SIDER: Single-Image Neural Optimization for Facial Geometric Detail Recovery

Aggelina Chatziagapi[1*] ShahRukh Athar[1*] Francesc Moreno-Noguer[2]
Dimitris Samaras[1]

[1]Stony Brook University
[2]Institut de Robòtica i Informàtica Industrial, CSIC-UPC

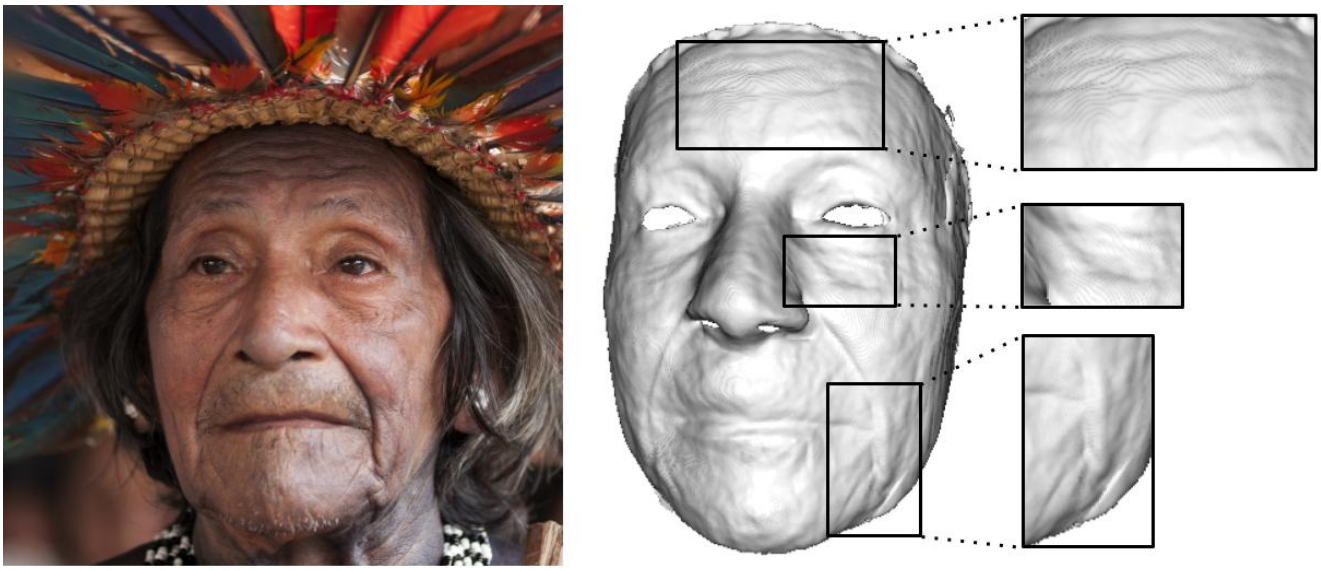

Figure 1: **SIDER** is a novel photometric optimization method that recovers, from a single image, detailed facial geometry without any 3D, multi-view or multiple image supervision. As shown above, the details recovered by SIDER such as wrinkles, skin folds and skin bumps are realistic and have high fidelity to the input image (see insets).

Abstract

*We present SIDER (**S**ingle-**I**mage neural optimization for facial geometric **DE**tail **R**ecovery), a novel photometric optimization method that recovers detailed facial geometry from a single image in an unsupervised manner. Inspired by classical techniques of coarse-to-fine optimization and recent advances in implicit neural representations of 3D shape, SIDER combines a geometry prior based on statistical models and Signed Distance Functions (SDFs) to recover facial details from single images. First, it estimates a coarse geometry using a morphable model represented as an SDF. Next, it reconstructs facial geometry details by optimizing a photometric loss with respect to the ground truth image. In contrast to prior work, SIDER does not rely on any dataset priors and does not require additional supervision from multiple views, lighting changes or ground truth 3D shape. Extensive qualitative and quantitative evaluation demonstrates that our method achieves state-of-the-art on facial geometric detail recovery, using only a single in-the-wild image.*

1. Introduction

The study of the 3D geometry of the human face is a problem of great interest in computer graphics and vision communities. Early approaches to recover the 3D structure of the human face were based on morphable models [1], where the face shape, expression and texture are optimized

*Equal Contribution

with respect to a given image. However, due to the low dimensionality of the identity and expression subspaces, morphable models are unable to capture facial geometric details, such as wrinkles and skin-folds. With the advent of deep learning, it has now become possible to train networks on large datasets [6, 23, 26], in order to regress the 3D shape of the face along with its geometric details. Still, the generalization ability of these methods is limited, mainly due to the lack of diversity of the data they are trained on [8].

Recent advances in implicit neural representations of 3D shape [9, 24, 31] have made it possible to learn rich details of the 3D geometry of an object. The geometry can be represented by Signed Distance Functions (SDFs) [9, 24, 31], Unsigned Distance Functions (USDFs) [3] or occupancies [17] that are parameterized using Multilayer Perceptrons (MLPs). The high representational power of MLPs, along with the use of positional encoding [18], facilitates the reconstruction of rich geometric details. Additionally, if the geometry is represented as an SDF, recent works [19, 31] have proposed ways that allow the cheap calculation of derivatives of the geometry through implicit differentiation, making gradient learning significantly more tractable. These implicit representations are learnt using either multi-view supervision [19, 31] or partial 3D data [9, 24]. However, in the absence of 3D data or multi-view supervision, like when one has access only to a single face image, it might not be possible to train such models as they may collapse into trivial solutions.

