Deep Facial Non-Rigid Multi-View Stereo

Ziqian Bai[1] Zhaopeng Cui[2] Jamal Ahmed Rahim[1] Xiaoming Liu[3] Ping Tan[1]
[1] Simon Fraser University [2] ETH Zürich [3] Michigan State University
{ziqianb, jrahim, pingtan}@sfu.ca, zhpcui@gmail.com, liuxm@cse.msu.edu

Abstract

We present a method for 3D face reconstruction from multi-view images with different expressions. We formulate this problem from the perspective of non-rigid multi-view stereo (NRMVS). Unlike previous learning-based methods, which often regress the face shape directly, our method optimizes the 3D face shape by explicitly enforcing multi-view appearance consistency, which is known to be effective in recovering shape details according to conventional multi-view stereo methods. Furthermore, by estimating face shape through optimization based on multi-view consistency, our method can potentially have better generalization to unseen data. However, this optimization is challenging since each input image has a different expression. We facilitate it with a CNN network that learns to regularize the non-rigid 3D face according to the input image and preliminary optimization results. Extensive experiments show that our method achieves the state-of-the-art performance on various datasets and generalizes well to in-the-wild data.

1. Introduction

3D face reconstruction from images has been extensively studied in computer vision and graphics due to its wide range of applications in face recognition, entertainment, and medical analysis. Multi-view approaches have been one of the typical choices for high-end products of face reconstruction. With the images captured under well calibrated multi-view systems like camera arrays, faithful 3D geometry can be recovered with algorithms leveraging multi-view geometric constraints [5, 10]. However, this class of methods heavily rely on synchronized multi-view data which could be expensive, or sometimes even impossible, to acquire due to either the bulky equipment setup or the static face assumption. This drawback severely limits the possible application realms, especially in daily entertainment and communication. To handle this limitation, non-rigid multi-view approaches, *i.e.*, Non-Rigid Structure-from-Motion (NRSfM), are proposed to leverage multi-view geometry constraints for reconstructing deformable

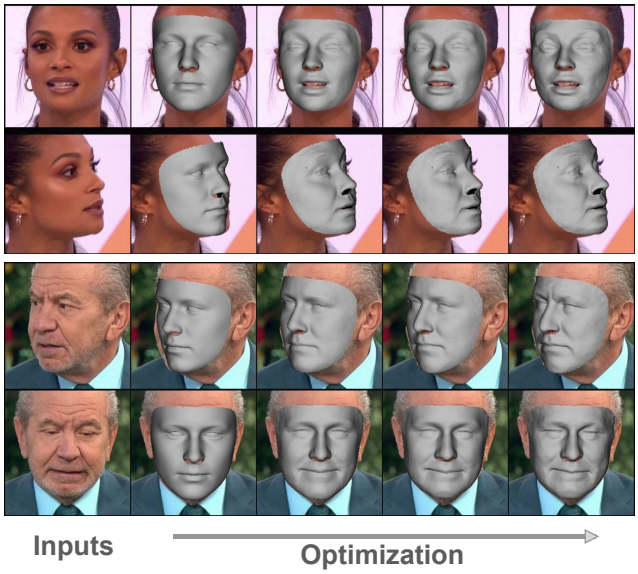

Figure 1: We present Deep Facial Non-Rigid Multi-View Stereo (DFNRMVS) to recover high-quality 3D models from multiple images of dynamic faces through multi-view optimization. From left to right are input images, initial 3D models, and 3D models after three-level optimization. Our DFNRMVS can gradually improves 3D models.

subjects, *e.g.*, faces with expression changes. However, the majority of works in this category only deal with sparse feature points [15, 31, 65]. Recently, dense approaches have been proposed [22, 33, 34], but usually contain complex modeling of geometry and rarely utilize data-driven priors.

With recent progress on deep learning, many face reconstruction methods are proposed to directly learn the image-to-parameter or image-to-geometry mapping purely from training data (*i.e.* regression), which makes their models data hungry. As a result, they normally resort to synthetic data [42], pre-computed 3DMM fitting [56, 64], or self-supervised learning [25, 49], which harms either the generalization or the reconstruction accuracy.

In this paper, we introduce the Deep Facial Non-Rigid Multi-View Stereo (DFNRMVS), which is the first end-to-end trainable network learning non-rigid multi-view stereo reconstruction for faces. This is achieved by first for-

mulating the dynamic face reconstruction task as a Non-Rigid Multi-View Stereo (NRMVS) optimization, which optimizes the 3D face shape by explicitly enforcing multi-view appearance consistency. Then, deep learning components are injected into this optimization pipeline to alleviate the problem difficulty with data-driven priors. Different from all previous NRSfM as well as learning-based face reconstruction methods, our model learns the process of parameter optimization explicitly in the network. It not only reduces the burden of the network and leads to better reconstructions, but also increases the generalization ability of our model, since the optimization brings in the domain knowledge of multi-view geometry. As a result, our model can be trained on limited but informative 3D scans to boost the performance while generalizing well to different 3D scan datasets and even to in-the-wild images.

To better regularize the ill-posed non-rigid setup as well as deal with the limited representation power of 3DMM, we also propose to learn an adaptive face model, which generates additive shape bases according to input images and preliminary optimization results. Comparing to generic face model, this additional information helps to customize the solution space case-by-case, making it more suitable for the optimization to produce better reconstructions. Our DFNRMVS achieves the state-of-the-art performance on various datasets with good generalization to the data in the wild. The code is available at https://github.com/zqbai-jeremy/DFNRMVS.

2. Related Work

Geometry-based methods. Geometry-based methods reconstruct the 3D face model based on multi-view stereo (MVS) [20] and photometric stereo [59]. Beeler *et al*. [5] proposed a stereo system to capture the 3D face geometry using seven cameras under standard light sources. Bradley *et al*. [10] presented a facial capture approach using a camera array which was able to reconstruct high resolution time-varying face meshes at 30 frames per second. These pure passive methods normally have poor reconstruction quality in the textureless regions as the stereo methods heavily rely on the feature matching. Many methods [26, 30, 44] also use the photometric stereo [59] for face reconstruction. Given the images captured under different illuminations, the surface normal is estimated first and then the 3D mesh is recovered through normal integration. These methods normally suffer from the convex/concave ambiguity. Approaches [21, 27] have been proposed to utilize the best of both worlds, where MVS recovers the base shape and photometric stereo recovers fine details. One major drawback of all mentioned methods is that they require either the images are synchronized or the subject is static during data capturing, which limits the applicable scenarios.

