OFER: Occluded Face Expression Reconstruction

Pratheba Selvaraju[1] Victoria Fernández Abrevaya[2] Timo Bolkart[3] Rick Akkerman[2]
Tianyu Ding[4] Faezeh Amjadi[4] Ilya Zharkov[4]

[1]University of Massachusetts, Amherst [2]Max Planck Institute for Intelligent Systems, Tübingen
[3]Google Research, Zürich [4]Microsoft Research, Redmond

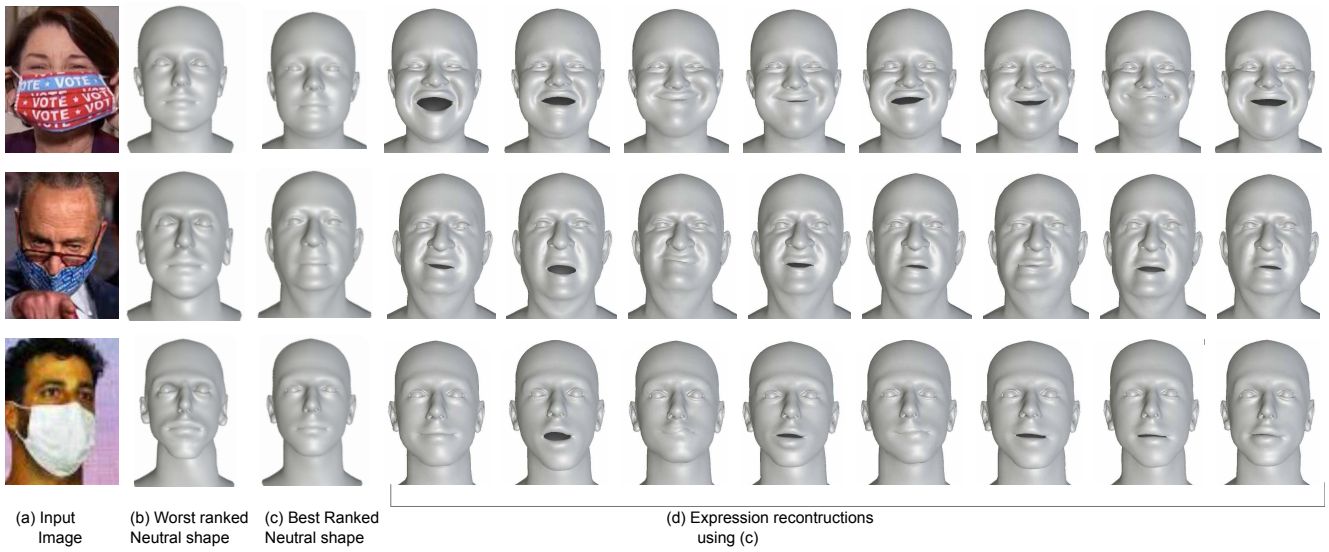

Figure 1. **Reconstructions generated by OFER.** Our method can reconstruct faces from a single image under hard occlusions (a), providing multiple solutions with diverse expressions that align with the input image (d). We use two diffusion models that denoise the shape and expression parameters of FLAME conditioned on the image. A novel ranking mechanism selects an optimal identity (c) from the generated set of shapes, on top of which the expression variants are applied to obtain the final results.

Abstract

Reconstructing 3D face models from a single image is an inherently ill-posed problem, which becomes even more challenging in the presence of occlusions. In addition to fewer available observations, occlusions introduce an extra source of ambiguity where multiple reconstructions can be equally valid. Despite the ubiquity of the problem, very few methods address its multi-hypothesis nature. In this paper we introduce OFER, a novel approach for single-image 3D face reconstruction that can generate plausible, diverse, and expressive 3D faces, even under strong occlusions. Specifically, we train two diffusion models to generate a shape and expression coefficients of face parametric model, conditioned on the input image. This approach captures the multi-modal nature of the problem, generating a distribution of solutions as output. However, to maintain consistency across diverse expressions, the challenge is to select the best matching shape. To achieve this, we propose a novel ranking mechanism that sorts the outputs of the shape diffusion network based on predicted shape accuracy scores. We evaluate our method using standard benchmarks and introduce CO-545, a new protocol and dataset designed to assess the accuracy of expressive faces under occlusion. Our results show improved performance over occlusion-based methods, while also enabling the generation of diverse expressions for a given image.

[1]Project website: https://ofer.is.tue.mpg.de/

1. Introduction

3D face reconstruction from a single image is crucial for creating life-like digital avatars and is used in numerous applications such as illumination-invariant recognition [62], medical imaging [38], and telepresence [31]. The task is an inherently challenging inverse problem, with difficulties posed by depth ambiguity, variations in lighting, diverse facial expressions, and pose [13].

The problem becomes even more challenging when images are subjected to *occlusions*, as shown in Fig. 1. The main difficulty arises from the face being only partially visible, introducing an additional source of ambiguity: the occluded areas can now correspond to an infinite number of valid face shapes, making it a *multi-hypothesis reconstruction* problem. Occlusions are very common in images captured in uncontrolled environments due to factors such as hair, accessories, medical masks, or even strong profile poses and head rotations (*e.g*. Fig. 9). Despite its prevalence, reconstruction under such conditions has seldom been addressed.

3D face reconstruction typically involves recovering the parameters of a statistical 3D facial model, such as the 3D Morphable Model (3DMM) [3], either via model fitting [1, 3, 23] or regression [8, 11, 21, 53]. While these works may handle milder occlusions depending on their training data, their performance deteriorates significantly under more severe obstructions (see *e.g*. Fig. 2). A few methods have been specifically proposed to address the challenge of occluded faces [12, 16, 17, 32]. Most offer a unique solution due to their deterministic nature, ignoring the multi-hypothesis aspect of the problem [16, 17, 32]. An exception to this is Diverse3D [12], which employs Determinantal Point Process (DPP) [29] to sample a diverse set of outputs. While DPP is designed to capture the diversity of the data, it fails to adequately represent the underlying distribution of facial geometry, often resulting in unrealistic and exaggerated reconstructions as shown in Fig. 2.

To ensure plausible face reconstructions while addressing the challenges of diverse sampling, we introduce OFER, a novel method for reconstructing 3D faces under occlusions. At its core, OFER employs two denoising diffusion probabilistic models (DDPMs) [26] to generate the shape and expression coefficients of the FLAME [33] parametric face model, conditioned on an input image. The ability of diffusion models to learn the underlying data distribution enables OFER to produce multiple plausible hypotheses. We leverage two pre-trained face recognition networks as image encoders, and following MICA [61], we train our models using only a small dataset of paired 2D-3D data.