In this work, we propose SIDER, a novel photometric optimization method that recovers facial geometric details from a single face image. SIDER uses an unsupervised coarse-to-fine optimization scheme that does not require any ground truth 3D, multi-view or varying light-source supervision. Since optimization of SDFs using a single image is prone to trivial solutions, SIDER first learns a coarse approximation of the 3D face geometry using a morphable model fit to the input image by a standard landmark fitting pipeline. Next, this SDF is optimized by minimizing the photometric loss with respect to the given image via implicit differentiation [19, 31]. After converging, SIDER outputs an SDF that represents the 3D face shape along with its geometric details (see Fig 1). We show, both quantitatively and qualitatively, that SIDER significantly outperforms the current state-of-the-art in detailed face reconstruction by recovering facial geometric details that are realistic and have high fidelity to the input image.

To summarize, our contributions are as follows:

- We propose SIDER, a method that recovers facial geometric details from a single face image in an unsupervised manner.
- We propose a novel coarse-to-fine optimization scheme that leverages a classical morphable model representation as a prior to prevent degenerate solutions of the SDF and is optimized using an unsupervised photometric loss.
- We achieve state-of-the-art performance in facial geometry reconstruction from single in-the-wild images.

2. Related Work

In this section, we describe recent related works in facial geometry estimation, facial geometric details recovery and implicit representations of 3D shapes.

Facial Geometry Estimation. One of the first widely used methods for 3D shape reconstruction of the human face was based on statistical 3D face models that can fit a given image, going back to the original 3D Face Morphable Model [1] (3DMMs). However, the underlying PCA-based representation for shape and expression of 3DMMs is not flexible enough to represent fine facial geometric details such as wrinkles, skin folds and skin bumps. Recent methods [2, 5, 7, 11, 12, 14, 26, 27, 28, 29, 33] leverage the power of deep-learning and large-scale image and video datasets to regress parameters that generate realistic 3D reconstructions or learn complex representations for face shape, expression and texture. These methods produce more realistic results than traditional 3DMM fitting, but they still cannot capture fine facial details and need large datasets to train on. In contrast, we propose a neural coarse-to-fine optimization scheme to recover facial geometric details from single images and without supervision.

Geometric Facial Details Estimation. The past few years have witnessed a significant improvement in the realism of reconstructed 3D face geometry and the quality of facial details. In [21], a CNN (CoarseNet) first regresses a rough geometry of the face, and then another CNN (FineNet) estimates facial details using a coarse depth map and input images. In [23], the regressed correspondence and depth maps are registered onto a template mesh, which is further refined to generate the detailed facial geometry. In [30] facial details are modelled with bump maps on top of a 3DMM base. DF^{2}Net [32] uses multiple refinement steps to reconstruct detailed geometric structure. First, a coarse depth is predicted, which is then refined by an F-Net. The refined depth along with the input image is then given as input to a specially designed Finer-Net that outputs the final recovered facial details. DECA [6] uses a differentiable renderer to perform 3D face reconstruction and recover the detailed face geometry. The facial geometric details are represented as a UV-map of vertex displacements of a FLAME mesh[15]. DECA is trained on a large dataset of close to 2 million images with a subset of them being paired. In contrast to the aforementioned methods, SIDER does not require an exorbitantly large dataset for training and can be used on single

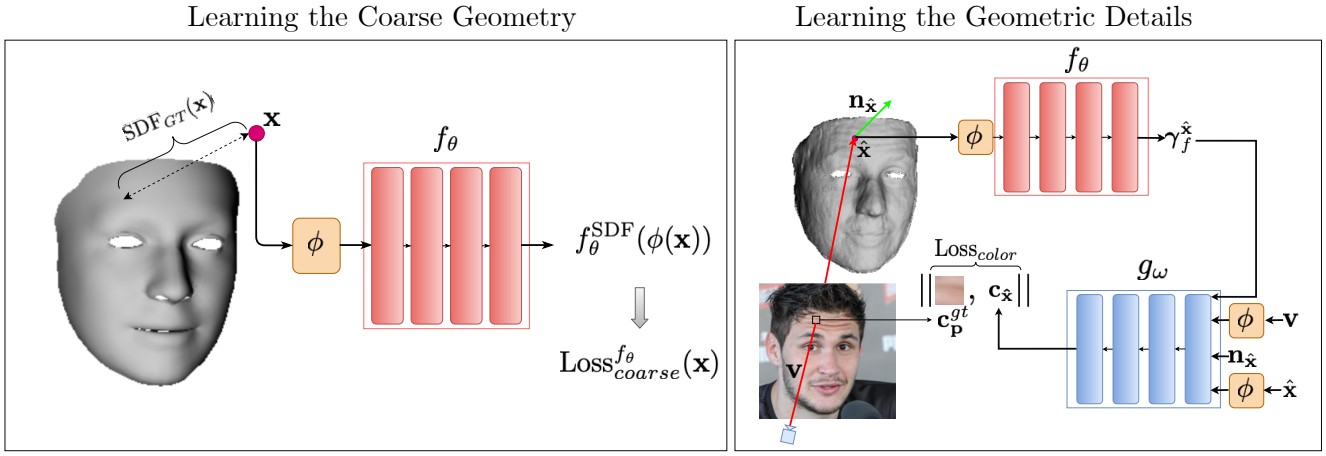

Figure 2: **Overview of SIDER: Left:** First SIDER learns a coarse geometry, represented as an SDF, by using FLAME mesh [15] fit to the input image as supervision **Right:** Next, it recovers facial geometric details by optimizing the SDF w.r.t. the photometric loss on the input image.

in-the-wild images.