To address this problem, Non-Rigid Structure-from-Motion (NRSfM) has proposed to reconstruct objects with non-rigid deformations, such as face with different expressions. Bregler *et al*. [11] proposed to use linear subspace of low rank to represent the non-rigid 3D shape. Dai *et al*. [15] proved that the ill-posedness of NRSfM can be solved by only the low rank assumption, which was extended to temporal domain [3, 19]. More recently, progresses have also been made for methods based on union-of-subspace [2, 65], sparse prior [31], and deep learning [32]. However, this set of methods mainly focus on sparse points. Very recently, dense NRSfM becomes possible with variational approach [22], and Grassmann manifold [33, 34]. However, these methods rarely utilize the strong capability of data-driven priors captured by deep learning.

Different from all previous dense non-rigid methods, our method injects deep learning techniques into the reconstruction pipeline to alleviate the difficulty of the problem via priors learned from informative ground truth.

Learning-based methods. The data-driven prior of facial geometry is also exploited for face reconstruction from images. 3D Morphable Model (3DMM) [7] is a classic example, which is widely used to parameterize the shapes of human faces. Given input images, the optimal 3DMM parameters that fit the input are usually estimated by analysis-by-synthesis optimization [6, 43, 52]. As the optimization depends on the initialization, these methods are not quite robust in practice. Moreover, these methods are limited by the representation power of 3DMM. So extensive facial databases have been published recently to deal with complex expressions [8, 28, 35, 58, 62]. More recent works take a step further to also recover medium- and fine-scale details via corrective basis [9, 24] and Shape-from-Shading [23]. However, their models are usually computationally expensive due to the large amount of parameters to be optimized.

With the recent advances of deep learning, many methods are proposed for monocular face reconstruction. Various networks are designed to regress parameters of face models or 3D geometry with supervision from synthetic data [25, 42, 47], pre-computed 3DMM fitting [17, 56], RGB images [16, 50, 51, 54, 55], and identity labels [45]. To handle complex face geometry more flexibly, methods [13, 53, 57] regress the geometric residuals to recover fine-scale details. However, these methods mainly focus on single-view reconstruction. Only very recently, multiple images based methods are proposed [49, 60].

Different from prior learning-based multi-view methods, where reconstruction is generally formulated as regression, our method explicitly incorporates multi-view geometric constraints inside the learning framework via end-to-end trainable optimization. Thus, our model is a novel fusion of geometry- and learning-aspects leveraging the best of both worlds: the *quality* and *generalization* of geometry-based methods and the *robustness* of learning-based methods.

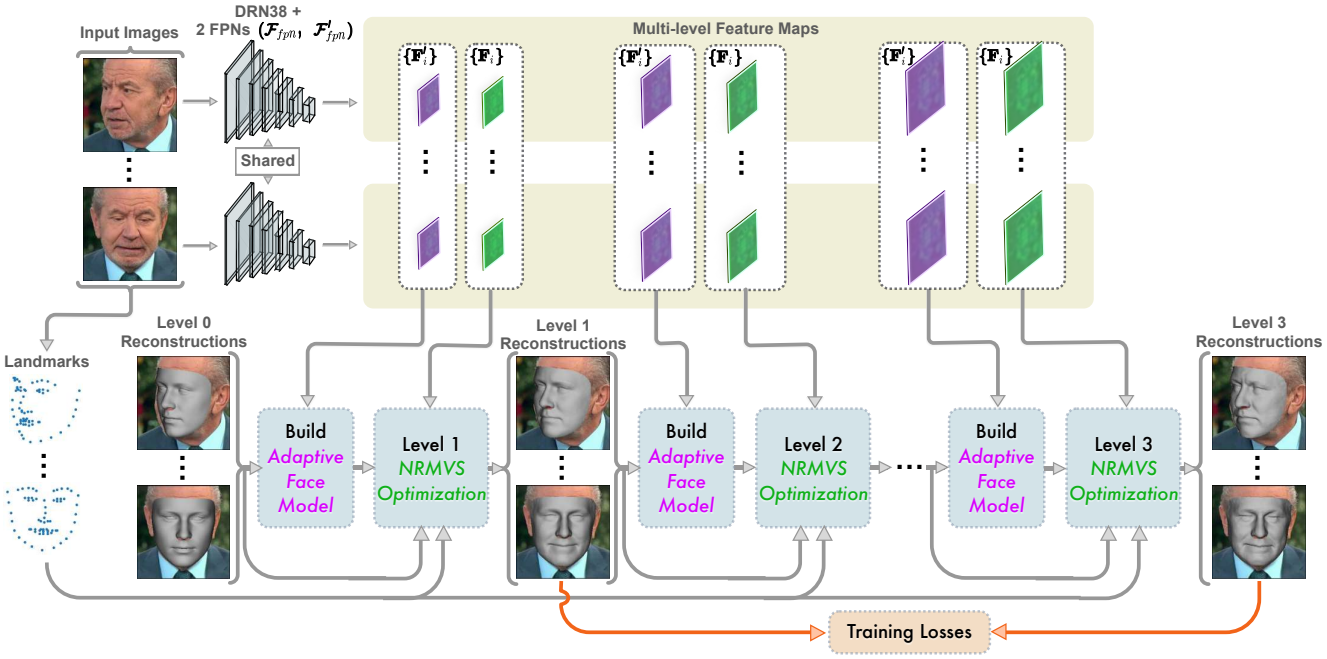

Figure 2: An overall of our method including (1) Non-Rigid Multi-View Stereo (NRMVS) optimization (Sec. 3.1, Sec. 3.2); (2) Adaptive Face Model (Sec. 3.3); (3) Multi-level reconstruction scheme (Sec. 3.4); and (4) Training losses (Sec. 3.5).

