Disjoint Pose and Shape for 3D Face Reconstruction

Raja Kumar*   Jiahao Luo*   Alex Pang   James Davis

University of California Santa Cruz
{rkumar38, jluo53, pang, davisje}@ucsc.edu

Abstract

Existing methods for 3D face reconstruction from a few casually captured images employ deep learning based models along with a 3D Morphable Model(3DMM) as face geometry prior. Structure From Motion(SFM), followed by Multi-View Stereo (MVS), on the other hand, uses dozens of high-resolution images to reconstruct accurate 3D faces. However, it produces noisy and stretched-out results with only two views available. In this paper, taking inspiration from both these methods, we propose an end-to-end pipeline that disjointly solves for pose and shape to make the optimization stable and accurate. We use a face shape prior to estimate face pose and use stereo matching followed by a 3DMM to solve for the shape. The proposed method achieves end-to-end topological consistency, enables iterative face pose refinement procedure, and show remarkable improvement on both quantitative and qualitative results over existing state-of-the-art methods.

1. Introduction

3D face models find a wide range of applications in scenarios such as 3D avatars [53], biometric identification [9], photo editing [49] and film production [10]. Traditionally, specialized setups and hardware [15, 10] have been used to generate high-fidelity 3D faces. However, reconstructing accurate 3D faces from a few casually captured uncalibrated images remains a challenging problem.

Recent learning based methods use a Deep Neural Network (DNN) model to reconstruct 3D faces. Most often, a 3D Morphable Model (3DMM) is employed to represent the face shape using a vector space representation that is learned from a linear combination of principle components of a collection of face scans. These methods aim to recover the 3DMM parameters from given facial images, often by analysis-by-synthesis optimization or by employing multi-

---
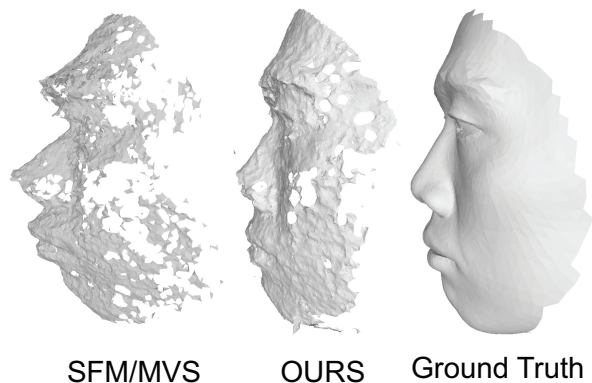*authors contributed equally

SFM/MVS   OURS   Ground Truth

Figure 1. Comparison of raw stereo output from SFM/MVS [37] [36] (left), our method (middle), and ground truth (right). Clearly, SFM/MVS produces a stretched-out face with more noise due to under-constrained optimization with only 2-views. The proposed face shape prior solves this problem by providing a strong prior.

view constraints. However, it requires a complicated, non-linear optimization that has difficulty converging in practice.

In contrast, the highest quality 3D faces are generated using multi-view stereo methods with dozens of calibrated high-resolution images captured in a laboratory setting. Unfortunately pre-calibrated camera pose is not readily available in a casual capture setting. Recent learning-based methods consisting of Structure From Motion (SFM) followed by Multi-View Stereo (MVS) have been proposed for 3D reconstruction from uncalibrated sets of images. However, solving for camera poses using this method when only a few views are available are under-constrained and results in inaccurate estimate as shown in Figure 1 above.

The key observation in this paper is that when only two views are available jointly optimizing for pose parameters, shape parameters, and correspondence between image and model is theoretically optimal, but results in sub-optimal results in practice. Instead we propose a disjoint solution in which pose and shape are solved separately. This allows
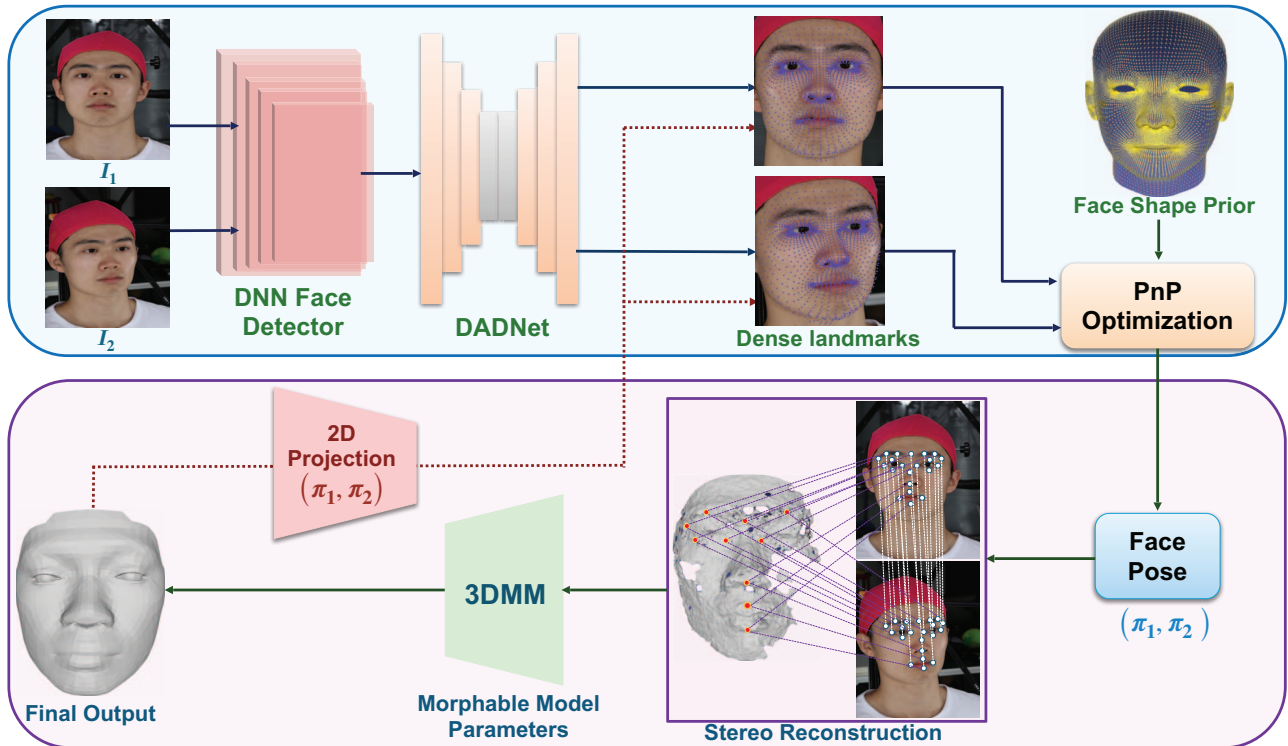
Figure 2. An overview of the proposed method. In the proposed method, we solve for pose and shape disjointly for an accurate reconstruction. The blue box shows the proposed face pose estimation using face shape prior method, and pink box shows the 3D face reconstruction pipeline. π_1 and π_2 represent the projection matrices.