Generating diverse solutions is crucial for reconstructing faces under occlusions. Equally important is selecting a unique, consistent shape that best represents the face in the image across all generated expressions. This is feasible since identity is generally well defined even under occlusions. To achieve this, we propose a novel **ranking mechanism** that evaluates and ranks the samples generated by the shape diffusion network. Specifically, given an input image and N generated shape samples, the ranking network scores and ranks each of the candidates, selecting the top-ranked one. To the best of our knowledge, such a ranking mechanism as a selection process for diffusion models has not been considered before.

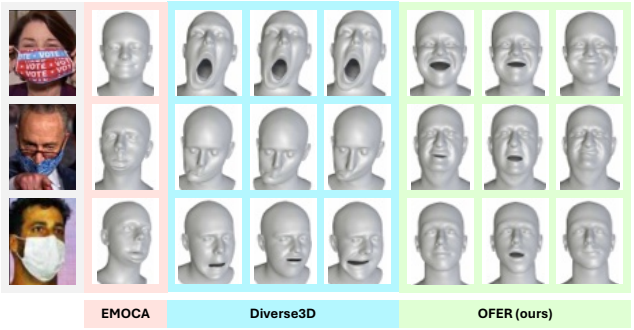

Figure 2. **Comparison against expression reconstruction methods.** We show results from EMOCA [8] (pink); samples generated by Diverse3D [12] (blue); and samples generated by our method (green). EMOCA can only reconstruct a single solution due to its deterministic nature, while Diverse3D shows non-plausible faces. OFER (our method) generates diverse 3D faces with plausible expressions.

To evaluate our approach, we introduce a new dataset and evaluation protocol, CO-545, derived from the CoMA dataset [44]. CO-545 includes 545 ground-truth pairs of occluded images and corresponding FLAME non-occluded 3D vertices. Experimental results show that our method effectively generates multiple hypotheses for single in-the-wild occluded images, achieving superior quality and diversity on CO-545. Additionally, when evaluated on the NoW benchmark [49], our method demonstrates improved performance with the ranking-selection mechanism.

In summary, our contributions are the following:

- A new method for occluded 3D face reconstruction using DDPMs, which outputs multiple 3D face shape and expression hypotheses, overcoming the limitations of deterministic methods.
- A novel ranking mechanism that scores and selects the optimal 3D face shape from reconstruction candidates generated by the shape network.
- A new validation dataset, CO-545, for the quantitative evaluation of occluded face reconstruction, addressing the lack of existing evaluation protocols.

2. Related Work

3D face reconstruction from a single image. 3D face reconstruction from a single image has been a key research

area for several decades [18], with approaches broadly categorized into model-free [10, 15, 20, 24, 27, 47, 50, 57] and model-based methods [8, 21, 45, 58, 61]. Model-free approaches estimate 3D geometry directly from images, while model-based ones recover the low-dimensional parameters of a statistical model of the 3D face such as BFM [40] or FLAME [33]. To overcome the lack of large-scale datasets with paired images and 3D models, recent trends have shifted towards self-supervised learning [8, 11, 21, 45, 52] using landmark re-projection and/or photometric error. These works can yield suboptimal results when landmarks are missing or when color information is compromised due to occlusions. Alternatively, MICA [61] uses a small dataset of paired 2D-3D data to map an image embedding from a face recognition network [9] to the FLAME shape parameters. This technique achieves state-of-the-art results for neutral face shape reconstruction but does not support expressive faces.

3D face reconstruction from occluded images. A few works have specifically addressed the problem of reconstructing faces under occlusion. Egger *et al*. [16, 17] proposed a probabilistic optimization approach that simultaneously solves for model parameters and segmentation regions using an expectation–maximization approach. FOCUS [32] follows a similar idea within a self-supervised learning strategy. These methods prioritize obtaining valid reconstructions for the non-occluded regions but do not address the ambiguity of the problem, which requires a distribution of solutions. Recently, Diverse3D [12] tackled this by employing a mesh-based variational auto-encoder (VAE) for shape completion, combined with Determinantal Point Process (DPP) [29] for sampling diverse solutions in expression space. However, this method often results in unrealistic and exaggerated reconstructions (see Fig. 2), since DPP is not designed to capture the statistical properties of the data.

Learning to Rank. Ranking plays an essential role in Information Retrieval [35], commonly used for sorting documents by relevance [25, 28], image search [39], and recommendation systems [36]. Learning to Rank methods [35] perform this task by using supervised machine learning approach. RankGAN [34] uses ranking as a rewarding mechanism to train a language generator to produce higher-quality descriptions. Some recent works such as T5 [60] focus on text ranking with large language models, and incorporate ranking losses as a fine-tuning tool to optimize model performance. To the best of our knowledge, ranking has not been explored in the context of diffusion models to select an optimal sample.

3. Method

Our method takes as input a single-view image of an occluded face and generates a *set* of 3D faces as output. The goal is to produce reconstructions that explore a *diverse range of expressions* in the occluded areas, while accurately capturing the visible regions of the input image.

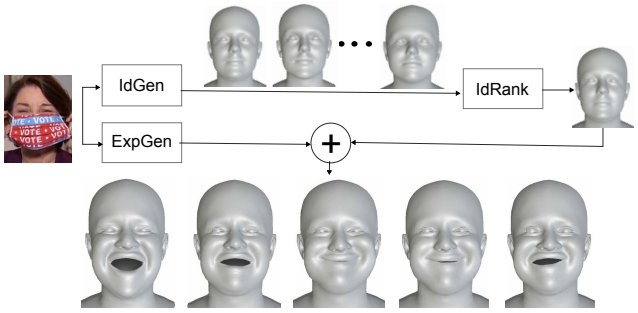

Figure 3. **Overview of OFER.** Given an input image, the Identity Generative Network (IdGen) samples N shape parameters. The reconstructed shapes are then passed to the Identity Ranking Network (IdRank) to select a unique identity. Finally, the Expression Generative Network (ExpGen) generates N expression parameters, which are combined with the selected shape to output *diverse and expressive face reconstructions* (bottom row).

OFER is structured around three key components. First, the *Identity Generative Network* (IdGen, Fig. 4 and Sec. 3.2), a DDPM conditional diffusion model, outputs a set of FLAME [33] shape coefficients capturing a distribution of plausible neutral 3D faces. Next, the *Identity Ranking Network* (IdRank, Fig. 5 and Sec. 3.3), a small MLP, evaluates and ranks the generated shape samples from IdGen, selecting the one that best matches the image. Finally, the *Expression Generative Network* (ExpGen, Fig. 4 and Sec. 3.4), another conditional DDPM network, generates a diverse set of FLAME expression coefficients. The three networks are conditioned on the same input image. By combining the selected shape coefficient with the set of expression hypotheses, we obtain the final set of 3D reconstructions. The overall architecture of OFER is shown in Fig. 3. We detail each of its components in the following sections.

