important role in the reconstruction task. In Tab. 6 we show ablation results for the embedding used in IdGen. The results indicate that using only the ArcFace [9] embedding improves the error metric of this network, compared to employing FaRL [59] or a combination of both as conditions. This may be because ArcFace [9] is used to distinguish identity-defining features, while FaRL [59] is designed for downstream facial analysis tasks that require capturing facial subtleties. In addition, we compare the performance of ArcFace with state-of-the-art image embeddings, DINOv2 [37] and CLIP [42]. The results are presented in Tab. 7. The reason we hypothesize for the lower performance of CLIP and DINOv2 is that they are trained on multi-domain models and not specifically trained on faces like ArcFace.

	Occluded subset			Unoccluded subset		
Embedding	Med ↓	Mean ↓	std ↓	Med ↓	Mean ↓	std ↓
ArcFace [9]	**1.03**	**1.26**	**1.06**	**1.01**	**1.22**	**1.01**
ArcFace [9] + FaRL [59]	1.05	1.30	1.09	1.03	1.28	1.06
FaRL [59]	1.15	1.40	1.12	1.07	1.30	1.06

Table 6. **Ablation on image embeddings for the IdGen network.** MSE error on the NoW validation dataset, using ArcFace [9] embedding, FaRL [59] embedding, or a combination of both.

B.2. Identity Ranking Network (IdRank)

Ranking selects low-error samples. We showcase the ability of the ranking network to identify subtle differences to select top-ranked samples with lower MSE error than the lower-ranked samples in Fig. 10.

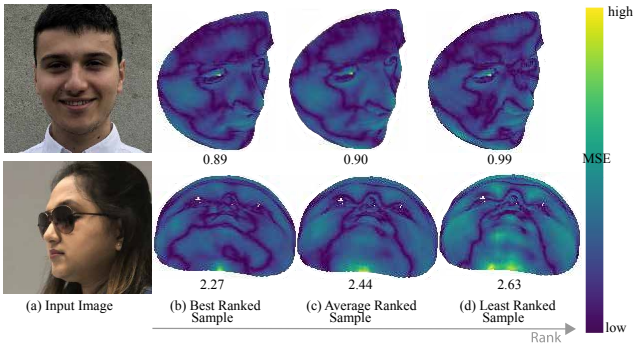

Figure 10. **Ranking on NoW validation images.** For the image in column (a), we generate 100 shape coefficients with IdGen and rank them using IdRank. Column (b) shows the top-ranked sample, (c) a mid-ranked sample (50th), and (d) the worst-ranked sample. Columns (b)-(d) show viewpoints highlighting the sample errors, with its median MSE shown underneath each sample.

Sampling Quantity for Input. Naturally, increasing the number of generated samples results in a more thorough coverage of the data distribution. As the sample set size increases and leads to denser mode coverage, we expect the

Method	ArcFace	CLIP	DINOv2
Median	1.01	1.65	1.68
Mean	1.24	2.07	2.08

Table 7. **Ablation study for IdGen embedding.** on NoW validation benchmark. ArcFace trained for identity clustering performs better in comparison to other image embedding for identity recognition reconstruction tasks.

#samples	Ideal lowest error sample		
	Med ↓	Mean ↓	std ↓
FLAME [33] (Baseline)	1.02	1.30	1.11
10	0.89	1.13	0.97
50	0.83	1.07	0.94
100	0.81	1.05	0.92
500	0.78	1.02	0.90

Table 8. **Ablation on the effect of sampling quantity as input to IdRank.** To evaluate the tradeoff between the number of generated samples by IdGen and the "ideal" lowest error sample in the set. The values are the lowest median error sample evaluated against the NoW validation benchmark.

average error to get closer to the ground truth. However, for the ranking network to identify that optimal element, it must be trained on a sufficiently large number of samples to effectively rank them and select the best one. This requirement comes with a significant computational cost, especially due to the slower inference time of diffusion models. Thus, finding the right balance in the number of samples used for training is important. We conducted an ablation study on different sample sizes generated by IdGen and assessed the median error of the optimal sample from each set to determine the ideal number of samples for training IdRank. The results are presented in Tab. 8.