appropriate regularization priors to be used in each stage, allowing a more stable overall solution.

Our end-to-end pipeline consists of three stages, face pose estimation using a face shape prior, 3D reconstruction using stereo matching, and iterative camera pose refinement. In the first stage, given two views, we detect topologically consistent dense 2D landmarks for both views and use a strongly constrained 3D face shape prior in the same topology as the dense 2D landmarks. Given 2D landmarks and the 3D face shape prior, we solve for camera poses using Quadratically Constrained Quadratic Program (QCQP) based optimization proposed in Terzakis *et al.* [40]. In the second stage of our pipeline, we perform stereo matching to generate a 3D point cloud. This step uses no face shape prior in order to allow a full range of shape variation. This is followed by a fit to a modern 3DMM model called FLAME to fill in the missing region. The 3DMM acts as a face prior on the raw 3D point cloud, however it allows a great deal of variation in shape and is thus much less strongly constrained than the face prior used to find pose.

Since our output accuracy is determined in part by the estimate of pose derived from 2D landmarks, we perform an iterative refinement on these. The 2D landmarks are refined using the current estimates of pose and shape. This refinement converges in only a few iterations. Our analysis

on FaceScape and Stirling datasets show that the proposed pipeline outperforms the state-of-the-art multi-view methods in quantitative as well as qualitative comparison.

The contribution of this paper is an end-to-end pipeline for 3D face reconstruction by solving pose and shape disjointly using two uncalibrated images, achieving state-of-the-art accuracy.

2. Related Work

2.1. 3D Morphable Face Model

3DMM is a statistical model which transforms the shape and texture into a vector space representation that is derived from hundreds of 3D face scans. It was first proposed by Blanz *et al.* [8], and numerous variations such as [33, 29] were introduced to include identity, expression, and pose factors. This allowed for separate control of the model, and using these attributes, it can be transformed into a set of blendshapes [28] which can then be rigged to create unique animations for each individual. More recent deep learning based methods [3, 42, 44, 43] have been proposed to enhance the representation power of 3DMM. We recommend referring to the recent survey [18] for a comprehensive review. In this paper, we use the FLAME model [29] for its simplicity and wide applicability.
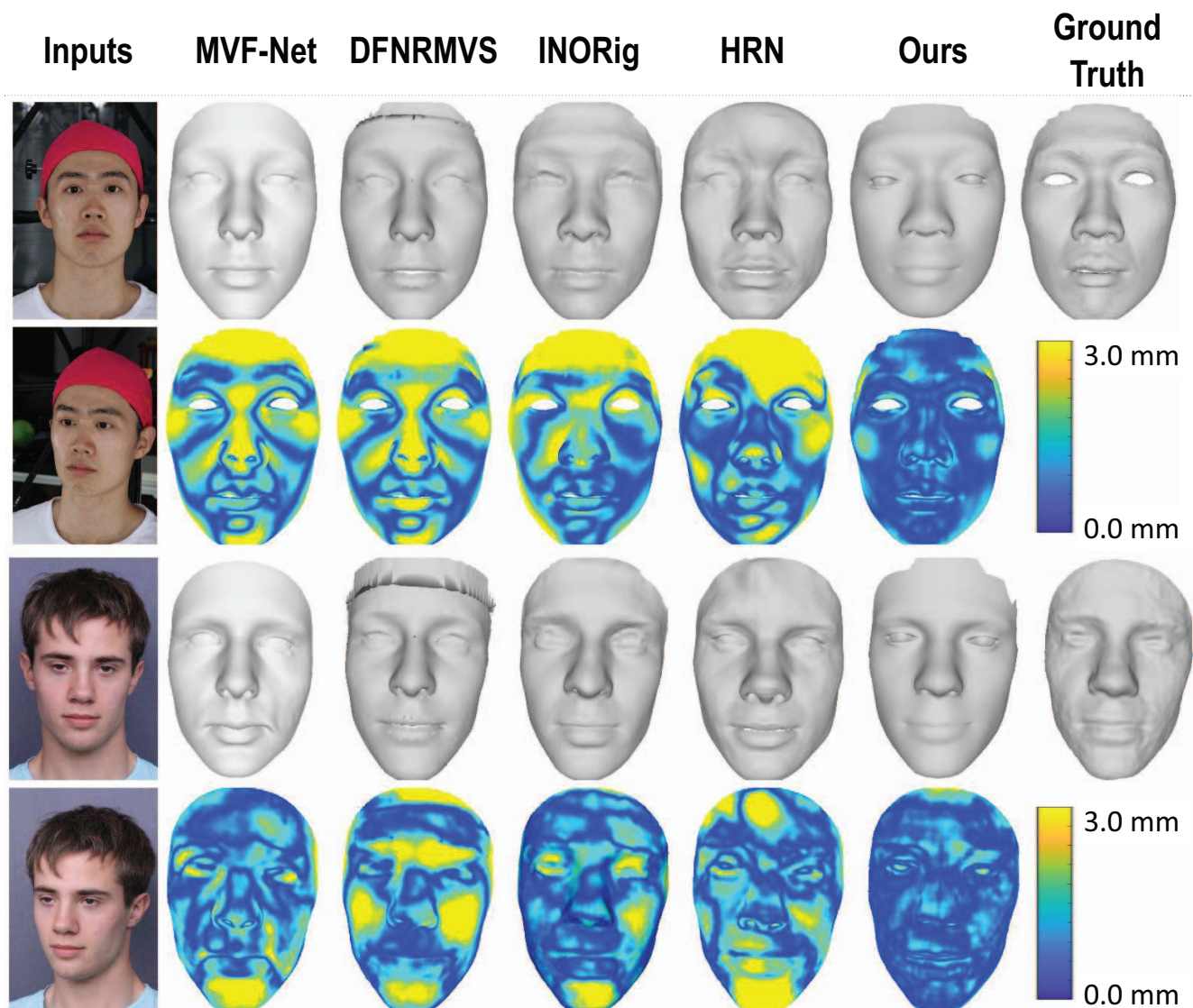
Figure 3. A qualitative comparison of output rendering and error map to the SOTA methods. The first and second rows show the output and error map on a sample input from the FaceScape dataset, while the third and fourth show the output and error map on a sample of the Stirling dataset. The error maps visualize a range from blue (0 mm) to yellow (3 mm). It is evident that the proposed method produces a smooth reconstruction with lower error.