3.1. Preliminaries: FLAME model

FLAME [33] is a parametric 3D face model combining separate shape and expression spaces to form a 3D head mesh M with $n = 5023$ vertices. Given the shape $S \in \mathbb{R}^{|S|}$, expression $E \in \mathbb{R}^{|E|}$, and pose $P \in \mathbb{R}^{|P|}$ parameters, FLAME produces a mesh M as

$$M(S,E,P) = \text{LBS}\big(\mathcal{T}(S,E,P), \mathcal{J}(S), P, \mathcal{W}\big), \quad (1)$$

where LBS is the linear blend skinning function, weights $\mathcal{W} \in \mathbb{R}^{4\times n}$ and joint regressor $\mathcal{J}(S) \in \mathbb{R}^{3K}$. $T(S,E,P) = \mathcal{T} + \mathcal{B}_S(S) + \mathcal{B}_E(E) + \mathcal{B}_P(P)$ deforms a template mesh $\mathcal{T}$ using the shape $\mathcal{B}_S$, expression $\mathcal{B}_E$ and pose-corrective $\mathcal{B}_P$ blendshapes.

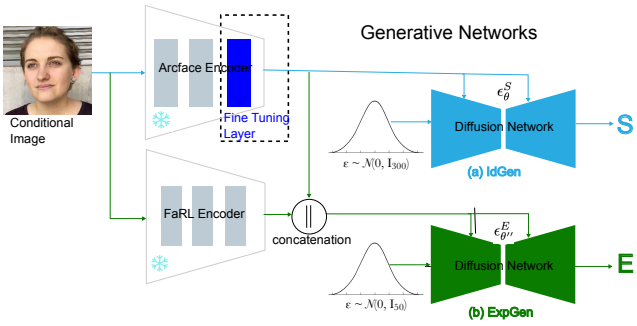

Figure 4. **Identity and expression generative networks.** (IdGen, in **blue**, and ExpGen, in **green**). For **IdGen**, the input image is encoded into a 512-dimensional embedding using ArcFace [9]. This serves as a condition for the 1D U-Net diffusion network, which is trained to denoise 300-dimensional noise into FLAME shape coefficients, S. For **ExpGen**, the input image is encoded into a 1024-dimensional embedding using the FaRL [59] and ArcFace [9] encoders. The embedding serves as a condition for the 1D U-Net diffusion network, which is trained to denoise 50-dimensional noise into FLAME expression coefficients, E.

3.2. Identity Generative Network (IdGen)

The first step of OFER is to generate a set of FLAME shape coefficients, S. The identity network ϵ_θ^S (Fig. 4, top) consists of a 1-dimensional U-Net [46] with self-attention layers that is trained to transform unit Gaussian noise $s_T \sim \mathcal{N}(0, I_{300})$ into $s_0 = S$.

The input image is used as conditioning by encoding it into a 512-dimensional feature vector $c_a \in \mathbb{R}^{512}$ using the ArcFace [9] encoder. ArcFace, being specifically trained to enhance face recognition, produces deep features that are highly robust to differences in poses, illumination, and external distractors, making it a more suitable choice for training under occlusions compared to landmark detectors. Following MICA [61], we freeze all layers of the ArcFace encoder except the last, such that the encoded features can be fine-tuned during training. During sampling, we generate a set of N noise vectors $\{s_T^i \in \mathbb{R}^{300}\}_{i=1}^N$ and iteratively denoise them using ϵ_θ^S to obtain the FLAME shape coefficients, $\{s_0^i = \hat{S}_i \in \mathbb{R}^{300}\}_{i=1}^N$.

We optimize the parameters θ by minimizing the following loss function L_ε:

$$L_\varepsilon(\theta) = \mathbb{E}_t\Big[\mathbb{E}_{s_0,\varepsilon}[|\epsilon_\theta^S(s_t) - \varepsilon_t|]\Big], \tag{2}$$

where ε_t is fixed Gaussian noise with variance proportional to the linear time step t [26]. Due to the nature of diffusion models, each generated sample $\hat{S}_i$ exhibits plausible variations in the occluded regions, capturing the variance in the input data.

3.3. Identity Ranking Network (IdRank)

Our goal is to generate a set of 3D faces with diverse expressions that align with the input occluded image. While expressions can vary, we would like to preserve a unique and consistent identity across all samples. This is motivated by three main observations. First, we noticed that shape coefficients display much less variability than expression coefficients, as images typically contain more cues for the expression-independent facial features. Second, we noticed that, when computing the reconstruction error for the N generated samples, the minimum value is on par with state-of-the-art methods (see supp.mat.). Since identity is typically more defined than expression, a method capable of selecting this optimal sample would be advantageous. Finally, even in cases of severe occlusions where identity features may be less discernible, having a method that effectively filters out poorly generated samples is important for practical applications. Towards this end, we introduce IdRank, a network designed to rank the outputs of IdGen in order to select those that best match the input image.

Ranking framework. The overall architecture of IdRank is shown in Fig. 5. Given a pool of N candidate shapes generated by IdGen, the goal is to select the one closest in distance to the ground-truth neutral shape (available during training). To achieve this, we train a *ranking* network $R_{\theta'}$ that goes from a list of v face vertices (derived from the shape coefficients) to a scalar p representing the probability that the sample matches the ground truth. We then apply the network $R_{\theta'}$ over the N samples, ranking each based on their scores and selecting the highest-scored sample.

Before passing the mesh to the network, we remove the vertices corresponding to the back of the head since the focus is on facial reconstruction, and their inclusion in the error computation might lead to spurious results. The number of candidates N is an important hyper-parameter, as it defines a trade-off between computational complexity and reconstruction error – a large value of N means that the actual ground-truth shape has more chances of being generated, but incurs a significantly higher computational cost. We empirically choose here $N = 100$.