Face vertices as Input. Selecting the most accurate sample requires a network design that excludes irrelevant information (See Sec. 3.3 for more details). To validate this claim, we conducted an ablation study comparing three inputs: (a) shape coefficients, which are a PCA model representing the entire head shape; (b) all landmark vertices of the reconstructed head shape; and (c) front-face vertices from the reconstructed head shape. The results of this comparison are presented in Tab. 9. The high errors associated with the coefficients and vertex-based reconstructions likely stem from the inclusion of irrelevant information.

Loss function. The challenge in selecting the best representative shape from the samples generated by the IdGen is that, even with mild occlusions where most of the facial structure is preserved, variability can still occur. This means we need an appropriate loss function that mitigates the selection of sub-optimal samples. We considered two loss functions for this network: binary cross-entropy and soft-

Input to rank network	precision ↑			IoU ↑			error (GT/Pred) (ideal = 1)		
	1 %	10 %	20 %	10 %	20 %	small30 %	avg1	avg5	avg10
(a) $\hat{S}$	3.3	21.3	33.2	26.7	40.0	53.3	0.54	0.57	0.64
(b) M^{frontal}	3.3	11.3	24.8	13.3	30.0	53.3	0.49	0.58	0.61
(c) X (μ_M, M')	3.3	**44.5**	**64.7**	**33.3**	**80.0**	**93.3**	**0.68**	**0.78**	**0.83**

Table 9. **Ablation for input to ranking network.** $\hat{S}$ is the 300-dimensional shape coefficients generated from IdGen; M^{frontal} is the front face vertices of FLAME mesh M reconstructed from ($\hat{S}$; X=(μ_M, M') is the mean, residual pair defined in Sec. 3.3

Loss	precision ↑			IoU ↑			error (GT/Pred)↓		
	1 %	10 %	20 %	10 %	20 %	30 %	avg1	avg5	avg10
BCE loss	0.0	3.0	20.8	0.0	0.0	16.7	0.56/0.86	0.60/0.88	0.63/0.85
Softmax loss	0.0	**30.0**	**51.7**	**50.0**	**66.7**	**66.7**	**0.58/0.76**	**0.62/0.77**	**0.65/0.79**

Table 10. **Ablation for loss function for IdRank.** We trained the network with Binary Cross Entropy (BCE) loss and Cross Entropy on Softmax (Softmax) loss using 100 Stirling(HQ) [22] frontal face images. For validation, we used 20 Florence [2] dataset frontal face images. IoU represents the Intersection over Union of predicted ranking order and ground truth ranking order for the first 10, 20, and 30 sorted rank index. error(GT/Pred) shows the average error for the first 1, 5, and 10 ground truth rank samples and that of predicted rank samples.

max loss, which approximates the ground truth error distribution considering all samples. To assess the performance of both, we conducted an ablation study, the results of which are presented in Tab. 10. Since more than one sample can be optimal, ranking and selection using softmax loss yields better precision with a set of higher-ranked samples.

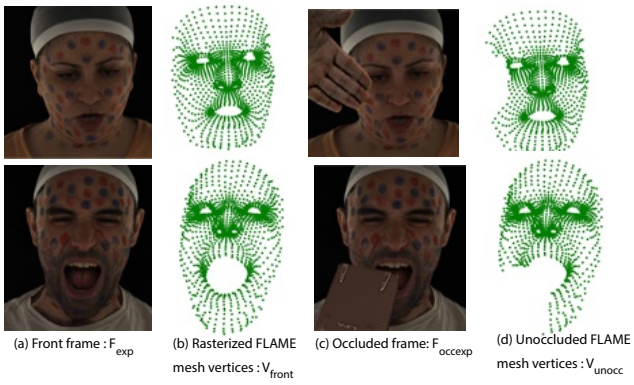

Figure 11. **CO-545.** Column (a) the expressive frames with a frontal view, F_{exp}; Column (b) the rasterized mesh vertices V_{front}; Column (c) the occluded frames with synthetic objects,F_{occexp}; Column (d) the unoccluded vertices which belong to unoccluded pixels, V_{unocc}. The pairs (F_{occexp}, V_{unocc}) make CO-545 dataset.