Implicit Neural Representations. Representing 3D shapes implicitly with neural networks, more specifically using Multi-Layer Perceptrons (MLPs), has led to the development of methods that are able to reconstruct a large variety of 3D shape with rich details [3, 4, 9, 16, 17, 19, 20, 24, 25, 31]. In [20], authors train an SDF using instance specific meshes. Once trained, a latent space for the instances is learnt, making it possible to sample a large variety of shapes represented as SDFs. In [9], the authors propose a regularizer, the eikonal constraint, that allows learning SDFs from sparse 3D samples in the form of a point cloud. The authors of [31] use the eikonal constraint proposed by [9] to learn 3D shapes of objects using multi-view supervision along with object masks. They also propose an expression for the derivative of the intersection point of sphere tracing with respect to the MLP parameters that matches the real derivative up to the first order. SIDER uses the expression for the derivative of the intersection point proposed by [31], in order to back-propagate gradients to its geometry network and recover the facial geometric details. However, unlike the aforementioned methods, SIDER leverages a prior estimated from statistical models, which constrains the optimization and allows recovery of facial geometric details from single images without any multi-view supervision or ground truth 3D data.

3. SIDER

Given a single image, SIDER aims to extract from it facial geometric details, such as wrinkles and skin folds. SIDER uses a two-stage neural optimization approach to extract these details. In the first stage, we leverage the FLAME morphable model [15] prior, in order to learn the coarse geometry of the face, which is represented as an SDF. Next, we optimize this SDF w.r.t the photometric loss of the provided image, $\mathbf{I}$, in order to learn the facial geometric details. In the following, we elaborate the workings and the training process of SIDER.

3.1. Architecture

As shown in the overview Fig. 2, SIDER consists of two MLPs: a geometry network f_θ and a rendering network g_ω. The geometry MLP, $f_\theta(\cdot)$ represents the face shape (along with the facial geometric details) as an SDF. More specifically, for any point $\mathbf{x}$:

$$f_\theta(\mathbf{x}) = \{f_\theta^{\text{SDF}}(\phi(\mathbf{x})), \boldsymbol{\gamma}_f^{\mathbf{x}}\} \tag{1}$$

where $f_\theta^{\text{SDF}}(\mathbf{x})$ is the SDF predicted at $\mathbf{x}$, $\boldsymbol{\gamma}_f^{\mathbf{x}}$ is a feature vector predicted at $\mathbf{x}$ that is used as input to the rendering network g_ω, and ϕ is the positional encoding.

The rendering network, $g_\omega(\cdot)$, predicts the RGB value of a point $\mathbf{x}$ as follows:

$$g_\omega(\phi(\mathbf{x}), \mathbf{n}_\mathbf{x}, \phi(\mathbf{v}), \boldsymbol{\gamma}_f^{\mathbf{x}}) = \{R, G, B\} \tag{2}$$

where $\mathbf{x}$ are the point's coordinates, $\mathbf{n}_\mathbf{x}$ is the normal at point $\mathbf{x}$, $\mathbf{v}$ is the viewing direction and $\boldsymbol{\gamma}_f^{\mathbf{x}} = f_\theta(\mathbf{x})$ is a feature vector predicted by the geometry network f_θ at $\mathbf{x}$.

3.2. Learning the coarse geometry

Given a face image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, SIDER first learns a coarse geometry of the face using a morphable model prior. FLAME is fit to the face in $\mathbf{I}$ using standard landmark fitting [6, 10]

$$\min_{\boldsymbol{\alpha}_{\text{shape}}, \boldsymbol{\alpha}_{\text{exp}}, \boldsymbol{\alpha}_{\text{pose}}, \mathbf{cam}} ||L_{FLAME}^i - L_{gt}^i||; \quad \forall i \tag{3}$$

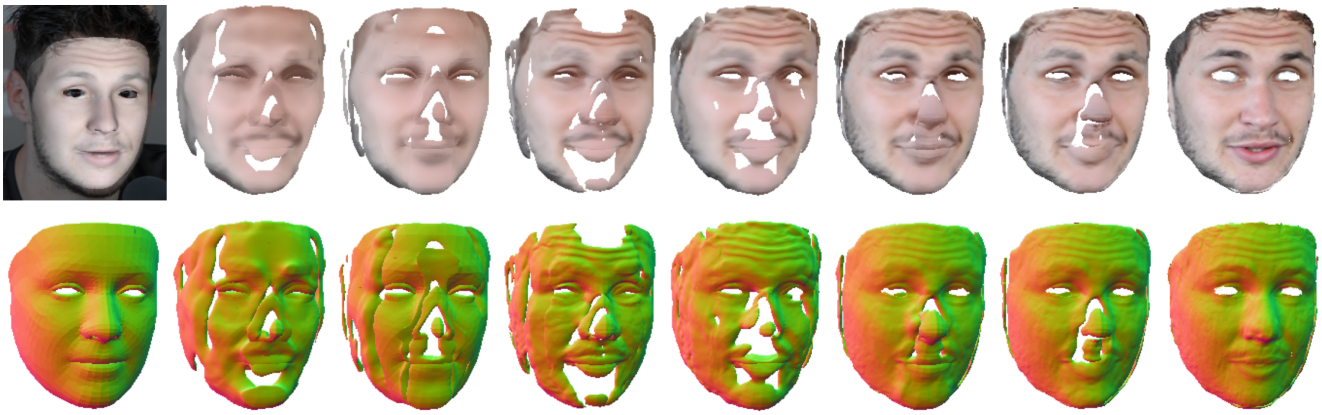

Figure 3: **SIDER in action:** Here we show the intermediate output of SIDER as it is trained. The first column contains an overlay of the initial FLAME [15] geometry on top of the image (top row) and the normals of the initial geometry. The subsequent columns show the rendered RGB (top row) and the normals of the learnt geometry (bottom row) as SIDER is trained. Once converged, we get a facial geometry with high quality details (last column, bottom row) and its photrealistic rendering (last column, top row).