Generating ground-truth data. During training, we generate the ground-truth data for ranking in an online fashion as follows. Given an image I together with its ground-truth FLAME mesh $M^{GT} \in \mathbb{R}^{5023}$, we first sample N shape parameters $\{\beta_i\}_{i=1}^N$ using IdGen (with gradient back-propagation disabled). We then reconstruct the neutral face vertices using Eq. (1), yielding a set of mesh candidates $\mathcal{M} = \{M_i = M(\beta, \mathbf{0}, \mathbf{0}) \in \mathbb{R}^{5023}\}_{i=1}^N$. From the set $\mathcal{M}$ as well as the ground-truth mesh M^{GT} we retain only the frontal vertices using a pre-computed mask, resulting in $\mathcal{M}_{\text{frontal}} = \{\hat{M}_i \in \mathbb{R}^{n'}\}_{i=1}^N$ and $M^{GT}_{\text{frontal}} \in \mathbb{R}^{n'}$ ($n' < 5023$). Finally, we compute the error between each $\hat{M}_i$ and M^{GT}_{frontal}

to produce the following set:

$$\mathcal{D}_{GT} = \{|\hat{M}_i - M^{GT}_{\text{frontal}}|\}_{i=1}^{N}. \quad (3)$$

with $|\cdot|$ the L1 norm. We empirically observed that removing the mean $\mu_{\mathcal{M}} = \frac{1}{N}\sum_{i=1}^{N}\hat{M}_i$ from the set $\mathcal{M}_{\text{frontal}}$ reduces redundancy and helps convergence of the network. Hence, we transform $\mathcal{M}_{\text{frontal}}$ into a zero-centered set of residual meshes $\mathcal{M}_{\text{center}}$ via:

$$\mathcal{M}_{\text{center}} = \{M'_i = (\hat{M}_i - \mu_{\mathcal{M}})\}_{i=1}^{N}. \quad (4)$$

The final training set for image I is defined as the list of pairs $\mathbf{X} = [(\mu_{\mathcal{M}}, M'_i)]_{i=1}^{N}$, obtained by sorting the values in $\mathcal{D}_{GT}$ in ascending order.

Learning to rank facial shapes. The ranking network $R_{\theta'}$ is a small MLP that takes as input a single pair $(\mu_{\mathcal{M}}, M'_i)$ and is trained to predict a sample score d_i. It is conditioned on a 1024-dimensional feature vector $c_R = (c_a \| c_f)$, formed by concatenating features from the ArcFace [9] (c_a) and FaRL [59] (c_f) encoders, designed for facial analysis tasks. The network is run N times producing $\mathcal{D} = \{d_1, \ldots, d_N\}$ scores, which are then passed through a softmax operator to transform them into probabilities, $P = \text{softmax}(\mathcal{D}) = \{p_i = \frac{e^{d_i}}{\sum_{j=1}^{N} e^{d_j}}\}_{i=1}^{N}$. The ground-truth distance values $\mathcal{D}_{GT}$ are also passed through a softmax operator, $\mathcal{D}'_{GT} = \text{softmax}(\mathcal{D}_{GT})$. Finally we compute the cross-entropy loss between the two distributions P and $\mathcal{D}'_{GT}$:

$$L_{\mathcal{D}'_{GT},P}(\theta') = -\sum_{i=1}^{N} \mathcal{D}'_{GT}(i)\log(p_i). \quad (5)$$

Inference. During inference, we select the sample with the highest probability score as the optimal sample, denoted as S_R. While multiple samples may have equal probability scores, their ordering in such cases would be rather random. In such cases, we choose one of the highest ranked equal probables samples.

3.4. Expression Generative Network (ExpGen)

The expression network $\epsilon^{E}_{\theta''}$ architecture (Fig. 4, bottom) closely follows that of IdGen, employing the conditional DDPM framework to generate the first 50 components of the FLAME expression coefficients. It is conditioned on a 1024-dimensional feature vector, similar to IdRank. During sampling, the network initializes from unit Gaussian noise $\varepsilon_T \in \mathbb{R}^{50}$ and iteratively denoises towards the FLAME expression parameters, $e_0 \in \mathbb{R}^{50}$. The parameters θ'' are optimized by minimizing the loss function:

$$L_{\varepsilon}(\theta'') = \mathbb{E}_t\Big[\mathbb{E}_{e_0,\varepsilon}[|\epsilon^{E}_{\theta}(e_t) - \varepsilon_t|]\Big]. \quad (6)$$

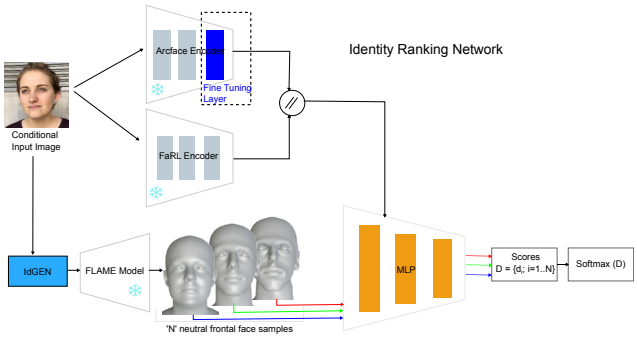

Figure 5. **Identity Ranking Network.** Given the N shape coefficients from IdGen, we first reconstruct the neutral meshes using FLAME. Each mesh is passed through a 5-layer MLP to compute a score, conditioned on the input image. The N scores are then converted into probabilities using softmax. The ranking order of the sorted scores is compared against the ranking of the sorted reconstruction errors, and the network is trained to match them.

4. Experiments

In this section, we present qualitative and quantitative evaluation results for OFER. We begin by introducing the datasets used for training and testing, including the proposed CO-545 dataset, which we specifically designed to evaluate performance on occluded images (Sec. 4.2). Next, we provide evaluation metrics and experimental results on IdGen and IdRank (Sec. 4.3), followed by a similar analysis for the final reconstructions with ExpGen (Sec. 4.4).

4.1. Setup

We use the DDPM framework [26] using the DDPM sampler with 1000 sampling time steps for both IdGen and ExpGen. We trained IdGen and ExpGen for 400 epochs with a batch size of 128 on an RTX 8000 GPU. The training process spans one GPU day. IdRank was trained for 120 epochs with a batch size of 32 on a Titan X GPU. The training process ran for 20 epochs per day, converging over 6 days.

4.2. Datasets

Training data. We train IdGen with the Stirling [22], FaceWarehouse [6], LYHM [7], and Florence [2] datasets, which contain 2D-3D ground-truth pairs of images and corresponding registered FLAME shape coefficients. The same four datasets are used to train IdRank. For ExpGen we use the Facial Motion across Subjects (FaMoS) [4] dataset, which offers paired images and FLAME 3D meshes of faces in motion. The shape and expression coefficients were provided by the FaMoS authors.

The CO-545 dataset. To evaluate the reconstruction of expressive faces under occlusions, we propose a new dataset built on top of the CoMA dataset [44]. The CO-545 vali-

dation dataset comprises 545 frontal face images (no profile or rotated head views) synthetically occluded (hand, face, and random) across 11 subjects taken from CoMA [44]. For each image, we identify the sets of occluded and unoccluded vertices and evaluate them using separate metrics. More details can be found in the supplementary material.