3. Proposed Method

Given multiple facial images, we aim to design a deep neural network to generate detailed 3D face models by exploiting the multi-view appearance consistency explicitly in the network. As shown in Fig. 2, our framework mainly consists of three modules: feature extraction, adaptive face model generation, and non-rigid multi-view stereo optimization. We will first present the non-rigid multi-view stereo optimization (Sec. 3.1) and explain how this optimization is integrated with deep learning through a learnable objective and solver (Sec. 3.2). Then we will introduce the adaptive face model generation (Sec. 3.3). Finally, we present our multi-level reconstruction scheme (Sec. 3.4) and training losses (Sec. 3.5).

3.1. Non-Rigid Multi-View Stereo

Given a set of M facial images $\{\mathbf{I}_i\}_{i=1}^{M}$ capturing the same person but under different expressions and views, the estimation of 3D facial geometry $\mathbf{V}_i$ and 6 DoF rigid head pose $\mathbf{p}_i$ for each image can be formulated as a Non-Rigid Multi-View Stereo (NRMVS) optimization by minimizing the appearance-consistency error and landmark fitting error.

Parameterization. For head pose $\mathbf{p}$, we parameterize it with $\mathbf{p} = (s, \mathbf{R}, \mathbf{t})$ under the weak perspective camera model assumption, where s is a scale factor, $\mathbf{R} \in SO(3)$ is the rotation matrix, and $\mathbf{t} \in \mathbb{R}^2$ is the 2D translation on the image plane. Thus, the projection $\mathbf{\Pi} : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ mapping a 3D point $\mathbf{v} \in \mathbb{R}^3$ to the image plane is,

$$\mathbf{\Pi}(\mathbf{v}) = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \end{bmatrix} \mathbf{R}\mathbf{v} + \mathbf{t}. \tag{1}$$

Similar to linear 3DMM [7], we parameterize the face shape as $\mathbf{V} = f(\mathbf{x})$, where the generator function $f(\mathbf{x})$ maps the low-dimensional parameter vector $\mathbf{x} \in \mathbb{R}^K$ to the vector comprising 3D coordinates of all N vertices, $\mathbf{V} \in \mathbb{R}^{3N}$. Therefore, the parameters of the non-rigid multi-view stereo optimization can be represented as $\mathbf{X} = (s, \mathbf{R}, \mathbf{t}, \mathbf{x})$.

Objective Function. The objective function of our non-rigid multi-view stereo optimization is as the following,

$$\mathbf{E} = \lambda_a \mathbf{E}_a + \lambda_l \mathbf{E}_l, \tag{2}$$

where $\mathbf{E}_a$ is the appearance consistency error across views, and $\mathbf{E}_l$ is the facial landmarks alignment error. λ_a and λ_l balance the importance of two objectives.

For appearance consistency $\mathbf{E}_a$, a naïve option is to use image intensity difference as the consistency metric. For each view i, we project the current reconstruction $(\mathbf{V}_i, \mathbf{p}_i)$ onto the image $\mathbf{I}_i$ by Eq. (1), and sample the intensity via bilinear interpolation. As a result, each vertex will have an intensity value $I(\mathbf{v}_i)$. Then, for each pair of views (i, j) where $i \neq j$, we compute the intensity difference of corresponding vertices and average across all vertices and views. To sum up, we have

$$\mathbf{E}_a = \frac{2}{M(M-1)} \sum_{i \neq j} \frac{1}{N} \sum_{k=1}^{N} \|I(\mathbf{v}_i^k) - I(\mathbf{v}_j^k)\|_2^2, \tag{3}$$

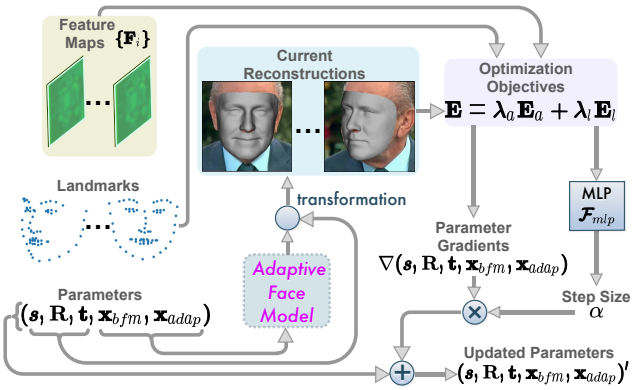

Figure 3: Overview of one iteration of the NRMVS optimization described in Sec. 3.1 and Sec. 3.2.

where $\mathbf{v}_i^k$ ($\mathbf{v}_j^k$) denotes the k-th vertex of view i (view j). We also take vertex visibility into account and approximate it by backface culling as in [50].

In terms of landmark alignment $\mathbf{E}_l$, we adopt a similar objective as Tewari *et al*. [49, 50], which takes sliding landmarks on the face contour into account. Given the 68 detected landmarks $\{\mathbf{u}_i^k\}_{k=1}^{68}$ on each image $\mathbf{I}_i$ via an off-the-shell method [12], the objective reads as:

$$\mathbf{E}_l = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{68}\sum_{k=1}^{68}\|\mathbf{u}_i^k - \mathbf{\Pi}(\mathbf{v}_i^{\mathbf{u}^k})\|_2^2, \tag{4}$$

where $\mathbf{v}_i^{\mathbf{u}^k}$ denotes the mesh vertex corresponding to k-th landmark for view (or image) i.

Optimization Solver. Due to the differentiability, the objective $\mathbf{E}$ can be minimized by gradient based solvers. For simplicity, we choose the first order optimization solver (*i.e*., gradient descent). The reconstruction parameters can be updated iteratively by

$$\mathbf{X} \leftarrow \mathbf{X} + \alpha \nabla_{\mathbf{X}} \mathbf{E}(\mathbf{X}), \tag{5}$$

where α is the step size.

3.2. Learnable Objective and Solver

Directly solving the proposed NRMVS optimization (Sec. 3.1) could be extremely difficult due to the highly non-convex intensity space. Inspired by recent works on **rigid** 3D reconstruction [48, 61] and motion estimation [38], we inject deep learning from two aspects to alleviate the difficulty: (1) more robust objective; (2) more flexible solver.