where L^i_{FLAME} is the position of the i'th landmark of the FLAME model [15], L^i_{gt} is the position of the i'th landmark predicted by 3DDFA [10] (which we treat as ground truth), $\boldsymbol{\alpha}_{\text{shape}}, \boldsymbol{\alpha}_{\text{exp}}, \boldsymbol{\alpha}_{\text{pose}}$ are the shape, expression and pose parameters of the FLAME model [15] and **cam** are the camera parameters. Once FLAME [15] is fit to the image, we train an MLP, f_θ, to represent this coarse mesh as an SDF by minimizing the following:

$$\text{Loss}_{geo}(\mathbf{x}) = ||f^{\text{SDF}}_\theta(\phi(\mathbf{x})) - \text{SDF}_{GT}(\mathbf{x})||; \quad \forall \mathbf{x} \in \mathcal{P} \quad (4)$$

where $\mathcal{P}$ is a set of randomly chosen points in space in the neighborhood of the FLAME mesh, $\mathbf{x} = \{x, y, z\}$ is a point in space, ϕ is the positional encoding and $\text{SDF}_{GT}(\cdot)$ is the ground truth SDF to the coarse mesh. Since the FLAME face mesh is an open and single surface layer, an SDF cannot be defined on it directly (there is no region where the distance is negative, since there is no 'inside'). Therefore, in order to define the SDF, we consider the face mesh to be volume of 'thickness' $\boldsymbol{\epsilon}$. This allows us to define the ground-truth SDF as follows

$$\text{SDF}_{GT}(\mathbf{x}) = \text{Point2Mesh}(\mathbf{x}) - \frac{\boldsymbol{\epsilon}}{2} \quad (5)$$

where Point2Mesh is the point-to-mesh distance function and $\boldsymbol{\epsilon}$ is a small number denoting the thickness of the mesh. Additionally, the geometry network is regularized using the eikonal loss:

$$\text{Loss}_{eik}(\mathbf{x}) = \mathbb{E}_{\mathbf{x}}(||\nabla_{\mathbf{x}} f_\theta(\mathbf{x})|| - 1)^2 \quad (6)$$

The full loss on the geometry network is

$$\text{Loss}^{f_\theta}_{coarse}(\mathbf{x}) = \text{Loss}_{geo}(\mathbf{x}) + \lambda \text{Loss}_{eik}(\mathbf{x}) \quad (7)$$

where $\lambda = 1\text{e}^{-4}$ is a regularization coefficient.

3.3. Recovering the Facial Geometric Details

Once f_θ has been trained to approximate the coarse geometry, we use the provided image, $\mathbf{I}$, to fine-tune f_θ and recover the facial geometric details. In order to render the SDF f_θ, we use sphere-tracing along with a rendering network g_ω. Rays are shot into the scene from the camera center $\mathbf{o}$, and the intersection of these rays with the face mesh is estimated using sphere-tracing and implicit differentiation [31]. The colors of the intersecting points are predicted using g_ω. More specifically, consider a ray, $\mathbf{r} = \mathbf{o} + \mathbf{v}t$, with viewing direction $\mathbf{v}$, parametrization t, and surface intersection point $\hat{\mathbf{x}}$. The RGB color at $\hat{\mathbf{x}}$ is calculated as follows:

$$\mathbf{c}_{\hat{\mathbf{x}}} = g_\omega(\phi(\hat{\mathbf{x}}), \mathbf{n}_{\hat{\mathbf{x}}}, \phi(\mathbf{v}), \boldsymbol{\gamma}^{\hat{\mathbf{x}}}_f) \quad (8)$$

where $\mathbf{c}_{\hat{\mathbf{x}}}$ is the predicted RGB color at point $\hat{\mathbf{x}}$, $\mathbf{n}_{\hat{\mathbf{x}}}$ is the normal at point $\hat{\mathbf{x}}$ and $\boldsymbol{\gamma}^{\hat{\mathbf{x}}}_f$ is a feature vector predicted by the geometry network, f_θ at $\hat{\mathbf{x}}$. The facial geometric details are recovered by jointly optimizing the geometry network f_θ and the rendering network g_ω with respect to the photometric loss as follows:

$$\min_{\theta,\omega} \text{Loss}_{color}(\mathbf{c}_{\hat{\mathbf{x}}}, \mathbf{c}^{gt}_{\mathbf{p}}) = ||\mathbf{c}_{\hat{\mathbf{x}}} - \mathbf{c}^{gt}_{\mathbf{p}}|| \quad (9)$$

where $\mathbf{c}^{gt}_{\mathbf{p}}$ is the ground truth pixel color. The gradients to the geometry network f_θ, are calculated using implicit differentiation [31]. Additionally, in order to ensure that f_θ does not drift too far away from the face shape, it is regularized using the coarse SDF from Sect 3.2. The complete loss of the geometry network is:

$$\begin{aligned}\text{Loss}^{f_\theta}_{detail}(\mathbf{c}_{\hat{\mathbf{x}}}, \mathbf{c}^{gt}_{\mathbf{p}}) = &||\mathbf{c}_{\hat{\mathbf{x}}} - \mathbf{c}^{gt}_{\mathbf{p}}|| \\ &+ \lambda_1 ||f_\theta(\phi(\mathbf{x})) - \text{SDF}_{GT}(\mathbf{x})|| \\ &+ \lambda_2 \text{Loss}_{eik}\end{aligned} \quad (10)$$