Testing data. We evaluate the final expressive reconstructions using the dataset provided by Dey *et al*. [12] ("Dey Dataset"), and the CO-545 dataset. We additionally evaluate the quality of neutral shape reconstructions using the occluded and unoccluded subsets of the NoW benchmark [49]. Note that we do not have access to the occluded/unoccluded split for the testing set, and hence provide results mainly for the validation set. Additional comparisons on the test set are included in the supplementary material.

4.3. Evaluation on IdGen and IdRank

Baselines and Metrics. We quantify the errors using the occluded and unoccluded subsets of the NoW validation dataset, in terms of mean, median, and standard deviation of the mean square error (MSE) relative to the ground-truth mesh. Since OFER is designed to provide multiple solutions, we establish a baseline by randomly sampling FLAME shape coefficients from its parametric space. We compare our method to standard reconstruction techniques as well as state-of-the-art occlusion-based approaches, Diverse3D [12] and FOCUS [32].

Results. Fig. 6 and Fig. 7 present qualitative results of 3D shape reconstruction from [49] and [19]. We compare our method alongside the state-of-the-art neutral face reconstruction method MICA [61] and the occluded face reconstruction method by Dey *et al*. [12], Diverse3D. In Fig. 6 we observe that, despite mild occlusions, OFER can recover accurate shapes with distinctive features (*e.g*., cheeks in the first row and the head length in the second row). We show an example of extreme occlusion in Fig. 7, where we reconstruct two images of the same person wearing a face mask, alongside a minimally occluded reference image. We observe that OFER, and in particular the top-ranked sample, produces plausible results for the two masked images.

We present quantitative results in Tab. 2. Our method achieves a lower average error across all reconstructed samples compared to other state-of-the-art occlusion-based approaches. In addition, the shapes generated by IdGen can achieve competitive results with standard reconstruction methods, as shown by the minimum reconstruction error obtained over all samples ("OFER-min").

Ablation. To highlight the effectiveness of ranking in improving the search space and in selecting high-quality samples, we conducted an ablation study, with results presented in Tab. 1. In row (a), we show the results of generating 200 and 1000 random FLAME samples, while row (d) presents the results from OFER. The "ideal lowest error" column

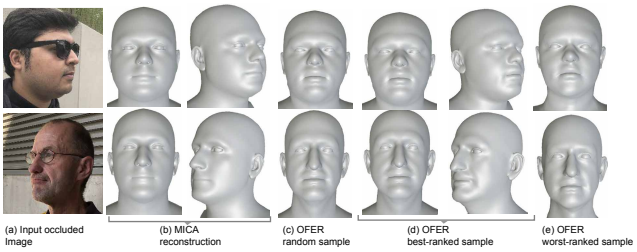

Figure 6. **Neutral face reconstruction on occluded images.** For (a) the given occluded input image, (b) shows the reconstructed shape provided by MICA [61]; (c) is one of the generated samples from our method; (d) and (e) are the best and worst-ranked samples, respectively, as selected by the ranking network.

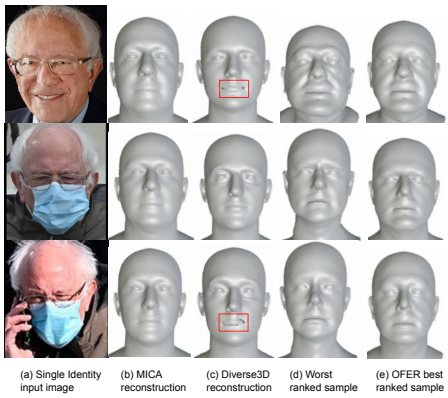

Figure 7. **Neutral face reconstruction on a single identity.** For (a) the given input image, (b) shows the reconstructed shape provided by MICA [61]; (c) shows the reconstructed shape provided by Diverse3D [12]; (d) and (e) are the worst and best-ranked samples as selected by the ranking method.

corresponds to an ideal ranking (*i.e*., the minimum error within the samples). Even with an increased number of random hypotheses (1000 samples), finding a good reconstruction in this space remains less likely, as indicated by the higher error in the ideal ranking scenario (0.81 with OFER vs 0.90 with FLAME), showcasing the improved search space when ranking is applied to IdGen. In row (b), we show the results of ranking a set consisting of 50% random FLAME samples and 50% OFER reconstructions; row (c) reflects the same for an $80/20$ split. This mimics a scenario where the samples to rank include both correct solutions and high-error ones. Here, the ranked results (1st and 2nd columns) are consistently lower than the average error (3rd column), highlighting its effectiveness in selecting low-error samples. This is further supported by qualitative evidence in Fig. 8.

4.4. Evaluation on reconstructions with ExpGen

Baselines and Metrics. We compare our method against the state-of-the-art diverse expression reconstruction ap-

Samples	Ideal Lowest Error Sample			Ranked Sample			Average of All Samples		
	Med ↓	Mean ↓	std ↓	Med ↓	Mean ↓	std ↓	Med ↓	Mean ↓	std ↓
(a)FLAME(200 - 1k)	(1.00 - 0.90)	(1.31 - 1.20)	(1.18 - 1.11)	NA	NA	NA	1.75	2.19	1.86
(b)FLAME(50)+OFER(50)	0.84	1.09	0.95	1.08	1.44	1.32	1.33	1.79	1.62
(c)FLAME(80)+OFER(20)	0.86	1.10	0.97	1.34	1.81	1.65	1.60	2.06	1.84
(d)OFER(100)	**0.81**	**1.05**	**0.92**	**0.98**	**1.21**	**1.01**	**1.02**	**1.25**	**1.04**

Table 1. **Ablation study on the importance of ranking.** *Ideal Lowest Error Sample* refers to the sample with the lowest median MSE. *Ranked Sample* is the sample selected by the Ranking Network. *Average of All Samples* represents the average error across all generated samples. "NA" in the Ranked Sample column indicates that our network cannot validate these larger sample sizes.

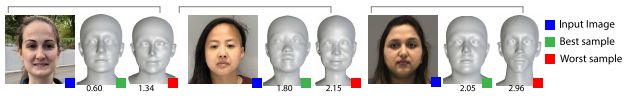

Figure 8. **Qualitative and quantitative results for best and least ranked samples** for row (b) in Tab. 1 evaluated on NoW validation images. The median MSE error between the ground-truth scan and the reconstructed 3D shape is displayed below each rendering. These results demonstrate that ranking as a selection method enhances the quality of sample selection.