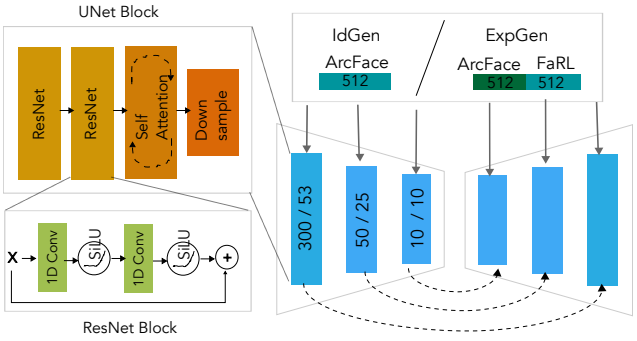

Figure 12. **1D U-Net Transformer Hybrid architecture.** of IdGen and ExpGen. Each block of the U-Net is comprised of two ResNet blocks, followed by a self-attention module and a downsampling/upsampling module for the encoder and decoder respectively. Each ResNet block consists of two 1D ConvNet followed by SiLU activation and residual connection. For IdGen, the conditional embedding ($\mathbb{R}^{512}$) is passed to every layer of the U-Net. The training input to IdGen is FLAME [33] shape coefficient, $S \in \mathbb{R}^{300}$ which gets downsampled to 50 and 10 at the bottleneck. For ExpGen, the conditional embedding is ($\mathbb{R}^{1024}$) and the input is FLAME expression (including 3 jaw coefficients) $E \in \mathbb{R}^{53}$ which is downsampled to 25 and 10 at the bottleneck.

C. Design Choices

C.1. Ranking by distribution matching

The objective of the ranking model is to sort the output from the network and optimize a loss based on ranking order. We choose a list-wise ranking method due to its demonstrated effectiveness [5]. Since sorting and ranking are non-differentiable operations, we reformulate the sort-rank problem into a probability distribution alignment problem aiming to minimize the softmax loss L_R between the ground truth distribution g and the predicted distribution h.

C.2. Exclusion of Ranking in ExpGen

Our ranking network is trained to rank only *neutral* shapes and not expressions. This is because shape geometry remains consistent despite occlusions or variations in expressions in the input image. However, ranking the expression hypotheses is harder since multiple hypotheses can be equally valid for occluded regions.

D. Architecture

The detailed overview of the 1D U-Net-transformer hybrid architecture of our IdGen and ExpGen networks is shown in Fig. 12. Both share similar architecture, differing in the embeddings and the inputs. The conditional embedding of IdGen obtained from ArcFace [9] is $\mathbb{R}^{512}$, and the embedding of ExpGen obtained by concatenating ArcFace [9] and FaRL [59] embeddings is $\mathbb{R}^{1024}$.

Dataset	Num Subjects	Num Images	with exp
Stirling [22]	133	1322	✗
Florence [2]	53	1239	✗
FaceWarehouse [6]	150	3000	✗
LYHM [7]	1211	7118	✗
FaMoS [4]	95	1.5M	✓

Table 11. **Datasets used for training Networks.** FaMoS expression coefficients are used to train ExpGen. The shape coefficients from the remaining four datasets were used to train IdGen and IdRank

E. Datasets

E.1. Training Dataset

In Tab. 11, we list the datasets used to train the networks in our framework. FaMoS [4] dataset comprises 3D registered FLAME meshes. We utilized only a subset of them such that it covered all the expression variations in the entire dataset. Please note that our network trains in parametric space rather than on meshes. Therefore, we obtained the corresponding parameters for this subset directly from the authors of FaMoS. We used the expression coefficients from this subset to train the ExpGen. The remaining four datasets listed in the table (Stirling [22], Florence [2], FaceWarehouse [6], LYHM [7]) were used to train IdGen and IdRank.