Feature-metric Appearance Objective. We replace the raw images $\{\mathbf{I}_i\}_{i=1}^M$ with feature maps $\{\mathbf{F}_i\}_{i=1}^M$ extracted by a *Feature Pyramid Network* (FPN) [36] $\mathcal{F}_{fpn}$ shown in Fig. 2 when computing the appearance consistency $\mathbf{E}_a$. Thus, the objective (3) becomes:

$$\mathbf{E}_a = \frac{2}{M(M-1)}\sum_{i\neq j}\frac{1}{N}\sum_{k=1}^{N}\|F(\mathbf{v}_i^k) - F(\mathbf{v}_j^k)\|_2^2, \tag{6}$$

where $F(\mathbf{v}_i)$ is the per-vertex feature vector sampled from the feature map $\mathbf{F}_i$, replacing the per-vertex image intensity $I(\mathbf{v}_i)$ sampled from raw image $\mathbf{I}_i$.

Step Size Prediction. Traditionally, the step size α for gradient descent is heavily tuned to ensure good performance. Instead, we use an MLP $\mathcal{F}_{mlp}$ shown in Fig. 3 that learns to predict α given the absolute residuals of the objectives averaged across vertices and views.

End-to-end learnable. The parameters of both networks $\mathcal{F}_{fpn}$ (in Fig. 2) and $\mathcal{F}_{mlp}$ (in Fig. 3) can be updated during end-to-end training. In principle, $\mathcal{F}_{fpn}$ learns to extract feature maps that are suitable for the optimization (*i.e*., more smooth and convex), while $\mathcal{F}_{mlp}$ learns to predict better step sizes that expedite the convergence (*i.e*., larger step size with larger magnitudes of objectives), reducing the difficulty of the optimization.

3.3. Adaptive Face Model

To better leverage existing 3DMM while not limited by its representation power, we propose an *Adaptive Face Model* that contains two linear subspaces $\mathbf{x} = (\mathbf{x}_{bfm}, \mathbf{x}_{adap})$. The final facial shape $\mathbf{V}$ is represented as,

$$\mathbf{V} = f(\mathbf{x}) = \overline{\mathbf{V}} + \mathbf{B}_{bfm}\mathbf{x}_{bfm} + \mathbf{B}_{adap}\mathbf{x}_{adap}, \tag{7}$$

where $\overline{\mathbf{V}} \in \mathbb{R}^{3N}$ is the mean shape, $\mathbf{B}_{bfm} \in \mathbb{R}^{3N\times K_{bfm}}$ is the PCA basis from *Basel Face Model* (BFM) [40] that is common to all faces, and $\mathbf{B}_{adap} \in \mathbb{R}^{3N\times K_{adap}}$ is the adaptive basis that is built from the input images and preliminary reconstructions. The coefficient $\mathbf{x}_{bfm} \in \mathbb{R}^{K_{bfm}}$, termed as the BFM parameter, is constant across different views, while $\mathbf{x}_{adap} \in \mathbb{R}^{K_{adap}}$, termed as the adaptive parameter, varies across views.

Since the adaptive basis $\mathbf{B}_{adap}$ is built according to the initial (or intermediate) pose $\{\hat{\mathbf{p}}_i\}_{i=1}^M$ and geometry $\{\hat{\mathbf{V}}_i\}_{i=1}^M$, it can hopefully capture the aspects where preliminary reconstructions fail to explain the input. To achieve this goal, it needs to be built right before the NRMVS optimization, as shown in Fig. 2.

For each view i, a feature map $\mathbf{F}_i'$ is firstly extracted from the image $\mathbf{I}_i$ via a separate FPN [36] $\mathcal{F}_{fpn}'$ shown in Fig. 2 as the following,

$$\mathbf{F}_i' = \mathcal{F}_{fpn}'(\mathbf{I}_i). \tag{8}$$

In total, we have M feature maps $\{\mathbf{F}_i'\}_{i=1}^M$, *i.e*., the left most column in Fig. 4. Then, we can obtain the adaptive basis $\mathbf{B}_{adap}$ by feeding these feature maps and the preliminary reconstructions $(\{\hat{\mathbf{V}}_i\}, \{\hat{\mathbf{p}}_i\})_{i=1}^M$ into the basis network $\mathcal{F}_{basis}$ shown in Fig. 4,

$$\mathbf{B}_{adap} = \mathcal{F}_{basis}\left(\{\mathbf{F}_i'\}_{i=1}^M, (\{\hat{\mathbf{V}}_i\}, \{\hat{\mathbf{p}}_i\})_{i=1}^M\right). \tag{9}$$

More specifically, we map these feature maps $\{\mathbf{F}_i'\}_{i=1}^M$ into the UV texture space according to the preliminary 3D

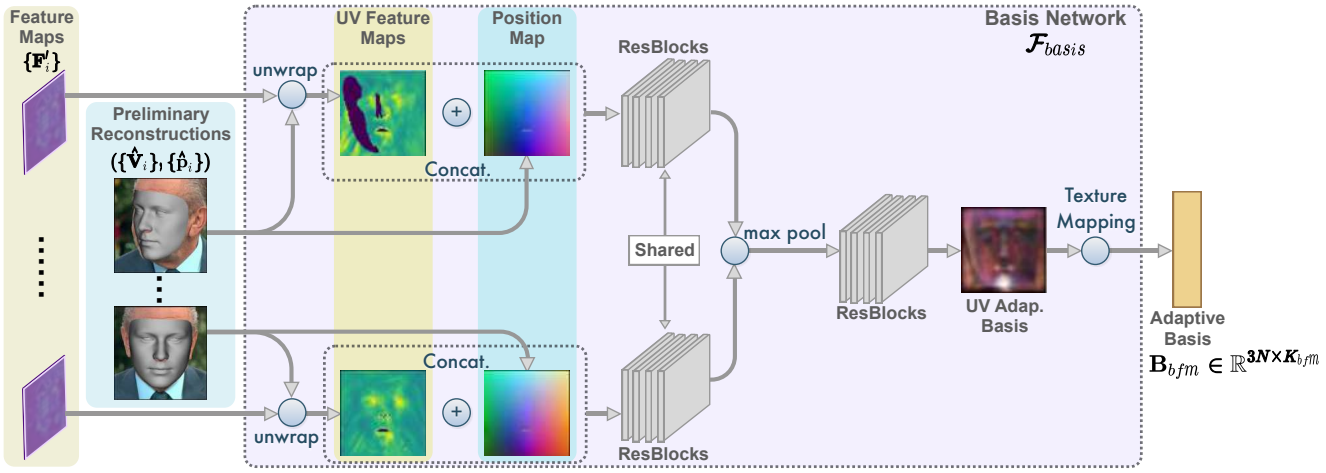

Figure 4: Pipeline of adaptive basis generation (Sec. 3.3).