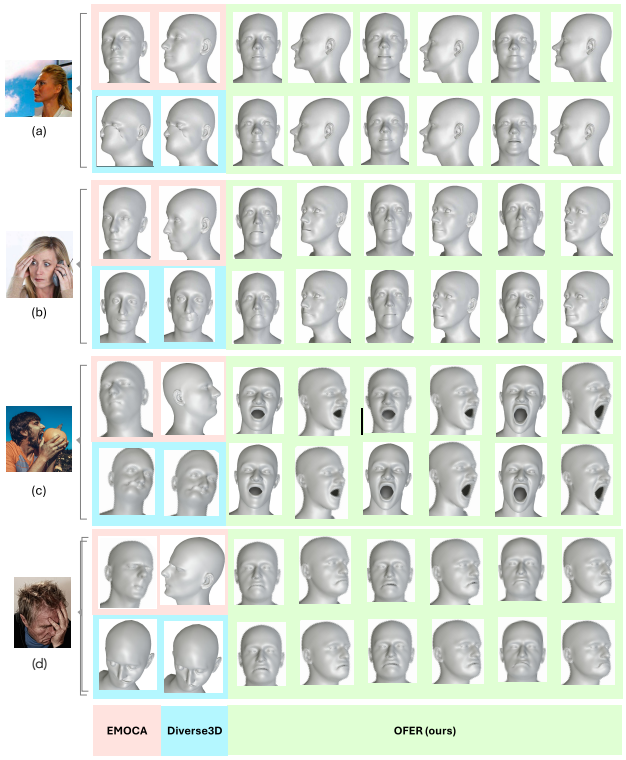

Figure 9. **Comparison of expression reconstruction for in-the-wild occluded images.** We compare against EMOCA [8] (front and side view in pink), two reconstructions from Diverse3D [12] (blue), and six samples (front and side view) from OFER (green).

proach by Dey *et al*. [12], and the single-hypothesis expression reconstruction method EMOCA [8]. For quantitative

Method	Unoccluded			Occluded			Both		
	Med	Mean	std	Med	Mean	std	Med	Mean	std
FLAME	1.79	2.24	1.89	1.80	2.26	1.91	1.79	2.25	1.90
Deep3D [11]	1.33	1.67	1.41	1.40	1.73	1.41	1.36	1.70	1.41
DECA [21]	1.18	1.47	1.24	1.29	1.56	1.29	1.17	1.46	1.25
MICA (4DS) [61]	n/a	n/a	n/a	n/a	n/a	n/a	1.02	1.25	1.05
MICA (8DS) [61]	n/a	n/a	n/a	n/a	n/a	n/a	0.90	1.11	0.92
TokenFace [58]	n/a	n/a	n/a	n/a	n/a	n/a	**0.79**	**0.99**	**0.85**
FOCUS [32]	1.03	1.25	1.03	1.07	1.34	1.19	1.05	1.31	1.14
FOCUS (MP) [32]	1.02	1.24	1.02	1.08	1.34	1.20	1.03	1.29	1.12
Diverse3D [12]	n/a	n/a	n/a	n/a	n/a	n/a	1.41	1.78	1.52
OFER-rank (ours)	**0.97**	**1.20**	**1.00**	**1.01**	**1.26**	**1.05**	**0.98**	**1.21**	**1.01**
OFER-min	0.81	1.04	0.91	0.84	1.10	0.96	0.81	1.05	0.92
OFER-avg	1.01	1.25	1.03	1.04	1.29	1.08	1.01	1.25	1.04

Table 2. **Quantitative evaluation on the NoW validation benchmark**, on the unoccluded, occluded and full set. "FLAME" is created by sampling random FLAME shape coefficients. Rows 2-6 show standard reconstruction methods, while the bottom rows show occlusion-based approaches. OFER-rank is our result using IdGen and IdRank networks. The bottom two rows show the minimum (OFER-min) and average (OFER-avg) reconstruction errors obtained over the 100 samples provided by IdGen. We show median (med), mean, and standard deviation (std) of the non-metrical MSE between the reconstructed and ground truth shapes.

evaluation on Dey *et al*. [12] we generate 15 expressions per image. It is important to note that Dey *et al*. [12] employs test-time optimization, whereas our approach does not. Additionally, their network is trained on the CoMA dataset, the same dataset used in the creation of CO-545. As a result, the evaluation data overlaps with their training set, potentially giving their results an additional advantage.

Building on [56] and [12] we evaluate using six metrics: (1) Sample Error (**SE**): the average Root Mean Square Error (RMSE) between the landmark vertices of the reconstructed samples and the ground truth; (2) Closest Sample Error (**CSE**): the SE of the closest reconstructed sample; (3) Average Self Distance-Visible (**ASD-V**): to ensure that the visible regions of the reconstructions remain close to the input, the maximum distance across samples should be minimized; **maxASD-V** measures the average per-vertex RMSE on the visible regions of a sample and its "farthest" neighboring sample across all instances, while **minASD-V** measures ASD-V with the "nearest" neighbour sample; (4) Average Self Distance-Occluded (**ASD-O**): the reconstructions of *occluded* regions should show high diversity, given by $d^o = \frac{\text{RMSE}(\text{gt}^v, \text{x}_i^v)}{\max(\text{MD})^o}$, while the visible regions must remain close to the ground truth, given by $d^v = \frac{\text{RMSE}(\text{x}_i^o, \text{x}_j^o)}{\sqrt{\max(\text{MD})^o}}$. Here, $\max(\text{MD})^o$ is the maximum mahalanobis distance threshold of the occluded vertices, and x_j is the nearest neighbour sample of x_i based on the L2 distance of occluded vertices. The superscript v and o point to the visible and occluded vertices. We compute ASD-O as the harmonic mean (HM) between d^v and d^o, given by $\frac{2 \cdot (1-d^v) \cdot d^o}{(1-d^v)+d^o}$; (5) Shape Standard Deviation (**STD-S**): in case of absence of ground-truth 3D data, we opt for the standard deviation metric of the

Method	N_C^{unocc}	N_C^{occ}	E_D^{occ}	E_C^{unocc}	E_C^{occ}
	RMSE ↓	RMSE ↓	RMSE ↓	RMSE ↓	RMSE ↓
Diverse3D[12]	2.44	2.16	6.4	7.46	4.72
OFER (ours)	1.90	1.95	**3.2**	**3.48**	**2.78**

Table 3. **Quantitative evaluation using SE** on unoccluded (*unocc*) and occluded images (*occ*) in Dey and CO-545 datasets. The letters represent: N for Neutral, E for expression, D for Dey dataset, and C for CO-545 dataset. Reconstructions evaluated on all front face vertices for *unocc* images and only on unoccluded vertices for *occ* images.