E.2. CO-545 Evaluation Dataset

We introduce a new dataset named CO-545 to quantitatively evaluate occluded expressions. First, we select the middle frame of each sequence in the CoMA dataset, F_{exp}, which exhibits frontal views with expressive features while excluding neutral expressions. Subsequently, we rasterize the FLAME mesh for each frame to eliminate naturally occluded vertices from the camera's perspective, selecting only facial vertices and excluding those from the back of the head, eyeballs, neck, and ears. This subset of vertices is denoted as V_{front}. Occlusion masks [54] are then applied to the selected frames F_{occexp}, removing additional vertices from V_{front} that fall within the masked pixel areas of the image. We thus obtain the set of unoccluded vertices, denoted as V_{unocc}, for each masked image. The (F_{occexp}, V_{unocc}) pairs form the dataset, allowing us to evaluate occluded samples only within the visible regions. This procedure enables the inclusion of additional evaluation data in the dataset. A few samples from this dataset are shown in Fig. 11.

Method	Median ↓ (mm)	Mean ↓ (mm)	Std ↓ (mm)
TokenFace [58]	**0.97**	**1.24**	**1.07**
MICA (8DS) [61]	1.08	1.37	1.17
3DDFA V2	1.53	2.06	1.95
DECA [21]	1.35	1.80	1.64
Dib *et al*. [14]	1.59	2.12	1.93
Dense landmarks [55]	1.36	1.73	1.47
FOCUS-MP [32]	1.41	1.85	1.70
Deng *et al*. [11]	1.62	2.21	2.08
RingNet [49]	1.50	1.98	1.77
OFER (4DS) (sample selected by ranking)	1.27	1.64	1.29

Table 12. **Neutral face 3D Metrical reconstruction error on the NoW test benchmark.** The results show a comparison of the accuracy of single-view reconstruction methods based on the NoW challenge.

F. Additional Results

F.1. Quantitative Results

Neutral face reconstruction. We provide a comprehensive comparison of neutral face reconstruction, including additional methods specifically focused on this task, alongside the occlusion-based reconstruction methods discussed in the main paper. Tab. 2 presents the evaluation results for non-metrical reconstruction error on the NoW [49] validation benchmark, while Tab. 12 presents metrical reconstruction error on the NoW test benchmark. Our method does not outperform TokenFace [58], which is explicitly trained with both 2D and 3D supervision–a limitation acknowledged and addressed as future work in the main paper. However, in the case of MICA, which when trained only on the four datasets, the results in Tab. 2 show that its performance is similar to our method. In addition, OFER demonstrates improved results when ranking is incorporated.

F.2. Qualitative Results

We show additional expression variations from the final reconstruction of our method in Fig. 13 and Fig. 14. For identity reconstruction, Fig. 15 presents more results from the ranking of samples evaluated on the NoW validation dataset. While most of the reconstructions appear visually similar, the variations are subtle (see the forehead patterns of rows (b) and (c)). In row (c), the chin area of the least ranked sample shows a high error compared to rank-1 and rank-5 samples. Since these subtle differences are hard to differentiate visually, ranking provides a way to automatically select high-quality samples without manual intervention.