face shape. We also convert each preliminary reconstruction to the UV space as *Position Map* [17] (*i.e.*, a 2D image recording 3D positions of all vertices in the UV space), and concatenate it with the corresponding unwrapped feature map along the channel dimension. The resulting M feature maps go through M Siamese branches separately, then are max pooled along the view dimension, and decoded to the UV texture representation of the adaptive basis. The final adaptive basis $\mathbf{B}_{adap}$ can be generated through standard texture mapping. Please refer to our supplementary for more details about the basis network $\mathcal{F}_{basis}$ for adaptive face model generation.

3.4. Multi-level Reconstruction

In order to better recover the details of 3D face shapes, we adopt a multi-level scheme. Specifically, we split the reconstruction process into 3 sequential levels $l = 1, 2, 3$, each of which solves a NRMVS optimization and outputs the reconstructions for all views $(\{\mathbf{V}_i^l\}, \{\mathbf{p}_i^l\})_{i=1}^M$.

At each level, the face model is updated as

$$\mathbf{V}^l = f^l(\mathbf{x}^l) = \overline{\mathbf{V}} + \mathbf{B}_{bfm}\mathbf{x}_{bfm} + \sum_{j=1}^{l} \mathbf{B}_{adap}^j \mathbf{x}_{adap}^j, \tag{10}$$

where the shape basis $\mathbf{B}_{bfm}$ is fixed for all levels, and the adaptive basis $\mathbf{B}_{adap}^l$ is updated from level to level. During NRMVS optimization at each level l, only head poses $\{\mathbf{p}_i\}_{i=1}^M$, BFM parameter $\mathbf{x}_{bfm}$, and the current level adaptive parameter $\{\mathbf{x}_{adap,i}^l\}_{i=1}^M$ will be optimized. Poses and BFM parameter are initialized with outputs from the previous level, and the adaptive parameter is initialized to zero at the beginning. At the very beginning (*i.e.*, level 0), the initial head pose is regressed by a pre-trained neural network (please refer to the supplementary material for details). The initial BFM parameter is set to zero. The initial preliminary reconstruction for each view is the mean face from BFM [40] transformed by the regressed pose. For later levels, the preliminary reconstructions are the output of the previous level.

3.5. Training Losses

Given ground truth meshes with the corresponding vertices of the reconstructed meshes, our network, *i.e.*, 2 FPNs $\mathcal{F}_{fpn}$ and $\mathcal{F}'_{fpn}$, the MLP $\mathcal{F}_{mlp}$ for step size prediction, and the basis network $\mathcal{F}_{basis}$, is trained in a supervised manner with standard losses. For each vertex, we compute the point-to-point L_2 distance between ground truth and reconstructed meshes (with poses) of all iterations, all views, and all levels after depth alignment and dense alignment separately (*i.e.*, 2 losses per vertex). For depth alignment, we compute the mean depth difference between predictions and ground truth, and add this difference to ground truth before computing the loss. For dense alignment, we use the correspondence of face mesh vertices to rigidly (with scale) align predictions to ground truth. These 2 vertex losses read as

$$L_{v\_dep} = \sum \|\mathbf{V}^{gt} - \mathbf{V}\|_2^2, \tag{11}$$

$$L_{v\_den} = \sum \|\mathbf{V}^{gt} - \mathbf{V}_{den}\|_2^2. \tag{12}$$

Intuitively speaking, the dense aligned loss L_{v_den} only measures the geometry error while the depth aligned loss L_{v_dep} also accounts for poses.

We also consider the supervision on per-vertex normal by measuring cosine similarity loss,

$$L_{norm} = \sum (1 - \cos(\mathbf{n}^{gt}, \mathbf{n}_{align})). \tag{13}$$

Following Liu *et al.* [37], an edge loss is also added,

$$L_{edge} = \frac{1}{\#E} \sum_{(i,j)\in E} \left| \frac{\|\mathbf{V}_i - \mathbf{V}_j\|}{\|\mathbf{V}_i^{gt} - \mathbf{V}_j^{gt}\|} - 1 \right|, \tag{14}$$

where E is the pre-defined edge graph of the template. The motivation of using L_{norm} and L_{edge} is to improve surface smoothness while preserve high-frequency details [37]. Finally, a landmark loss L_{land} similar to Eq. (4) is included. To sum, the total training loss is as the following,

$$L = \lambda_1 L_{v\_dep} + \lambda_2 L_{v\_den} + \lambda_3 L_{norm} + \lambda_4 L_{edge} + \lambda_5 L_{land}, \tag{15}$$

where $\lambda_{1,2,3,4,5}$ are hyperparameters that adjust the weights of different losses.

3.6. Optimization and Training

Here we further clarify the relationship between the NRMVS optimization and the training procedure. The NRMVS optimization belongs to the forward pass of our model. It can be analogized to a differentiable module, which takes in old reconstruction parameters and computes the updates to output new reconstructions iteratively. Then, the training losses are computed on the outputs of each iteration, whose gradient will be backwarded through the whole NRMVS optimization to update learnable weights (*i.e.*, the weights of $\mathcal{F}_{fpn}$ and $\mathcal{F}'_{fpn}$ in Fig. 2, $\mathcal{F}_{mlp}$ in Fig. 3, and $\mathcal{F}_{basis}$ in Fig. 4).

4. Experiments

4.1. Experimental Setup

Training data. We adopt the *Stirling/ESRC 3D face database* [1] to train our model. The dataset contains high quality 3D scans of more than 100 subjects. The majority of subjects have 3D scans for 8 different expressions. For each scan, 2 RGB images taken from ± 45 yaw angles are used as textures. We use the textured 3D scans to render images for training. More specifically, we select 85, 20, 35 non-overlapping subjects as training, validation, testing splits. To generate a training sample, two random expressions of the same subject are firstly selected. Then, we render one image for each expression with different poses and same global illumination using Spherical Harmonics (SH) [41]. As a result, around 8K training samples are generated.