Method	CSE	minASD-V	maxASD-V	ASD-O	ODE	ODE-O	ODE-V
	RMSE ↓	RMSE ↓	RMSE ↓	HM ↑	max md ↓	max md ↓	max md ↓
Diverse3D[12]	0.30	1.31	1.96	0.16	0.4	0.57	0.33
OFER (ours)	**0.17**	**0.65**	**1.39**	**0.18**	**0.01**	**0.01**	**0.01**

Table 4. **Quantitative evaluation of expression reconstructions using CSE, ASD-V and ODE** on the CO-545 dataset.

Method	STD-S			ODE		
	mask	*sunglasses*	*naturalocc*	*mask*	*sunglasses*	*naturalocc*
	RMSE ↑	RMSE ↑	RMSE ↑	RMSE ↓	RMSE ↓	RMSE ↓
Diverse3D[12]	11.81	21.28	20.84	0.95	0.34	0.41
OFER (ours)	**34.04**	**34.38**	**34.56**	**0.002**	**0.001**	**0.003**

Table 5. **Quantitative evaluation of expression reconstructions using STD-S and ODE metrics** evaluated on 200 images per subset (mask, sunglasses and natural occlusions) from [19].

entire shape, excluding data which falls outside the maximum Mahalanobis distance (MD) per vertex calculated using CoMA [44]; (6) Out-of-Distribution Error (**ODE**): ODE measures the average per-vertex Mahalanobis distance out-of-distribution error. It addresses the observation that reconstructions from Dey *et al*. [12] occasionally produce non-plausible expressions that do not fall within the distribution of ground truth expressions.

Results. The qualitative comparison of diverse expressive faces for occluded images is shown in Fig. 2 and Fig. 9, taken from [19] and [30, 41]. Our approach effectively generates plausible faces with varied expressions, even under challenging occlusions such as face masks. In contrast, EMOCA [8] tends to produce degenerate faces in similar hard-occlusion scenarios, while Dey *et al*. [12] results in unrealistic and extreme reconstructions. The quantitative evaluation shown in Tab. 3 and Tab. 4 also supports this, particularly highlighting the substantial improvement over Dey *et al*. [12]. Our method demonstrates a significant advantage, reliably reconstructing visible regions and producing varied expressions in occluded areas. Further, we tested on the real-world hard-occlusion dataset provided by Erakiotan *et al*. [19] using the STD-S and ODE metrics only, since it does not have any 3D ground truth. The results presented in Tab. 5 further shows that OFER generates plausible expressions within the face shape distribution.

5. Limitations and Future work

3D face reconstruction has many valuable applications, but if not carefully regulated poses risks like privacy violations and misuse in surveillance.

The generative capabilities of diffusion models with diverse inputs are attributed to the amount and variations in the training data [48]. In our case, we had access to only a small subset of the available 3D supervision dataset for both IdGen and ExpGen. As a result, our generated expressions may lack certain expressive details. Therefore, a promising direction for future research is to integrate both 2D and 3D supervision to improve the performance of the method.

When we refer to performance, we specifically measure it in terms of how closely the reconstructed model approximates the ground-truth scan. In spite of our ranking network selecting a sample that minimizes the error compared to a randomly chosen one from the generated set, it does not guarantee the selection of the absolute best match, *i.e*., the sample with the lowest error and the closest approximation to the ground truth. From this observation we identify two potential research directions for constructing an ideal ranking network. The first is to refine the ranking architecture itself by adopting a more suitable loss function – Softmax scoring is not ideal when presented with large number of sample sets since it dilutes the scores, making it challenging to identify the best among the higher-quality samples. Additionally, a more sophisticated ranking mechanism, e.g. [5], could complement the selection of quality samples. A second promising approach involves integrating the ranking process as a feedback mechanism during training of the diffusion model. Rather than using a standalone model, this integration could improve the quality of the generated samples, resulting in improved overall performance.

6. Conclusion

In this paper we introduced OFER, a conditional diffusion-based method for generating multiple hypotheses of expressive faces from a single-view, in-the-wild occluded image. Key to OFER is the use of two diffusion models to generate FLAME 3DMM shape and expression coefficients. To ensure a consistent geometric face shape for varied expression of a single identity, we introduced a probabilistic ranking method to select an optimal sample from the generated shape coefficients. By combining the statistical learning strengths and generative capabilities of diffusion models, along with the smooth face reconstruction provided by a parametric model, our method produces plausible 3D faces that accurately reflect the input image. OFER achieves state-of-the-art results in diverse expression reconstruction outperforming existing occlusion-based methods, and can generate plausible and diverse results for a given input.

References

[1] Oswald Aldrian and William A. P. Smith. Inverse rendering of faces with a 3D morphable model. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(5): 1080–1093, 2013. 2

[2] Andrew D. Bagdanov, Alberto Del Bimbo, and Iacopo Masi. The florence 2D/3D hybrid face dataset. In *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pages 79–80, 2011. 5, 2, 3

[3] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *International Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 187–194, 1999. 2

[4] Timo Bolkart, Tianye Li, and Michael J. Black. Instant multi-view head capture through learnable registration. In *Computer Vision and Pattern Recognition (CVPR)*, pages 768–779, 2023. 5, 3

[5] João Brito and João Mendes-Moreira. Instance ranking with multiple linear regression: Pointwise vs. listwise approaches. In *Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6, 2014. 8, 3

[6] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. FaceWarehouse: A 3D facial expression database for visual computing. *Transactions on Visualization and Computer Graphics (TVCG)*, 20(3):413–425, 2014. 5, 3

[7] Hang Dai, Nick E. Pears, William A. P. Smith, and Christian Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision (IJCV)*, 128 (2):547–571, 2020. 5, 3

[8] Radek Daněček, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 20279–20290, 2022. 2, 3, 7, 8, 5, 6

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 3, 4, 5, 1

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-shot multi-level face localisation in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5202–5211, 2020. 3

[11] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*, pages 285–295, 2019. 2, 3, 7, 4

[12] Rahul Dey and Vishnu Naresh Boddeti. Generating diverse 3D reconstructions from a single occluded face image. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1537–1547, 2022. 2, 3, 6, 7, 8, 5

[13] Haojie Diao, Xingguo Jiang, Yang Fan, Ming Li, and Hongcheng Wu. 3D Face reconstruction based on a single image: A review. *IEEE Access*, 12:59450–59473, 2024. 2

[14] Abdallah Dib, Cédric Thébault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *International Conference on Computer Vision (ICCV)*, pages 12799–12809, 2021. 4

[15] Pengfei Dou, Shishir K. Shah, and Ioannis A. Kakadiaris. End-to-end 3D face reconstruction with deep neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1503–1512, 2017. 3