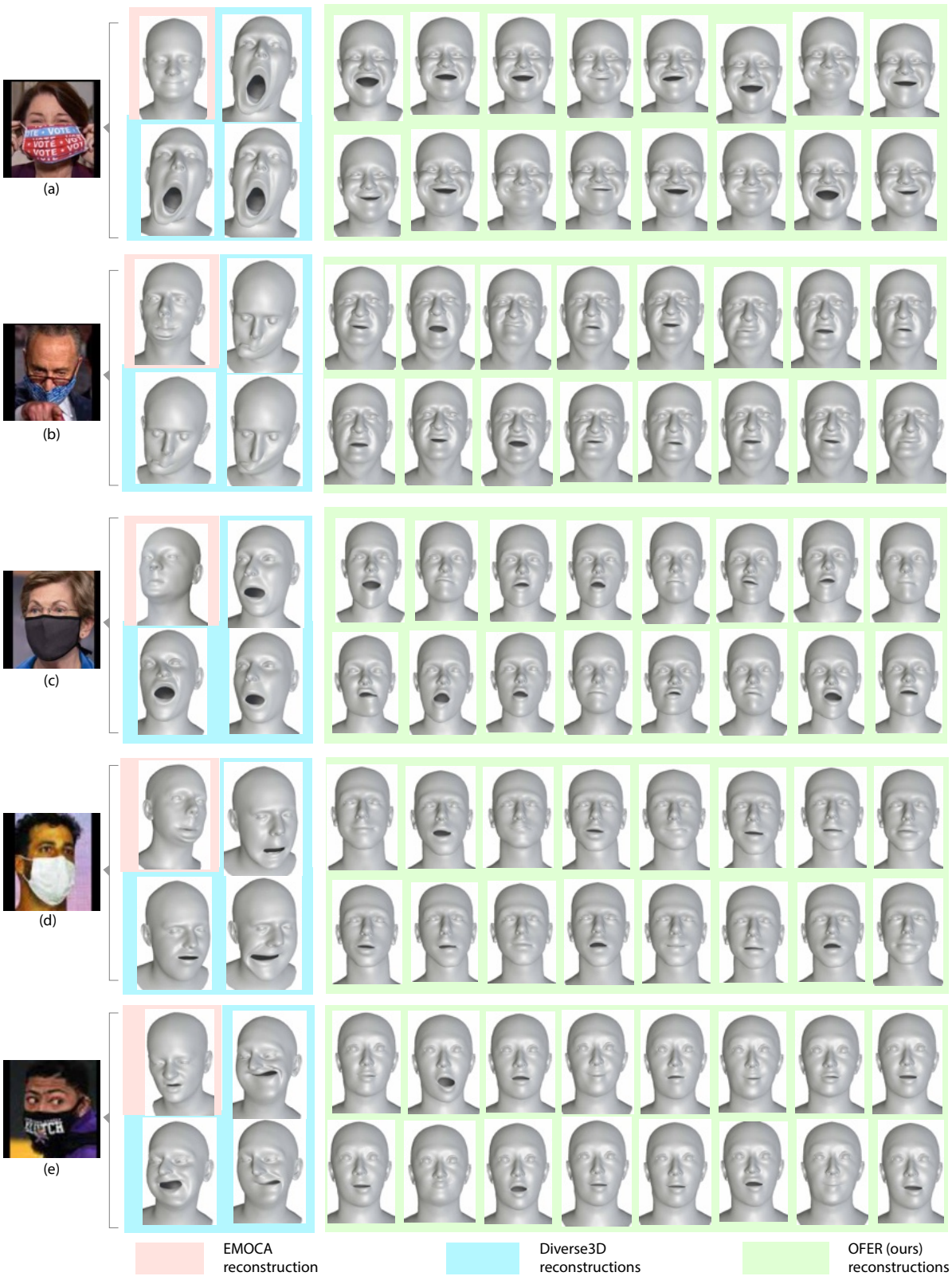

Figure 13. **Comparison of expression sampling on hard occlusions.** We compare against EMOCA [8] (pink), three samplings from Diverse3D [12] (blue) and 16 samples from our method (green).