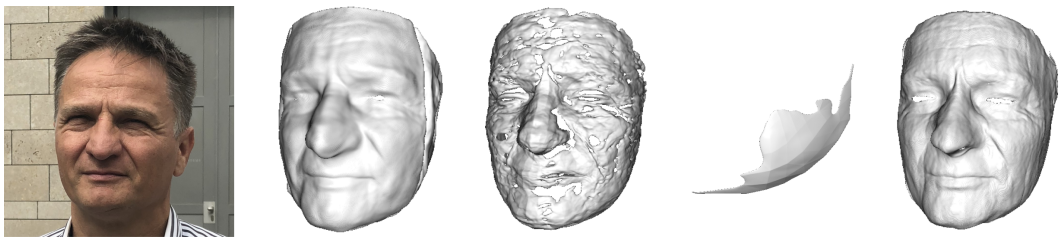

Figure 4: **Ablation Study**: Here we ablate the contributions of the normals and the feature vector as input to the gradient network (see Eq. (9)). The first column is the image on which SIDER is optimized. (a) Geometry learnt without using the feature vector as input. (b) Geometry learnt without using the normals as input, (c) Geometry learnt without using both feature vector and normals as input. The last column shows the result of SIDER that uses both the normals and the feature vector as input to the rendering network.

where $\lambda_1 = 1e^2, \lambda_2 = 1e^{-4}$ are the regularization coefficients.

4. Experiments

In this section, we evaluate SIDER's ability to recover facial geometric details from single in-the-wild images. We present both qualitative and quantitative results, and compare our proposed approach to the current state of the art.

4.1. Implementation Details

SIDER learns detailed facial geometry given only a single image. We perform extensive experiments on images from three publicly available datasets, namely FFHQ [13], ALFW2000 [34], and the NoW challenge dataset [22]. We choose challenging images that contain both subtle and more complicated facial geometric details, including wrinkles and skin folds. The images are resized to 256x256 resolution before fitting the FLAME model.

The geometry network, $f_\theta(\cdot)$, consists of 8 linear layers and a skip connection from the input to the fourth layer, similarly to [31]. The rendering network, $g_\omega(\cdot)$, consists of 4 linear layers. Its input is non-linearly mapped to learn high frequencies [18]. Each layer of both MLPs includes 512 hidden units. We first train the geometry network for 1000 epochs to learn the coarse geometry. Then, we fine-tune it via photometric optimization and jointly train the rendering network for around 200-300 epochs, until the loss is not decreasing further. We use Adam optimizer with a learning rate of 10^{-4}. It takes about 3 days on a single Titan RTX GPU for the method to converge.

4.2. Ablation Study

Gradients from the photometric loss of Eq. (9) are back-propagated into the geometry network f_θ via the feature vector, the normals and the intersection point $\hat{x}$ (see [31]). In this section, we ablate the contribution of the feature vector and the normals by alternatively removing one or the other, and finally both, from the input to the rendering network g_ω. Fig. 4 shows the results of the ablation study, the first image from the left is the input image on which SIDER was trained, (a) shows the results when the feature vector is removed from the input to g_ω, (b) shows the results when the normals are removed from the input, (c) shows the results when both are removed, and the last image shows the results of the full model. Without the feature vector as input (see Fig. 4 (a)) the geometry network does not recover any facial geometric details and the geometry is very smooth. In contrast, without the normals as input (see Fig. 4 (b)), the geometry network does seem to recover some details but the overall reconstruction suffers significantly. Without either the normals or the feature vector as input (see Fig. 4 (c)) learning completely fails. Using both the normals and the feature vector (see the last column of Fig. 4) as input to the rendering network, g_ω, allows SIDER to recover high quality facial geometric details without compromising overall reconstruction quality. The corresponding reconstruction errors (mm) for (a)/(b)/SIDER are: median 0.62/0.80/0.54, mean 0.73/0.87/0.69, std 0.60/0.59/0.58.

4.3. Quantitative Evaluation

We evaluate the accuracy of the detailed reconstruction of SIDER on the NoW challenge dataset [22]. We compare it with the performance of recent state-of-the-art methods, namely Pix2vertex [23], DF^{2}Net [32], and DECA [6]. We choose NoW for our quantitative evaluation, since it includes 3D ground truth scans and provides a standard evaluation protocol. It measures the distance from all reference scan vertices to the closest point in the reconstructed mesh surface, after rigidly aligning scans and reconstructions. Because Pix2vertex [23] and DF^{2}Net [32] reconstruct a smaller part of the face than the other methods, we ensure that the ground truth face scan is cropped to a circular area (same for all the methods) that would not exceed the smallest reconstructed mesh.

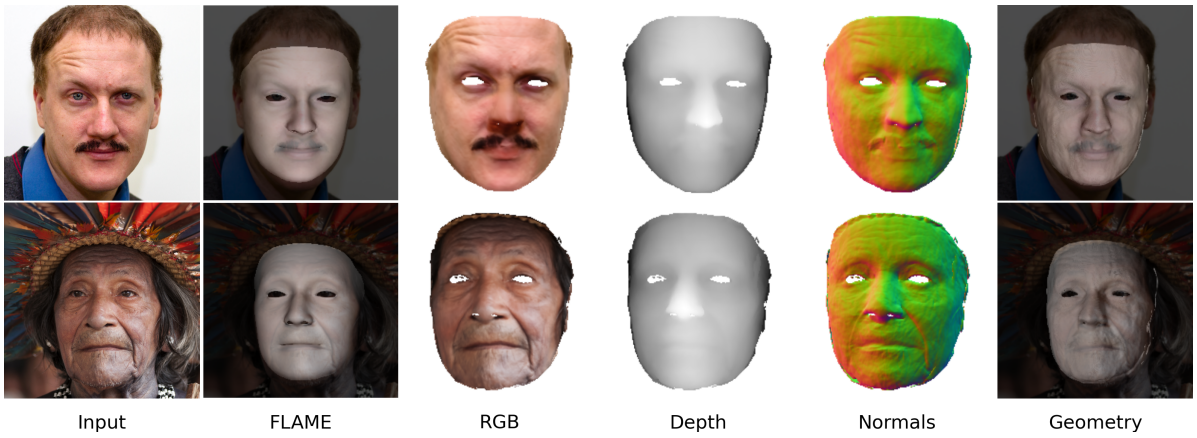

Figure 5: **Reconstructions by SIDER:** Above we show the results of SIDER on images from the FFHQ [13] dataset. The first column is the input image with respect to which SIDER is optimized. The second column is the overlay of coarse FLAME [15] mesh on the input image. The third column is the render of the detailed geoemtry learnt by SIDER. The fourth column is the depth. The fifth column are the normals of the deetailed face geometry learnt by SIDER and the last column is the overlay of the learnt detailed geometry on top of the input image.