2.2. Multi-View Stereo

Traditional MVS methods [20, 25, 27] estimate the 3D shape of the object from a set of calibrated multi-view images. These methods perform feature matching, followed by triangulation, and then the application of a depth map fusion algorithm to obtain 3D meshes [20, 11, 37, 41]. Parts of this pipeline that solve for specific missing information or improve the speed and accuracy of reconstructions have also been proposed. Depth fusion methods have been used to produce a valid 3D watertight surface [14, 32, 24]. Multiple stereo-based methods [46, 7] have been proposed specifically for faces. Valgaerts *et al.* [46] proposed a

lightweight passive facial performance capture approach that uses image-based scene flow computation, lighting estimation, and shading-based refinement algorithms to reconstruct the face. Beeler *et al.* [7] used a combination of smoothness, ordering, and uniqueness constraints to recover facial shape robustly. However, these methods generally require pre-calibrated cameras or many images. Our work focuses on face reconstruction from two uncalibrated images.

2.3. Deep Face Reconstruction

More recent learning-based methods [23, 13, 51] train DNN with different objective functions to improve 3D face reconstruction. Other methods [17, 34, 5] proposes to re-

cover facial geometry from uncalibrated images or videos. Duo *et al.* [17] uses DNN to disentangle facial and identity features and uses Recurrent Neural Network (RNN) to perform subspace representation of 3D shapes. Ramon *et al.* [34] uses a Siamese network to extract features from multi-views. Wu *et al.* [48] incorporates multi-view geometric constraints into the network by establishing dense correspondences between different views leveraging a novel self-supervised view alignment loss. Though these methods achieve good results, they require multiple views, and performance is sub-optimal when only two views are available.

3D face reconstruction from a single view has also been studied and is challenging due to its ill-posed nature. Generally, a deep neural network is trained to regress 3DMM model parameters [45, 52] reconstruct 3D geometries [22, 29, 35] or render image using analysis-by-synthesis [16, 21, 19]. Prior work [30, 48] has shown that multi-view face methods generally perform better than single-view. In this paper, we compare directly with several state of the art multi-view reconstruction methods.

3. Proposed Method

The proposed method consists of three stages, as shown in Figure 2. In the first stage, we estimate the face poses using topologically consistent dense 2D landmarks and a 3D face shape prior. In the second stage, an accurate 3D face is reconstructed using two-view stereo matching, which utilizes the face pose obtained in the first stage. This is followed by 3DMM model fitting to fill in the missing region. Finally, we perform Face Pose Refinement (FPR) by iteratively projecting the 3D face to image space using the face pose estimated in the previous iteration. We explain each stage in detail in the following sections.

3.1. Pose Estimation Using Face Shape Prior

In the first stage of our pipeline, we use topologically consistent dense 2D landmarks and a 3D face shape prior for face pose estimation. In this section, first we provide details about dense 2D landmark detection and face shape prior computation, followed by an explaination of the face pose estimation method.

Facial landmark detection is a well-explored field of research. However, most of the existing methods focus on a set of frequently used 68 landmarks. Martyniuk *et al.* [31] and Wood *et al.* [47] proposes dense landmark detection methods with 10x more landmark points. We first use a face detector [6] to find the facial region since landmark detection often fails to find accurate landmarks without it. DADNet [31] is used to find dense landmarks.

We use a face shape prior as a regularizer to solve for face poses. We use the mean face with no variability to provide a very strong prior. Given a set of m training set faces $\mathcal{F} = \{\mathbf{F_1}, \mathbf{F_2}, ..., \mathbf{F_m}\}$, we compute the face shape prior

$\mathbf{F_p}$ as the mean of all the face vertices in correspondence

$$\mathbf{F_p}^k = \frac{1}{m} \sum_{i=1}^{m} \mathbf{F_i}^k \qquad \forall k \in \{1, 2, ..., n\} \qquad (1)$$

where, $\mathbf{F_p}^k$ and $\mathbf{F_i}^k$ denote the kth vertex of face prior $\mathbf{F_p}$ and the ith face $\mathbf{F_i}$ respectively.