To obtain the ground truth dense correspondences, we run Non-Rigid ICP [4] to register the mean shape from BFM [40] to each 3D scan, and use the results as the ground truth dense correspondences. Note that even trained on this limited number of samples, our model can still generalize to other 3D scan datasets as well as in-the-wild images.

Implementation. Our model is implemented with Pytorch [39]. For the optimization, the objective weights are $\lambda_a = 0.25$, and $\lambda_l = 0.025$. We use 3 levels of optimization with the feature map resolution of 32×32, 64×64, and 128×128 respectively. In each level, 3 iterations of parameter updates are computed. During training, the losses are weighted as $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 100$, $\lambda_4 = 0.01$, and $\lambda_5 = 0.1$. The Adam [29] optimizer is applied for training with learning rate of 2.0×10^{-5}. The batch size is set to 2.

Figure 5: Qualitative comparison with Feng *et al.* [17], Chen *et al.* [13], and Tewari *et al.* [49]. For two-view methods, images of two consecutive rows are input together. Readers may zoom in and pay attention to details such as (1) face contour alignment (2^{nd} & 4^{th} row), (2) inputs with large expression change (5^{th} & 6^{th} row), and (3) medium-scale details such as large wrinkles (2^{nd} row), smiling line (6^{th} row), half-opened eyes (4^{th} row), and geometry around eyes (1^{st} & 3^{rd} & 7^{th} row).

Baseline Methods. We choose the following methods as baselines in the qualitative and quantitative evaluations. Tewari *et al.* [49] is a regression-based method that also tackles non-rigid multi-view face reconstruction, which is the most related one to ours. Thus, we treat it as an important baseline in both qualitative and quantitative evaluations. Several single-view reconstruction methods are also compared. Feng *et al.* [17] directly regress the face geometry in the form of *Position Maps*. Tewari *et al.* [50] learn a corrective basis on top of 3DMM and regress the basis parameters. These two baselines represent single-view methods that are not restricted by the 3DMM subspace. Deng *et al.* [16] act as the baseline for state-of-the-art 3DMM fitting method. We also include Chen *et al.* [13] in the comparison as it presents an interesting combination of optimization-

Table 1: Geometric errors on BU3DFE [62] dataset. [Key: Two input images with the same (S), or different expressions (D)]

	Two Views				Single View	
	Ours (S)	Ours (D)	Ours NoAdap (S)	Tewari *et al.* [49] (S)	Deng *et al.* [16]	Tewari *et al.* [50]
Mean (mm)	**1.11**	1.14	1.12	1.74	1.38	1.78
STD (mm)	**0.29**	**0.29**	0.33	0.45	0.37	0.49

Table 2: Geometric errors on Bosphorus database [46]. Inputs are 2-view image pairs with different expressions.

	Ours	Deng *et al.* [16]
Mean (mm)	**1.44**	1.47
STD (mm)	**0.38**	0.40

based landmark fitting and emotion priors captured by deep learning while can also synthesis facial details.

4.2. Qualitative Evaluation

We perform qualitative comparison on VoxCeleb2 [14], an in-the-wild facial video dataset collected from YouTube. We use the same set of images as Tewari *et al.* [49] uploaded in their website. For Tewari *et al.* [49], we directly use their uploaded results for comparison. For other baselines, we use their public implementations to generate the results. The visualizations are shown in Fig. 5.

Comparison with Tewari *et al.* [49]. Our results are better aligned to the faces due to the explicit optimization on the multi-view appearance consistency. Note that [49] is trained on a large amount of in-the-wild data, which should generalize better to the tested in-the-wild faces. However, even though we use limited 3D scans with rendered images during training, our model still achieves comparable or even better generalization on these in-the-wild faces. In addition, since our model is trained on high-quality 3D scans, it is able to capture medium-level details while [49] only gives coarse reconstructions.

Comparison with single-view methods. We also compared with single-view reconstructions [17] and [13]. Our method gives better results than [17] with the help of multi-view geometry. Our method also performs better than the emotion-driven proxy estimation in [13]. Although the additional detail synthesis in [13] can produce locally appealing results (*e.g.*, wrinkles on the forehead), it cannot correct the unfaithful geometry from proxy estimation.

4.3. Quantitative Evaluation

The BU3DFE [62] and the Bosphorus [46] datasets are used to quantitatively evaluate our method. The authors of [49] and [50] kindly provided their reconstructed meshes. For other baselines, the reconstructions are obtained with their public implementations. Then, we compute the geometric errors on all reconstructions in a consistent manner.

Evaluation on BU3DFE [62]. The BU3DFE dataset [62] includes 3D scans of 100 subjects with neutral face and 24 expressions. Each scan in BU3DFE is associated with 2 RGB images taken from $\pm 45^\circ$ yaw angles. We use the testing split provided by Tewari *et al.* [49] to evaluate our method with 2-view images of the same subject, either under the same expression or different ones.

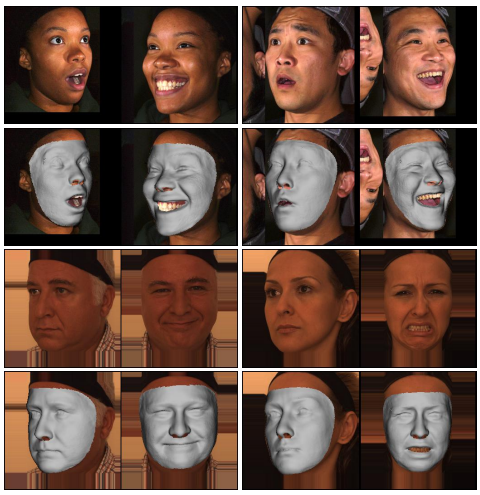

Figure 6: Visualization of our reconstructions on BU3DFE [62] (first 2 rows) and Bosphorus [46] (last 2 rows). Our method can capture different expression changes.