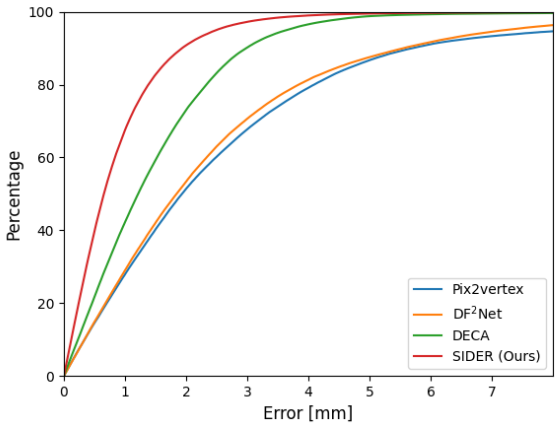

Figure 6: Cumulative error plot. Reconstruction error on the NoW validation set for the different methods.

As shown on Table 1, SIDER outperforms the current state-of-the-art, i.e. DECA [6], and other methods by a healthy margin. The mean and median error of SIDER is 0.89 mm and 0.66 mm respectively; both of which are less than that of DECA [6], which has a median error 1.19 mm and mean error of 1.47 mm. Additionally, the standard deviation of SIDER is lower than that of DECA [6], 0.88 mm as compared to 1.25 mm of DECA [6]. Similarly, the cumulative error plots in Fig. 6 demonstrate the significant advantage of SIDER compared to the other methods.

Method	Median (mm)	Mean (mm)	Std (mm)
Pix2vertex [23]	1.93	2.74	2.85
DF^{2}Net [32]	1.84	2.51	3.40
DECA [6]	1.19	1.47	1.25
SIDER (Ours)	**0.66**	**0.89**	**0.88**

Table 1: Quantitative results. Reconstruction error on the NoW validation set.

4.4. Qualitative Evaluation

We next provide an extensive qualitative evaluation of SIDER on in-the-wild images from datasets for which no ground truth is available.

In Fig. 5, we demonstrate qualitative results of SIDER on images from the FFHQ dataset [13]. The columns of the figure show correspondingly: the input image, the FLAME fitting, the learnt render (RGB output), the depth, the normals, and finally the overlay of the reconstructed face, with geometric details, on top of the original image. As can be seen, SIDER recovers accurately the geometric details for each input image. For example, in the first row of Fig. 5, the geometric details arising on the left side of the forehead due to the raised eyebrow are captured faithfully. Similarly, in the second row we see the wrinkles around the mouth are realistically recovered.

In Fig. 3, we show the RGB output and normals during the joint training of the geometry and rendering networks. We can see how the networks gradually learn a detailed facial geometry represented as an SDF.

In Fig. 7, we qualitatively compare the results of SIDER to state-of-the-art detailed facial reconstruction methods,