Algorithm 1 Iterative FPR and 3D Reconstruction

Input: Two view images (I_1, I_2), face shape prior (F_p) computed as explained in Section 3.1
Output: Reconstructed 3D face and Face Poses

1. **Init:** $L_1 = LD(I_1)$, $L_2 = LD(I_2)$, where, $L_1, L_2 \in \mathbb{R}^{Nx2}$ and LD is Landmark Detector

2. **for** $iter = 0, 1, \ldots$,till convergence

    (a) $(R_j, t_j) = argmin(R, t) \sum_{i=1}^{n} ||p_i - K(Rv_i + t)||^2 \quad \forall p_i \in L_j, v_i \in F_p \, and \, j \in \{1, 2\}$

    (b) $F_{sparse} = Stereo(I_j, R_j, t_j) \qquad \forall j \in \{1, 2\}$

    (c) $F_{flame} = 3DMM(F_{sparse})$

    (d) $L_j = K(R_j F_{flame} + t_j) \qquad \forall j \in \{1, 2\}$
    // 2D projection of 3D face

    (e) Next iteration with update landmarks (L_j)

Face pose estimation requires solving for the rotation $\mathbf{R} \in \mathrm{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$ for an image I. However, solving for face poses using only two views when the face shape is not yet known is a highly under-constraint problem and has shallow minima leading to many solutions. Hence, to constrain our hypothesis space, we use the face shape prior. Formally, Given an image I, we find the topologically consistent landmark points $\mathbf{L} = \{\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_n}\} \in \mathbb{R}^{nx2}$ and face shape prior $\mathbf{F_p} = \{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_n}\} \in \mathbb{R}^{nx3}$ and we seek to find the rotation $\mathbf{R}$ and $\mathbf{t}$ minimizing the cumulative squared projection error

$$S(\mathbf{R,t}) = \sum_{i=1}^{n} ||\mathbf{p_i} - \mathbf{K} \cdot (\mathbf{R} \cdot \mathbf{v_i} + \mathbf{t})||^2 = \sum_{i=1}^{n} ||\mathbf{p_i} - \pi \cdot v_i||^2$$
$$(2)$$

Where, $\mathbf{R}, \mathbf{t}$ are the rotation and translation parameters to be estimated, $\mathbf{K}$ is the scaling factor using weak perspective projection, π is the projection matrix and $\mathbf{p_i}$ and $\mathbf{v_i}$ represent the ith points in the image and 3D face prior space respectively. This is a well-known Perspective-n-Point (PnP) optimization problem, and we use a Quadratically Constrained Quadratic Program (QCQP) based implementation proposed in Terzakis *et al.* [40], to solve this optimization problem. Figure 2 (blue box) shows our face pose estimation pipeline.
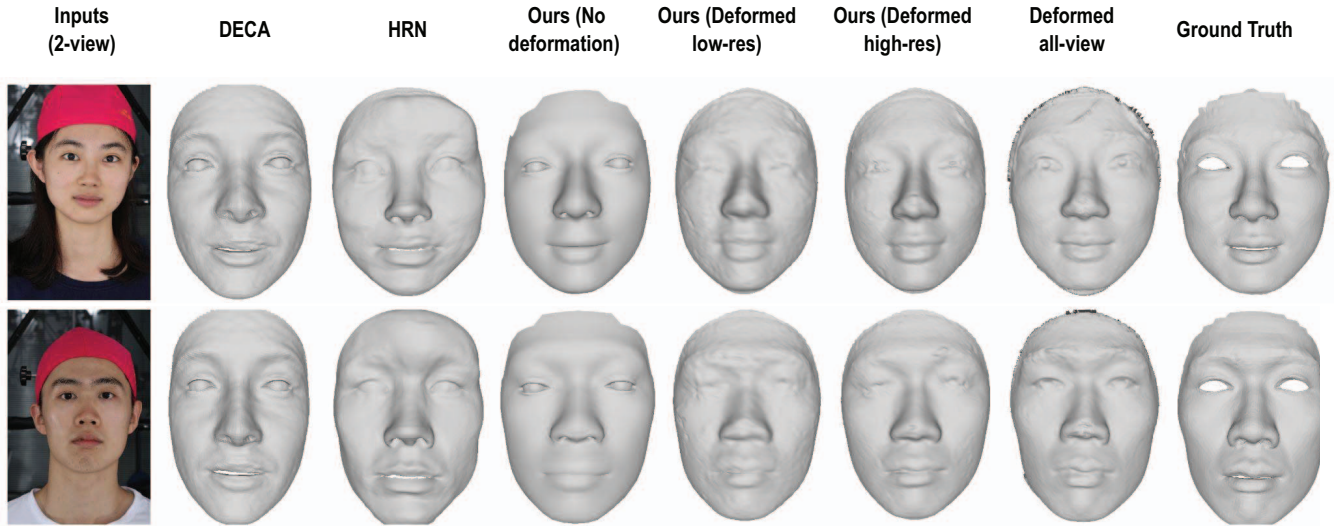
Figure 4. Our dense intermediate stereo output can be used to generate shape refinement beyond 3DMM. A qualitative comparison of deformed rendering to SOTA methods on FaceScape [50] dataset. (from left to right) DECA [19], HRN [26], ours without deformation, our deformed outputs, deformed outputs with high resolution inputs, deformed outputs with 50-view inputs, and the ground truth

3.2. 3D Face Reconstruction

Given two images and the associated face poses estimated using the method described in the previous section, we use the classical stereo method for 3D reconstruction. We do not employ any face shape prior during stereo matching. This allows maximum shape variability, however it produces noisy results and leaves gaps in areas not visible to both cameras.