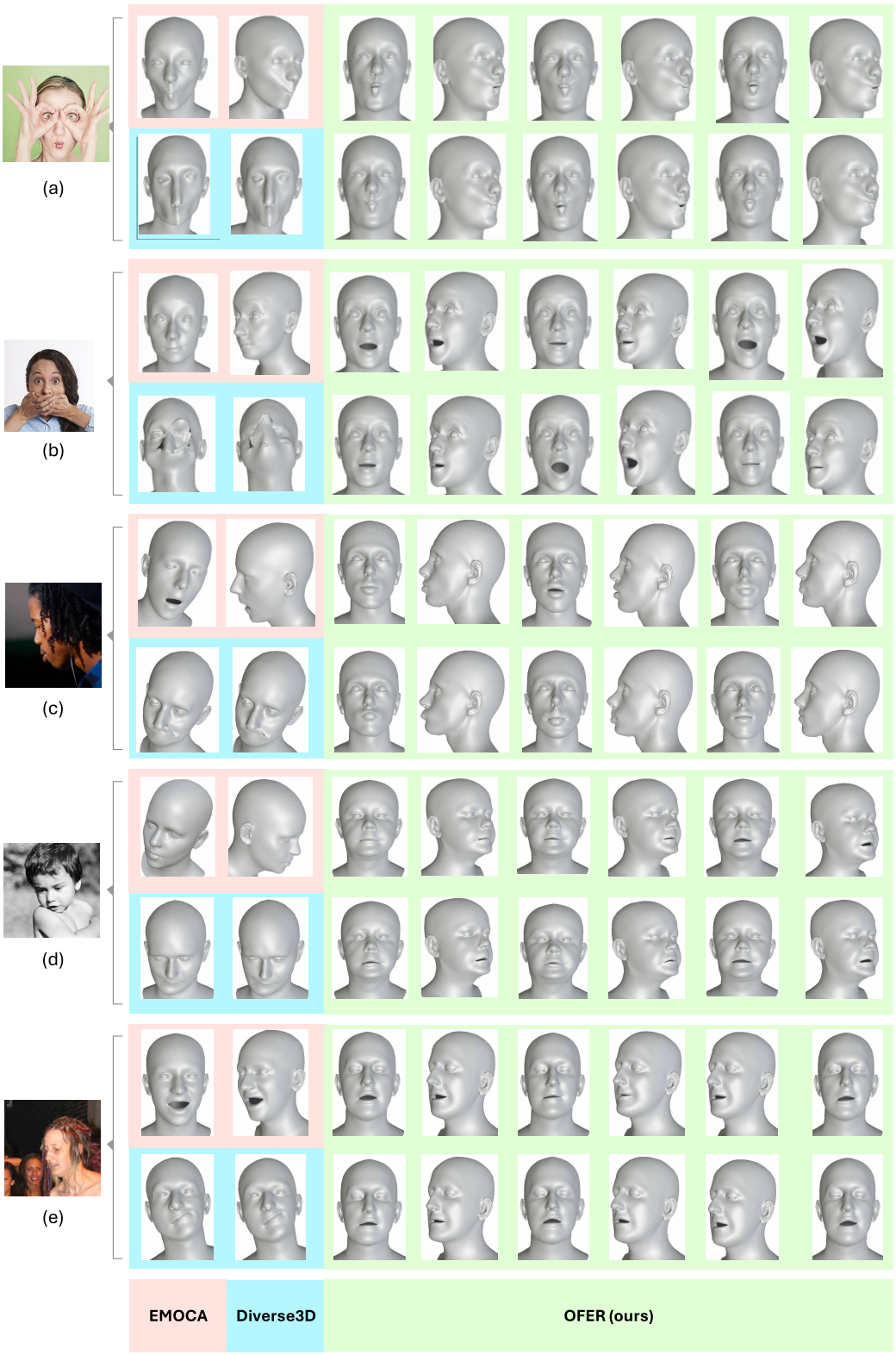

Figure 14. **Comparison of expression reconstruction for in-the-wild occluded images.** We compare against EMOCA [8] (showing front and side view, pink), two reconstructions from Diverse3D [12] (blue), and six samples (front and side view) from our method (green).

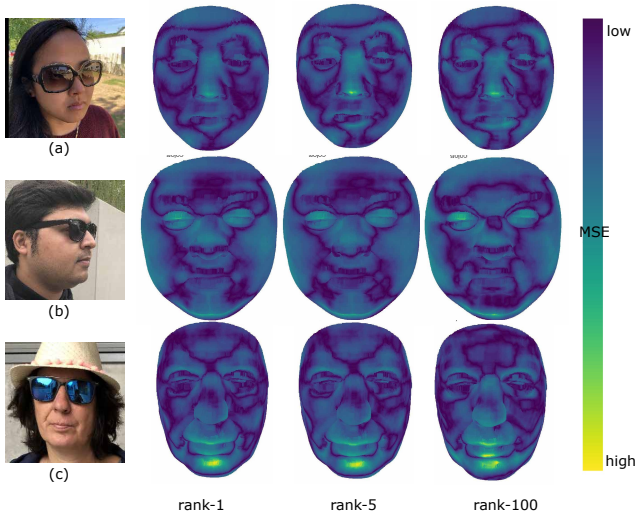

Figure 15. **Ranking on NoW validation.** Column 1 shows the optimal sample selected by IdRank, column 2 displays the sample ranked 5th, and the last column shows the lowest-ranked sample. Although error differences are subtle, variations can be observed between the higher-ranked samples (rank 1 and 5) and the lower ranked-sample (rank 100) in the nose and lip regions of image (a), the eye region of image (b), and the lip and chin region of image(c). This demonstrated the effectiveness of ranking in selecting higher-quality samples.