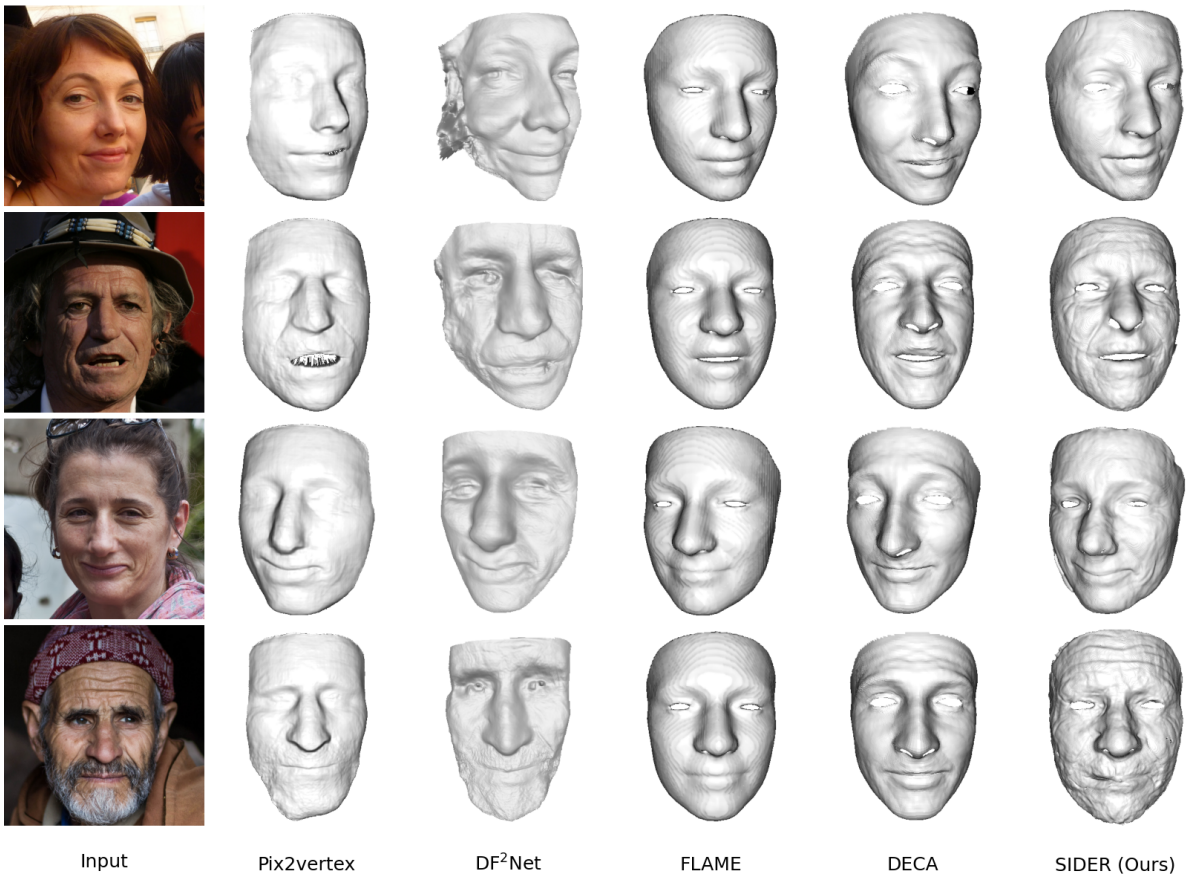

Figure 7: **Comparison to state-of-the-art methods:** Here we compare again prior art: Pix2vertex [23], DF^{2}Net [32], FLAME fitting [15] and DECA (w/ details) [6]. Input images taken from FFHQ.

namely Pix2vertex [23], DF^{2}Net [32], and DECA [6]. The first column corresponds to the input image, the second column contains the results of Pix2vertex [23], the third contains the results of DF^{2}Net [32], the fourth column contains the results of simple FLAME fitting [15] (used as ground truth of the coarse geometry during the first stage), the fifth column contains the results of DECA with details [6], the current state-of-the-art, and the last column contains the results of our method. Note that DECA results are masked with FLAME, in order to illustrate only the face region for fair comparison with the other methods. As can be seen, SIDER is able to recover significantly more detail than the other methods. Pix2vertex [23] generates detailed reconstructions that are quite smooth and misses large visible geometric details such at the skin fold on the right side of the lip of the image in row 1 or the wrinkles on the forehead of the image in row 4. The results of DF^{2}Net [32], while recovering greater geometric details than Pix2vertex [23], are still prone to errors. For example, it misses the wrinkles on the forehead of the image in row 2 and recovers incorrect details on the forehead from the image in row 4. The produced meshes also contain artifacts, e.g. on the right part of the head in rows 1 and 2, or include extreme curvature, e.g. in the eyes region. The face shape is also affected by lighting to a great extent (e.g. shaded part in row 2).

DECA [6], the current state-of-the-art in facial geometric details recovery, generates better reconstructions than both Pix2vertex [23] and DF^{2}Net [32], however it is unable to recover fine geometric details. For example, the skin fold on the right side of the mouth of the image in row 1 is captured by both our method, SIDER, and DF^{2}Net [32]. In contrast, DECA [6] is unable to recover it. Similarly, the geometric details under the right eye of the image in row 2 are not recovered by DECA [6]. SIDER, however, is able to accurately recover them. The skin folds around the eyes and around the lips of the image in row 3 are faithfully recovered by SIDER, but DECA [6] is unable to reconstruct them. In Fig. 8, we show a direct comparison of the facial geometric details generated by SIDER and DECA [6]. Insets in green show details generated by SIDER and insets in red show details generated by DECA [6]. For the image in row 1, we see that SIDER is faithfully able to recover the details on the forehead, around the eyes, and around the mouth. In contrast, the details recovered by DECA [6] are over-smoothed