To obtain a complete 3D facial model, a statistical blendshape 3D face model called FLAME is utilized as the geometry prior. This model consists of the head, neck, eyeball, and shoulder region and was developed from more than 33,000 scans. Despite its low dimensionality, the FLAME model is considered more expressive than other models, such as the FaceWarehouse [12] and Basel Face Model [29]. The model's parameters include both shape and expression, and this prior allows significantly more variation while solving for shape than the static prior used when solving for pose. The authors' original implementation of FLAME is used to determine the best-fitting parameters for the available data.

3.3. Iterative Face Pose Refinement

In the proposed pipeline, we find the 2D landmarks for two views independently using DADNet for initialization. However, the two views constraints introduced using stereo can help improve these landmarks, and consequently, the face poses. To this end, we perform an iterative Face Pose Refinement (FPR). Once we obtain the 3D face and face pose in the first iteration, we perform 2D projection and pick the corresponding landmarks. This is possible because
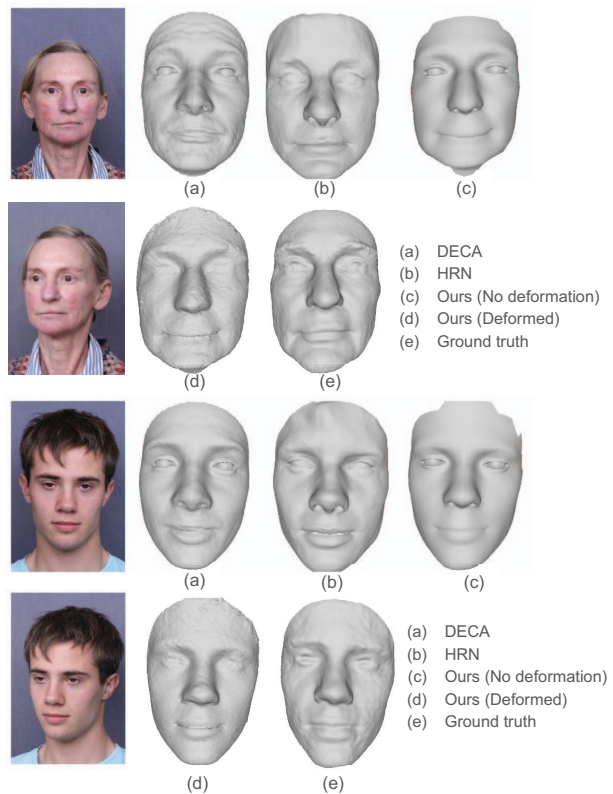

Figure 5. A qualitative comparison of deformed rendering to SOTA methods on Stirling [1] dataset. Our method can recover mid-frequency details which matches with groudtruth.

the landmarks of our 2D detector are topologically consistent with the vertices in our morphable model. More for-
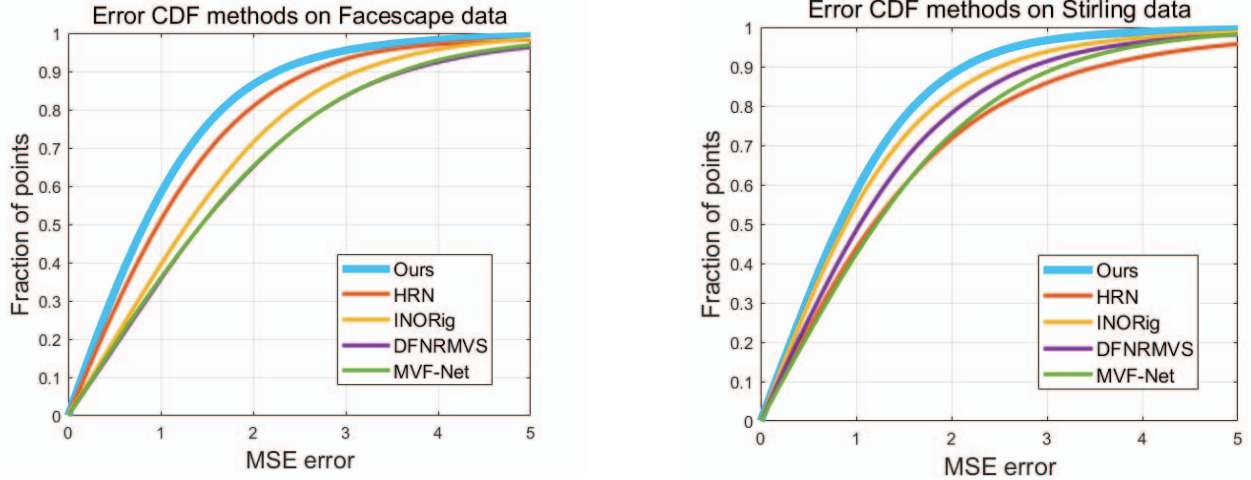
Figure 6. Comparison of performance plotted as a CDF with error on the x-axis and the fraction of reconstructed points with the error below this bound on the y-axis.Our proposed method has accuracy substantially better than the existing methods. A horizontal line through the plot at 0.9 is identical to the M90 error metric reported in results.

mally, given the generated face $\mathbf{F_g} = \{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_n}\}$ and face pose $(\mathbf{R}, \mathbf{t})$, we find the 2D projection as

$$\mathbf{L}^{'} = \mathbf{K} \cdot (\mathbf{R} \cdot \mathbf{F_g} + \mathbf{t}) = \pi \cdot \mathbf{F_g} \qquad (3)$$

where $\mathbf{L}^{'}$ is the new set of dense landmarks used for the next iteration of face pose estimation. This converges in just a few iterations. Algorithm 1 shows our FPR method in detail.