To compute the geometric errors, we first align the reconstructions with ground truth using 8 landmarks given by BU3DFE [62]. Then, ICP [63] is performed to further align the reconstructions to ground truth. Finally, we crop the ground truth based on landmarks with a similar strategy as [18] and compute the point-to-plane distance from ground truth to reconstruction. Our results are shown in Table 1 and compared to the state-of-the-art approaches. We also show qualitative examples of our results in Fig. 6.

Our learning-based optimization outperforms the regression-based approach [49]. Note that their model is trained in a self-supervised manner with in-the-wild videos for the sake of better generalization, which may affect the geometry accuracy. It is unclear whether their method can be trained with limited 3D scans without harming generalization. In contrast, our model can leverage 3D scans for better geometry accuracy without affecting generalization. Our method also outperforms state-of-the-art single-view approaches [16, 50] as we leverage the additional multi-view cue via learning-based optimization.

Evaluation on Bosphorus [46]. The Bosphorus database contains 105 subjects, each with expressive face images under frontal-view and neutral face images under various poses. For each subject, we select all images with emotion labels (only frontal-view provided), including expressions

Table 3: Geometric errors on rendered images using ESRC face database [1] for multi-view evaluation.

	2 Views	3 Views	4 Views
Mean (mm)	1.04	1.03	**1.02**
STD (mm)	0.33	0.30	**0.29**

of happy, surprise, fear, sadness, anger and disgust; then we select the neutral face image under -30° yaw angle to form 2-view image pairs. Note that for some subjects, only a subset of the mentioned expressions are available. In total, we collect 453 samples of 2-view images.

We use the mentioned protocol to compute the geometric errors, however, with a different set of 5 landmarks given by the database for alignment. The errors of neutral faces are down weighted accordingly as they are measured for multiple times. Results are shown in Table 2.

Our method achieves slightly better performance than Deng *et al*. [16] in both mean and standard deviation of errors. Note that the content difference between Bosphorus [46] and BU3DFE [62] leads to a non-trivial domain gap, which could affect the performance gain of our method comparing to *Ours (D)* in Table 1.

4.4. Ablation Study

More than Two Views. We evaluate how the number of views affect our method. For this evaluation, we use our test split of ESRC face database [1] with 35 subjects to render test images. To generate a test sample, we randomly select 4 expressions from a subject and render 4 images with an arbitrary global illumination and different poses. Here we order the images with increasing yaw angles for better illustration. When testing 2-view cases, the 1st and 4th images are used. For 3-view cases, (1st, 2nd, 4th) and (1st, 3rd, 4th) images are used separately. For 4-view cases, all images are used. We only measure the errors of the 1st and 4th images for fair comparison. Before testing with more that 2 views, we further finetune our model accordingly on the training split to better fit different number of views.

As shown in Table 3, by introducing more views during training and testing, our method achieves better performance in terms of both mean and standard deviation of geometric errors, which demonstrates the effectiveness of multi-view information.

Adaptive vs. Generic Basis. To demonstrate the benefit of adaptive basis, we design a baseline where we replace the adaptive basis with a generic one, which is common to different subjects and fixed during the optimization. Basically, we remove the basis network $\mathcal{F}_{basis}$ in Fig. 4 and set the UV texture representation of the basis (also shown in Fig. 4) as network parameters, which derives the generic basis via the same texture mapping. The quantitative result is shown in Table 1 as *Ours NoAdap (S)*. Although the mean error of generic basis is comparable to the one of adaptive basis, its standard deviation (STD) of error is larger, indicating that the results of generic basis are less stable. The qualitative comparison in Fig. 7 shows that generic basis tends to give more noisy reconstructions. Note that an additional smooth loss is applied when training the generic basis model, which has reduced the STD of geometric errors and alleviated the noisy pattern on outputs to some extent. The adaptive basis is able to derive the relationships among vertices from the spatial information of the images and the preliminary reconstructions, while generic basis treats each vertex independently. This property enables the adaptive basis to produce robust and smooth results.

Inputs Ours AdapB Ours GeneB

Figure 7: Qualitative comparison between Adaptive Basis (AdapB) and Generic Basis (GeneB).

Table 4: Geometric errors and running time on BU3DFE [62] dataset for multi-level scheme evaluation. Tested with RTX2080Ti.

	Level 1	Level 2	Level 3
Mean (mm)	1.29	1.18	**1.11**
STD (mm)	0.32	0.30	**0.29**
Time (s)	**0.12**	0.22	0.31

Multi-level Scheme. We also investigate the effectiveness of the multi-level scheme. Table 4 shows the quantitative results on BU3DFE [62]. The geometric error and its standard deviation (STD) consistently decrease as more levels of optimization are performed. Note that even our level 1 reconstructions outperform all baselines in Table 1. The last three columns in Fig. 1 are typical qualitative examples, showing that higher levels can better capture personalized details, *i.e*., wrinkles between eyebrows.

5. Conclusions

We solve 3D face reconstruction from multi-view images with different expressions by a novel Non-Rigid Multi-View Stereo (NRMVS) optimization framework. Our method introduces the traditional multi-view geometry (in terms of photo/feature-consistency) to the popular CNN-based face reconstruction. Solving 3D reconstruction by enforcing multi-view geometry constraints is effective in capturing shape details, and also improves the generalization to unseen data. Experiments demonstrate that our method achieves state-of-the-art performance and generalizes well to in-the-wild images, which proves the effectiveness of conventional multi-view geometry based optimization combined with modern CNNs. Although our NRMVS is still specific to faces, it is the first formulation of dense multi-view stereo with non-rigid motions, and hence, can be potentially applied to other non-rigid reconstruction problems.