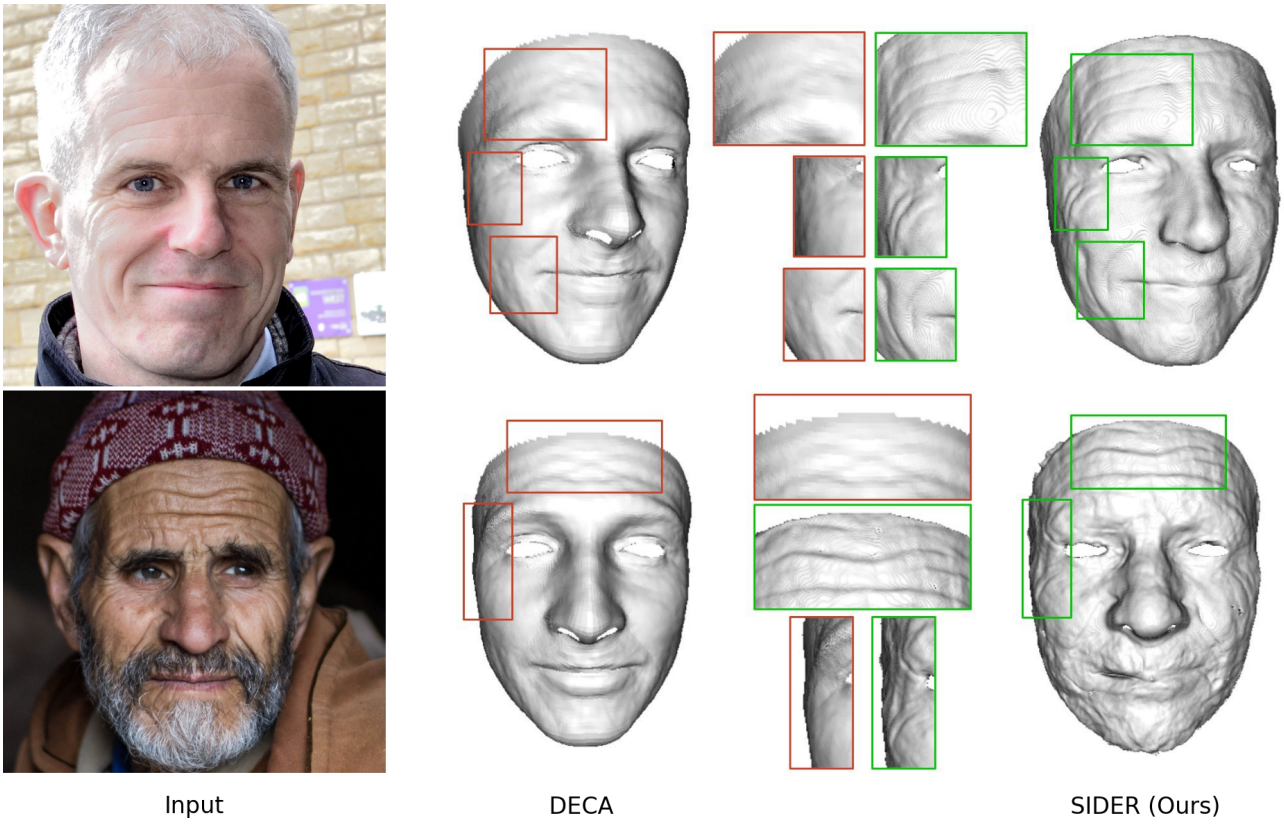

Figure 8: **Detailed comparison to DECA (w/ details) [6]:** Here we compare the facial geometric details generated by DECA (red insets) and SIDER (green insets). For the image in the top row, we see that SIDER faithfully recovers the details on the forehead, around the left eye and around the left corner of the lip. In contrast, DECA [6] smoothens the details on the forehead and fails to recover the details around the left eye and the left corner of the lip. Similarly, for the image in row 2, SIDER is able to accurately recover the details on the forehead and around the left eye, while DECA [6] fails to do so.

and inaccurate. Similarly, for the image in row 2, we see that the details recovered by SIDER have greater fidelity to the input than the details recovered by DECA [6].

In summary, as can be seen from the qualitative results in both Fig. 5, Fig. 8 and Fig. 7, SIDER is able to reconstruct high-quality facial geometric details from single images, more accurately and with greater fidelity to the input image than competing methods.

5. Conclusion

In this work we present SIDER, a method for high-fidelity detailed 3D face reconstruction from a single image that can be trained in an unsupervised manner. Our approach combines the best from classical statistical models and recent implicit neural representations. The former is used to obtain a coarse shape prior, and the latter provides high-frequency geometric detail, by only optimizing over a photometric loss computed w.r.t. the input image. A thorough quantitative and qualitative evaluation shows that SIDER outperforms current state-of-the-art by a significant margin. A limitation of our current approach is that it still cannot handle details like hair or beards and accessories such as glasses. This is because the photometric loss for these regions would require sub-pixel accuracy. In the future, we will explore alternatives for addressing this type of high-frequency details.

6. Acknowledgements

This work is partly supported by the Spanish government with the project MoHuCo PID2020-120049RB-I00 and María de Maeztu Seal of Excellence MDM-2016-0656. This work was also supported by a gift from Adobe, Partner University Fund 4DVision Project, and the SUNY2020 Infrastructure Transportation Security Center.

References

[1] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. 1999. 4321, 4322

[2] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and

Stefanos Zafeiriou. 3d face morphable models" in-the-wild". In *CVPR*, 2017. 4322

[3] Julian Chibane, A. Mir, and Gerard Pons-Moll. Neural unsigned distance fields for implicit function learning. In *NeurIPS*, 2020. 4322, 4323

[4] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, 2021. 4323

[5] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. End-to-end 3d face reconstruction with deep neural networks. In *CVPR*, 2017. 4322

[6] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. In *Proc. SIGGRAPH*, 2021. 4322, 4323, 4325, 4326, 4327, 4328

[7] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *CVPR*, 2018. 4322

[8] Markos Georgopoulos, Yannis Panagakis, and Maja Pantic. Investigating bias in deep face analysis: The kanface dataset and empirical study. In *arXiv*, 2020. 4322

[9] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*. 2020. 4322, 4323

[10] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 4323, 4324

[11] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *ICCV*, 2017. 4322

[12] Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 3d face reconstruction with geometry details from a single image. In *IEEE TIP*, 2018. 4322

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 4325, 4326

[14] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inversefacenet: Deep monocular inverse face rendering. In *CVPR*, 2018. 4322

[15] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. In *Proc. SIGGRAPH Asia*, 2017. 4322, 4323, 4324, 4326, 4327

[16] Shichen Liu, S. Saito, Weikai Chen, and H. Li. Learning to infer implicit surfaces without 3d supervision. In *NeurIPS*, 2019. 4323

[17] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 4322, 4323

[18] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 4322, 4325

[19] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020. 4322, 4323

[20] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, June 2019. 4323

[21] Elad Richardson, Matan Sela, Roy Or-El, and R. Kimmel. Face reconstruction from a single image. In *CVPR*, 2017. 4322

[22] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *CVPR*, 2019. 4325

[23] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *ICCV*, 2017. 4322, 4325, 4326, 4327

[24] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 4322, 4323

[25] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. In *CVPR*, 2021. 4323

[26] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Fml: Face model learning from videos. In *CVPR*, 2019. 4322

[27] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, 2018. 4322

[28] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019. 4322

[29] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, 2018. 4322

[30] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *CVPR*, 2018. 4322

[31] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *NeurIPS*, 2020. 4322, 4323, 4324, 4325

[32] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *ICCV*, 2019. 4322, 4325, 4326, 4327

[33] Wenbin Zhu, HsiangTao Wu, Zeyu Chen, Noranart Vesdapunt, and Baoyuan Wang. Reda:reinforced differentiable attribute for 3d face reconstruction. In *CVPR*, 2020. 4322

[34] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, 2015. 4325