4. Results

In this section, we first discuss the datasets, evaluation metrics, and implementation details for conducting the experiments. We then perform a qualitative and quantitative comparison to several State-Of-The-Art(SOTA) methods to show the effectiveness of our method. Finally, we provide an ablation study on different components to validate the proposed method.

4.1. Implementation Details

We perform our experiments on two widely used face datasets, FaceScape [50] and Stirling dataset [1]. The FaceScape dataset contains high-resolution scans of hundreds of identities, each with more than 50 high-resolution images captured by DSLR cameras. We downsample them to a lower resolution (512x341) for a fair comparison. We use 40 randomly selected individuals as our testset. The Stirling dataset contains high-resolution scans of more than 100 people, each with four views. We follow prior work [4] and use 31 individuals as our testset.

For error analysis, the predicted meshes are aligned to the ground truth using the Iterative Closest Point (ICP) algorithm. For each point on the ground truth scan, we calcu-

late the point-to-face distance in millimeters by finding the closest triangle in the predicted mesh. From this set of distances, we calculate summary statistics like mean-squared error (MSE), Median and a robust approximation of maximum error which discards 10% of high error points as outliers (M90). For the face shape prior used during pose estimation, we use 80% of the training set scans, none of which are a part of our test set. During the 3DMM fitting, we crop the face to include only the frontal region.

4.2. Comparison to the State-Of-The-Art

4.2.1 Qualitative Comparison

In this section, we perform a qualitative comparison of the proposed method with four SOTA methods: MVF-Net [48], DFNRMVS [5], INORig [4] and HRN [26] on FaceScape [50] and Stirling [1] datasets. We show the rendering of example 3D faces generated using different methods along with their error map distribution over the face in Figure 3. HRN [26] uses a geometry disentanglement and introduce the hierarchical representation to fulfill detailed face modeling. Although, it captures the high frequency details, the deformation introduces error near head, chin and cheekbones regions as evident from error map in Figure 3. MVF-Net [48] and DFNRMVS [5] trains convolutional networks to explicitly enforce multi-view appearance consistency and learns the pose and shape jointly. In contrast, we solve for the pose and shape disjointly, which enables our pipeline to enforce multi-view consistency using accurate multi-view stereo. The qualitative results shown in Figure 3 validates our claim. The first and second rows of Figure 3 show the output and error map on sample data from FaceScape, while the third and fourth rows show the same

Methods	FaceScape			Stirling		
	MSE	**Median**	**M90**	**MSE**	**Median**	**M90**
MVF-Net [48]	1.75	1.45	3.60	1.52	1.22	3.14
DFNRMVS [5]	1.79	1.45	3.64	1.36	1.04	2.79
INORig [4]	1.54	1.29	3.09	1.20	0.90	2.50
HRN [26]	1.31	0.97	2.61	1.69	1.18	3.53
Ours	**1.07**	**0.83**	**2.27**	**1.04**	**0.83**	**2.07**

Table 1. Quantitative comparison of face shape estimation with SOTA methods on FaceScape and Stirling datasets. MSE, Median, and max error after rejecting 10% outliers (M90) are provided over each test set. Our method clearly outperforms the existing methods on both FaceScape and Stirling datasets.

on a sample from the Stirling dataset. As evident from the error maps, our output face approximates the ground truth more closely than the other methods.

The integration of dense stereo reconstruction within our method enables the implementation of further non-rigid deformation. This particularly generates shape refinement beyond overly smooth FLAME face space. We upsample the FLAME face and perform the as-rigid-as-possible deformation [2] followed by nonrigid ICP [38]. We perform Taubin smoothing after, since 2-view stereo is still noisy [39]. We compare our qualitative results with DECA [19] and HRN [26] as these methods propose to recover higher frequency features beyond 3DMM models. As shown in the figure 4 and figure 5, we are able to recover mid-frequency features which looks visually closer to the ground truth.

In contrast to DNN which often downsample input images, stereo techniques excel in extracting features from high-resolution data. Therefore, in Figure 4 we show the scenario while we intentionally use two high resolution inputs. Stereo is generally used at a multi-view setting, so we also show where we increase the number of images (50+) to serve as upper-bound. Note that these two are already similar, we thus conclude that a mid-frequency detailed 3D face reconstruction can be achieved with our method.

In order to demonstrate the problem that arises when attempting to solve pose and shape jointly, we perform a qualitative comparison of our proposed method with a SFM/MVS method [36, 37] in Figure 1. An ambiguity exists between pose and shape, resulting in a stretched-out face when the optimization converges to a low error solution that is incorrect. Our disjoint pose and shape pipeline allows the use of a strong prior on shape while solving pose, and a known pose while solving shape. This results in finding the correct solution that closely matches the ground truth shape.

4.2.2 Quantitative Comparison

In this section, we present a qualitative comparison of the proposed method with SOTA multi-view methods. Table 1 shows the MSE, Median and M90 error numbers on the FaceScape testset. It can be seen that the proposed method significantly outperforms SOTA methods on all three error

metrics, achieving an MSE error of as low as 1.07mm.

Next, We provide the comparison on the Stirling dataset in Table 1 (right half). Although Stirling images have lower resolution and higher compression than FaceScape images, the proposed method outperforms SOTA methods, demonstrating the effectiveness of the proposed method on different datasets. Also, notice that our method outperforms DFNRMVS [5] and INORig [4], which use the Stirling data as their training set.

We have also evaluated the error as a cumulative distribution function, as shown in Figure 6, which provides a more complete summary than a single aggregate metric. A horizontal line through the plot at 0.9 is identical to the M90 error metric reported above. The proposed method exhibits a lower error for all fractions of points on the FaceScape dataset (Figure 6 left). The Stirling dataset (Figure 6 right) plot indicates that even with lower image quality, the proposed method outperforms other methods.