[16] Bernhard Egger, Andreas Schneider, Clemens Blumer, Andreas Forster, Sandro Schönborn, and Thomas Vetter. Occlusion-aware 3D morphable face models. In *British Machine Vision Conference (BMVC)*, 2016. 2, 3

[17] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3D morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision (IJCV)*, 126(12): 1269–1287, 2018. 2, 3

[18] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhöfer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D Morphable face models—past, present, and future. *Transactions on Graphics (TOG)*, 39(5):157:1–157:38, 2020. 3

[19] Mustafa Ekrem Erakın, Uğur Demir, and Hazım Kemal Ekenel. On recognizing occluded faces in the wild. In *International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 189–196, 2021. 6, 8

[20] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3D face reconstruction and dense alignment with position map regression network. In *European Conference on Computer Vision (ECCV)*, pages 557–574, 2018. 3

[21] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics (TOG)*, 40(4):88:1–88:13, 2021. 2, 3, 7, 4

[22] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3D reconstruction from 2D face images in the wild. In *International Conference on Automatic Face & Gesture Recognition (FG)*, pages 780–786, 2018. 5, 2, 3

[23] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. GANFIT: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1155–1164, 2019. 2

[24] Riza Alp Güler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. DenseReg: Fully convolutional dense shape regression in-the-wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2614–2623, 2017. 3

[25] Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. Learning-to-rank with BERT in TF-Ranking. *arXiv:2004.08476*, 2020. 3

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33:6840–6851, 2020. 2, 4, 5

[27] Harim Jung, Myeong-Seok Oh, and Seong-Whan Lee. Learning free-form deformation for 3D face reconstruction

from in-the-wild images. In *International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2737–2742, 2021. 3

[28] Hyeyoung Ko, Suyeon Lee, Yoonseo Park, and Anna Choi. A survey of recommendation systems: Recommendation models, techniques, and application fields. *Electronics (E)*, 11(1):141, 2022. 3

[29] Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning (FTML)*, 5(2-3):123–286, 2012. 2, 3

[30] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2144–2151, 2011. 8

[31] Kyungjin Lee, Juheon Yi, and Youngki Lee. FarfetchFusion: Towards fully mobile live 3D telepresence platform. In *International Conference on Mobile Computing and Networking (MCN)*, pages 20:1–20:15, 2023. 2

[32] Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. Robust model-based face reconstruction through weakly-supervised outlier segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 372–381, 2023. 2, 3, 6, 7, 4

[33] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *Transactions on Graphics (TOG)*, 36(6):194:1–194:17, 2017. 2, 3

[34] Kevin Lin, Dianqi Li, Xiaodong He, Ming-Ting Sun, and Zhengyou Zhang. Adversarial ranking for language generation. In *NeurIPS*, pages 3155–3165, 2017. 3

[35] Tie-Yan Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009. 3, 1

[36] Hai Thanh Nguyen, Thomas Almenningen, Martin Havig, Herman Schistad, Anders Kofod-Petersen, Helge Langseth, and Heri Ramampiaro. Learning to rank for personalised fashion recommender systems via implicit feedback. In *International Conference on Mining Intelligence and Knowledge Exploration (MIKE)*, pages 51–61, 2014. 3

[37] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. 2024. 1

[38] Zhaoxi Pan, Song Tian, Mengzhao Guo, Jianxun Zhang, Ningbo Yu, and Yunwei Xin. Comparison of medical image 3D reconstruction rendering methods for robot-assisted surgery. In *International Conference on Advanced Robotics and Mechatronics (ICARM)*, pages 94–99, 2017. 2

[39] Devi Parikh and Kristen Grauman. Relative attributes. In *International Conference on Computer Vision (ICCV)*, pages 503–510, 2011. 3

[40] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 296–301, 2009. 3

[41] PickPik. Pickpik - free stock photos. https://www.pickpik.com. 8

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh andSandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1

[43] Razieh Rahimi, Ali Montazeralghaem, and James Allan. Listwise neural ranking models. In *International Conference on the Theory of Information Retrieval (ICTIR)*, page 101–104, 2019. 1

[44] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 725–741, 2018. 2, 5, 6, 8

[45] George Retsinas, Panagiotis Paraskevas Filntisis, Radek Daněček, Victoria Fernández Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3D Facial expressions through analysis-by-neural-synthesis. In *Computer Vision and Pattern Recognition (CVPR)*, pages 2490–2501, 2024. 3

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015. 4

[47] Zeyu Ruan, Changqing Zou, Longhai Wu, Gangshan Wu, and Limin Wang. SADRNet: Self-aligned dual face regression networks for robust 3D dense face alignment and reconstruction. *Transactions on Image Processing (TIP)*, 30: 5793–5806, 2021. 3

[48] Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *AAAI Conference on Artificial Intelligence*, pages 4695–4703, 2024. 8

[49] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J. Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, 2019. 2, 6, 4

[50] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *International Conference on Computer Vision (ICCV)*, pages 1585–1594, 2017. 3

[51] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015. 1

[52] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian

Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 3735–3744, 2017. 3

[53] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard G. Medioni. Extreme 3D face reconstruction: Seeing through occlusions. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3935–3944, 2018. 2

[54] Kenny T. R. Voo, Liming Jiang, and Chen Change Loy. Delving into high-quality synthetic face occlusion segmentation datasets. *Computer Vision and Pattern Recognition Workshops (CVPRw)*, pages 4710–4719, 2022. 3

[55] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3D face reconstruction with dense landmarks. In *European Conference on Computer Vision (ECCV)*, pages 160–177, 2022. 4

[56] Ye Yuan and Kris M. Kitani. Diverse trajectory forecasting with determinantal point processes. In *International Conference on Learning Representations (ICLR)*, 2020. 7

[57] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. DF2Net: A dense-fine-finer network for detailed 3D face reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 2315–2324, 2019. 3

[58] Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, and Yu Li. Accurate 3D face reconstruction with facial component tokens. In *International Conference on Computer Vision (ICCV)*, pages 8999–9008, 2023. 3, 7, 4

[59] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. In *Computer Vision and Pattern Recognition (CVPR)*, pages 18676–18688, 2022. 4, 5, 1, 3

[60] Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. RankT5: Fine-tuning T5 for text ranking with ranking losses. In *International Conference on Research and Development in Information Retrieval (SIGIR)*, pages 2308–2313, 2023. 3

[61] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision (ECCV)*, pages 250–269, 2022. 2, 3, 4, 6, 7

[62] Xuan Zou, Josef Kittler, and Kieron Messer. Illumination invariant face recognition: A survey of passive methods. In *International Conference on Biometrics: Theory, Applications, and Systems (BTAS)*, pages 1–8, 2007. 2