References

[1] *Stirling/ESRC 3D face database*. 6, 8

[2] Antonio Agudo and Francesc Moreno-Noguer. Dust: Dual union of spatio-temporal subspaces for monocular multiple object 3d reconstruction. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 6262–6270, 2017. 2

[3] Ijaz Akhter, Yaser Sheikh, and Sohaib Khan. In defense of orthonormality constraints for nonrigid structure from motion. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1534–1541. IEEE, 2009. 2

[4] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2007. 6

[5] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. In *ACM Trans. on Graphics (TOG)*, volume 29, page 40. ACM, 2010. 1, 2

[6] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 25(9):1063–1074, 2003. 2

[7] Volker Blanz, Thomas Vetter, et al. A morphable model for the synthesis of 3d faces. In *Proc. of ACM SIGGRAPH*, volume 99, pages 187–194, 1999. 2, 3

[8] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision (IJCV)*, 126(2-4):233–254, 2018. 2

[9] Sofien Bouaziz, Yangang Wang, and Mark Pauly. Online modeling for realtime facial animation. *ACM Trans. on Graphics (TOG)*, 32(4):40, 2013. 2

[10] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. High resolution passive facial performance capture. In *ACM Trans. on Graphics (TOG)*, volume 29, page 41. ACM, 2010. 1, 2

[11] C Bregler, A Hertzmann, and H Biermann. Recovering non-rigid 3d shape from image streams. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 690–696. IEEE, 2000. 2

[12] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1021–1030, 2017. 4

[13] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2, 6, 7

[14] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. *Proc. Interspeech 2018*, pages 1086–1090, 2018. 7

[15] Yuchao Dai, Hongdong Li, and Mingyi He. A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision (IJCV)*, 107(2):101–122, 2014. 1, 2

[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 Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 2, 6, 7, 8

[17] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 534–551, 2018. 2, 5, 6, 7

[18] 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 *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 780–786. IEEE, 2018. 7

[19] Katerina Fragkiadaki, Marta Salas, Pablo Arbelaez, and Jitendra Malik. Grouping-based low-rank trajectory completion and 3d reconstruction. In *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, pages 55–63, 2014. 2

[20] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 32(8):1362–1376, 2009. 2

[21] Graham Fyffe, Paul Graham, Borom Tunwattanapong, Abhijeet Ghosh, and Paul Debevec. Near-instant capture of high-resolution facial geometry and reflectance. In *Computer Graphics Forum*, volume 35, pages 353–363. Wiley Online Library, 2016. 2

[22] Ravi Garg, Anastasios Roussos, and Lourdes Agapito. Dense variational reconstruction of non-rigid surfaces from monocular video. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1272–1279, 2013. 1, 2

[23] Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *Proc. of ACM SIGGRAPH*, 32(6):158–1, 2013. 2

[24] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Trans. on Graphics (TOG)*, 35(3):28, 2016. 2

[25] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 8377–8386, 2018. 1, 2

[26] Athinodoros S Georghiades. Recovering 3-d shape and reflectance from a small number of photographs. In *Proceedings of the 14th Eurographics workshop on Rendering*, pages 230–240. Eurographics Association, 2003. 2

[27] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. In *ACM Trans. on Graphics (TOG)*, volume 30, page 129. ACM, 2011. 2

[28] Patrik Huber, Guosheng Hu, Rafael Tena, Pouria Mortazavian, P Koppen, William J Christmas, Matthias Ratsch, and Josef Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016. 2

[29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[30] Martin Klaudiny and Adrian Hilton. High-detail 3d capture and non-sequential alignment of facial performance. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 17–24. IEEE, 2012. 2

[31] Chen Kong and Simon Lucey. Prior-less compressible structure from motion. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 4123–4131, 2016. 1, 2

[32] Chen Kong and Simon Lucey. Deep non-rigid structure from motion. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1558–1567, 2019. 2

[33] Suryansh Kumar. Jumping manifolds: Geometry aware dense non-rigid structure from motion. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 5346–5355, 2019. 1, 2

[34] Suryansh Kumar, Anoop Cherian, Yuchao Dai, and Hongdong Li. Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 254–263, 2018. 1, 2

[35] 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. on Graphics (TOG)*, 36(6):194, 2017. 2

[36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017. 4

[37] Feng Liu, Luan Tran, and Xiaoming Liu. 3d face modeling from diverse raw scan data. In *Proc. of International Conference on Computer Vision (ICCV)*, Seoul, South Korea, October 2019. 5, 6

[38] Zhaoyang Lv, Frank Dellaert, James M Rehg, and Andreas Geiger. Taking a deeper look at the inverse compositional algorithm. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 4581–4590, 2019. 4

[39] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[40] 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. 4, 5, 6

[41] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 117–128. ACM, 2001. 6

[42] Elad Richardson, Matan Sela, and Ron Kimmel. 3d face reconstruction by learning from synthetic data. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 460–469. IEEE, 2016. 1, 2

[43] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 986–993. IEEE, 2005. 2

[44] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 39(11):2127–2141, December 2016. 2

[45] 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 *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, 2019. 2

[46] Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. Bosphorus database for 3d face analysis. In *European Workshop on Biometrics and Identity Management*, pages 47–56. Springer, 2008. 7, 8

[47] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1576–1585, 2017. 2

[48] Chengzhou Tang and Ping Tan. Ba-net: Dense bundle adjustment network. *Intl. Conf. on Learning Representations (ICLR)*, 2019. 4

[49] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. Fml: face model learning from videos. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 10812–10822, 2019. 1, 2, 4, 6, 7

[50] 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 *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2549–2559, 2018. 2, 4, 6, 7

[51] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proc. of International Conference on Computer Vision (ICCV)*, pages 1274–1283, 2017. 2

[52] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 2

[53] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3D face morphable model. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[54] Luan Tran and Xiaoming Liu. Nonlinear 3D face morphable model. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[55] Luan Tran and Xiaoming Liu. On learning 3D face morphable model from in-the-wild images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. doi: 10.1109/TPAMI.2019.2927975. 2

[56] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 5163–5172, 2017. 1, 2

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

[58] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. Face transfer with multilinear models. In *ACM Trans. on Graphics (TOG)*, volume 24, pages 426–433. ACM, 2005. 2

[59] Robert J Woodham. Photometric stereo: A reflectance map technique for determining surface orientation from image intensity. In *Image Understanding Systems and Industrial Applications I*, volume 155, pages 136–143. International Society for Optics and Photonics, 1979. 2

[60] 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 *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 959–968, 2019. 2

[61] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 4

[62] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006. 2, 7, 8

[63] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 7

[64] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016. 1

[65] Yingying Zhu, Dong Huang, Fernando De La Torre, and Simon Lucey. Complex non-rigid motion 3d reconstruction by union of subspaces. In *Proc. of Computer Vision and Pattern Recognition (CVPR)*, pages 1542–1549, 2014. 1, 2