As mentioned in qualitative comparison, our intermediate stereo output can be used to deform the 3DMM output to recover detailed features. In table 2 we show the error before and after deformation. While the qualitative results appear more visually appealing, we note a modest enhancement in the error for FaceScape dataset.

Stirling	MSE	Median	M90
Ours (No Deformation)	1.04	0.83	2.07
Ours (Deformed)	0.97	0.78	1.98
FaceScape	**MSE**	**Median**	**M90**
Ours (No Deformation)	1.07	0.83	2.27
Ours (Deformed)	1.06	0.82	2.26

Table 2. Comparison after performing the deformation to 3DMM output using stereo output

The proposed method can easily be generalized to 3 and more views. We show the reconstruction error for 3-views in Table 3. It can be seen that the proposed method outperforms the existing methods even on 3-views.

4.3. Ablation Study

In this section, we provide an ablation study of different components of the proposed pipeline. In the pose estima-

	FaceScape (3-views)		
Methods	MSE	Median	M90
MVF-Net [48]	1.75	1.45	3.60
DFNRMVS [5]	1.73	1.40	3.56
INORig [4]	1.46	1.18	3.05
HRN [26]	1.28	0.96	2.54
Ours	**0.97**	**0.80**	**2.01**

Table 3. Quantitative comparison using 3-views input with SOTA methods on FaceScape datasets.

tion stage of the pipeline, results are affected by both 2D dense landmark inaccuracies and face prior inaccuracies. Mistakes made estimating pose will propagate to errors in shape reconstruction. Therefore, we perform an analysis of how the reconstruction accuracy is impacted by changes in these two factors.

In the first experiment, we aim to find the upper bound of the proposed method if we improve our face shape prior. We replace the mean face prior with an ideal prior, the ground truth 3D scan, when estimating the face pose. It can be observed in Table 4 that MSE error improves from 1.07 to 1.05mm. However, this is a very small improvement and we conclude that a mean face prior, though simple, is a sufficient prior while estimating pose.

Next, we perform an analysis to find the upper bound reconstruction accuracy with improvement in 2D landmarks. For this, we use ground truth 2D landmark to find the face pose. In Table 4, the first row shows the MSE using the proposed method without FPR, while the last row shows when ground truth landmarks are used, again without FPR. A significant improvement of ~ 0.3mm (from 1.16 to 0.87) is observed when landmark localization is improved. Based on this observation, we introduced the iterative FPR portion of our pipeline to improve landmarks. As shown in Table 4, second row, the proposed FPR method does improve accuracy, reducing error to 1.07. However, there is still a scope for improvement.

If both ground truth 3D face and ground truth 2D landmarks exist it is possible to solve for pose exactly. This is equivalent to pre-calibrated cameras. In this case the remaining error is due to stereo reconstruction and morphable model fitting. However this error is substantially lower at 0.77mm. We conclude that although pose has many fewer parameters than shape, finding correct pose is critical to future progress in 3D face reconstruction.

Finally, we also perform an analysis of how the reconstruction error improves after each iteration of the proposed FPR process. In Figure 7, we show the MSE for the first three iterations on a sample input. It can be observed that the proposed FPR improves the MSE from 1.16 mm to 1.07 mm. The iterative FPR converges in just three iterations due to good initialization and an end-to-end topologically consistent pipeline.

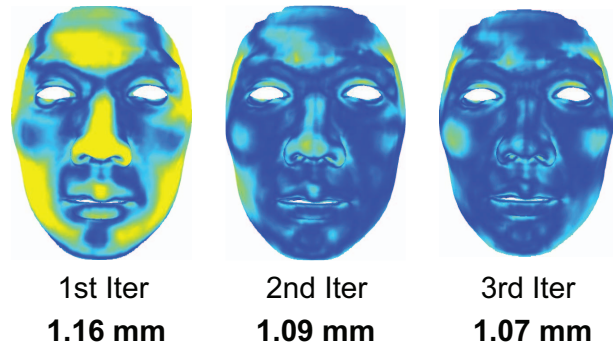|	1st Iter	2nd Iter	3rd Iter
	1.16 mm	**1.09 mm**	**1.07 mm**

Figure 7. (Top) error map for a sample input and (Bottom) average error on FaceScape dataset after each iteration of FPR. A clear improvement in reconstruction performance is observed using the proposed FPR.

	Mean Face	GT Face
Ours (No FPR)	1.16	1.09
Ours (Using FPR after 3 iter)	1.07	1.05
GT Landmarks (No FPR)	0.87	0.77

Table 4. Ablation study for change in reconstruction performance (MSE in mm) with change in 2D landmarks and face shape prior. It can be observed that using GT 2D landmark provides a significant improvement of ~0.3mm while GT Face provides less than .1 mm improvement. The proposed FPR method reduces error to 1.07mm

5. Limitations

In comparison to the state-of-the-art methods, our proposed method achieves significantly better 3D face reconstruction both quantitatively and qualitatively. Our method's reliance on photometric stereo as an intermediary step means that it faces challenges. For instance, it struggles when there's a significant angle between the input views, leading to difficulties in stereo matching due to large disparity and matching ambiguity. Additionally, the requirement for images to be captured under uniform conditions is essential for effective stereo matching.

6. Conclusion

This paper proposes an end-to-end method for 3D face reconstruction from two uncalibrated images. We introduce a strong face shape prior to the face pose estimation in order to make the optimization stable and accurate. This pose is then used for stereo reconstruction, followed by a 3DMM fitting to find the face shape. An iterative face pose refinement procedure improves the face pose and consequently reconstruction accuracy. The proposed method is evaluated on two widely used face datasets, and outperforms SOTA methods.

References

[1] Stirling/esrc 3D face database. 5, 6

[2] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid ICP algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 7

[3] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018. 2

[4] Ziqian Bai, Zhaopeng Cui, Xiaoming Liu, and Ping Tan. Riggable 3D face reconstruction via in-network optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6216–6225, 2021. 6, 7, 8

[5] Ziqian Bai, Zhaopeng Cui, Jamal Ahmed Rahim, Xiaoming Liu, and Ping Tan. Deep facial non-rigid multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5850–5860, 2020. 3, 6, 7, 8

[6] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019. 4

[7] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. In *ACM SIGGRAPH 2010 papers*, pages 1–9. 2010. 3

[8] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 351–358, 1995. 2

[9] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003. 1

[10] George Borshukov, Dan Piponi, Oystein Larsen, John P Lewis, and Christina Tempelaar-Lietz. Universal capture-image-based facial animation for" the matrix reloaded". In *ACM Siggraph 2005 Courses*, pages 16–es. 2005. 1

[11] Neill DF Campbell, George Vogiatzis, Carlos Hernández, and Roberto Cipolla. Using multiple hypotheses to improve depth-maps for multi-view stereo. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, pages 766–779. Springer, 2008. 3

[12] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2013. 5

[13] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1538–1547, 2019. 3

[14] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, pages 303–312, 1996. 3

[15] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, pages 145–156, 2000. 1

[16] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 4

[17] Pengfei Dou and Ioannis A Kakadiaris. Multi-view 3D face reconstruction with deep recurrent neural networks. *Image and Vision Computing*, 80:80–91, 2018. 3, 4

[18] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3D morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 2

[19] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021. 4, 5, 7

[20] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2009. 3

[21] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3D face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1155–1164, 2019. 4

[22] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3D dense face alignment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX*, pages 152–168. Springer, 2020. 4

[23] Wilfried Hartmann, Silvano Galliani, Michal Havlena, Luc Van Gool, and Konrad Schindler. Learned multi-patch similarity. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1586–1594, 2017. 3

[24] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, page 0, 2006. 3

[25] Kiriakos N Kutulakos and Steven M Seitz. A theory of shape by space carving. *International Journal of Computer Vision*, 38:199–218, 2000. 3

[26] Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, and Xuansong Xie. A hierarchical representation network for accurate and detailed face reconstruction from in-the-wild images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 5, 6, 7, 8

[27] Maxime Lhuillier and Long Quan. A quasi-dense approach to surface reconstruction from uncalibrated images. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 27(3):418–433, 2005. 3

[28] Hao Li, Thibaut Weise, and Mark Pauly. Example-based facial rigging. *Acm transactions on graphics (tog)*, 29(4):1–6, 2010. 2

[29] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017. 2, 4, 5

[30] Jiahao Luo, Fahim Hasan Khan, Issei Mori, Akila de Silva, Eric Sandoval Ruezga, Minghao Liu, Alex Pang, and James Davis. How much does input data type impact final face model accuracy? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18985–18994, 2022. 4

[31] Tetiana Martyniuk, Orest Kupyn, Yana Kurlyak, Igor Krashenyi, Jiři Matas, and Viktoriia Sharmanska. Dad-3dheads: A large-scale dense, accurate and diverse dataset for 3D head alignment from a single image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 4

[32] Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nistér, and Marc Pollefeys. Real-time visibility-based fusion of depth maps. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. Ieee, 2007. 3

[33] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE International Conference on Advanced Video and Signal based Surveillance*, pages 296–301. Ieee, 2009. 2

[34] Eduard Ramon, Janna Escur, and Xavier Giro-i Nieto. Multi-view 3D face reconstruction in the wild using siamese networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3, 4

[35] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 4

[36] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 7

[37] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 1, 3, 7

[38] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116. Citeseer, 2007. 7

[39] Gabriel Taubin. A signal processing approach to fair surface design. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999. 7

[40] George Terzakis and Manolis Lourakis. A consistently fast and globally optimal solution to the perspective-n-point problem. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 478–494. Springer, 2020. 2, 4

[41] Engin Tola, Christoph Strecha, and Pascal Fua. Efficient large-scale multi-view stereo for ultra high-resolution image sets. *Machine Vision and Applications*, 23:903–920, 2012. 3

[42] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3D face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019. 2

[43] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7346–7355, 2018. 2

[44] Luan Tran and Xiaoming Liu. On learning 3D face morphable model from in-the-wild images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):157–171, 2019. 2

[45] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5163–5172, 2017. 4

[46] Levi Valgaerts, Chenglei Wu, Andrés Bruhn, Hans-Peter Seidel, and Christian Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graph.*, 31(6):187–1, 2012. 3

[47] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 160–177. Springer, 2022. 4

[48] Fanzi Wu, Linchao Bao, Yajing Chen, Yonggen Ling, Yibing Song, Songnan Li, King Ngi Ngan, and Wei Liu. Mvf-net: Multi-view 3D face morphable model regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 959–968, 2019. 4, 6, 7, 8

[49] Fei Yang, Jue Wang, Eli Shechtman, Lubomir Bourdev, and Dimitri Metaxas. Expression flow for 3D-aware face component transfer. In *ACM SIGGRAPH 2011 papers*, pages 1–10. 2011. 1

[50] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 5, 6

[51] Hongwei Yi, Zizhuang Wei, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Pyramid multi-view stereo net with self-adaptive view aggregation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 766–782. Springer, 2020. 3

[52] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3D solu-

tion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 146–155, 2016. 4

[53] Michael Zollhöfer, Michael Martinek, Günther Greiner, Marc Stamminger, and Jochen Süßmuth. Automatic reconstruction of personalized avatars from 3D face scans. *Computer Animation and Virtual Worlds*, 22(2-3):195–202, 2011. 1