Learning 3D Face Reconstruction with a Pose Guidance Network

Pengpeng Liu[1⋆], Xintong Han[2], Michael Lyu[1], Irwin King[1], and Jia Xu[2]

[1] The Chinese University of Hong Kong
[2] Huya AI

Abstract. We present a self-supervised learning approach to learning monocular 3D face reconstruction with a pose guidance network (PGN). First, we unveil the bottleneck of pose estimation in prior parametric 3D face learning methods, and propose to utilize 3D face landmarks for estimating pose parameters. With our specially designed PGN, our model can learn from both faces with fully labeled 3D landmarks and unlimited unlabeled in-the-wild face images. Our network is further augmented with a self-supervised learning scheme, which exploits face geometry information embedded in multiple frames of the same person, to alleviate the ill-posed nature of regressing 3D face geometry from a single image. These three insights yield a single approach that combines the complementary strengths of parametric model learning and data-driven learning techniques. We conduct a rigorous evaluation on the challenging AFLW2000-3D, Florence and FaceWarehouse datasets, and show that our method outperforms the state-of-the-art for all metrics.

1 Introduction

Monocular 3D face reconstruction with precise geometric details serves as a foundation to a myriad of computer vision and graphics applications, including face recognition [1, 2], digital avatars [3, 4], face manipulation [5, 6], *etc.* However, this problem is extremely challenging due to its ill-posed nature, as well as difficulties to acquire accurate 3D face annotations.

Most successful attempts to tackle this problem are built on parametric face models, which usually contain three sets of parameters: identity, expression, and pose. The most famous one is 3D Morphable Model (3DMM) [7] and its variants [5, 8–10]. Recently, CNN-based methods that directly learn to regress the parameters of 3D face models [11–14], achieve state-of-the-art performance.

Are these parameters well disentangled and can they be accurately regressed by CNNs? To answer this question, we conduct a careful study on the AFLW2000-3D dataset [15]. Fig. 1(a) illustrates our setting. We first train a neural network that takes an RGB image as input to simultaneously regress the identity, expression and pose parameters. The `Baseline` 3DMM model is obtained by minimizing the 3D vertex error. Then, we independently replace the predicted identity,

⋆ Work mainly done during an internship at Huya AI.

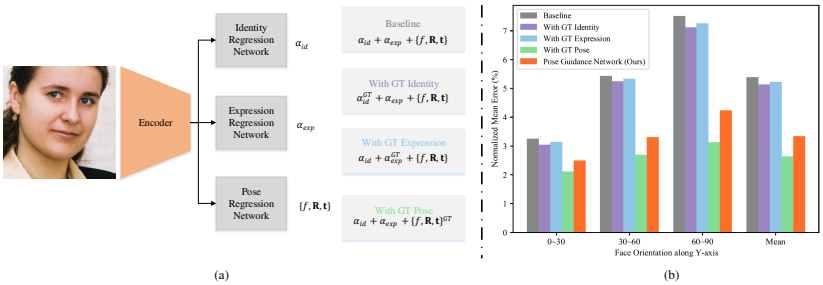


Fig. 1. Our CNN baseline takes an RGB image as input, and regresses identity, expression and pose parameters simultaneously. The three sets of parameters are obtained by minimizing the 3D vertex error. We compute the Normalized Mean Error (NME) of this face model and denote it as `Baseline`. Then we replace the predicted identity, expression, pose parameters with their ground truth, and recompute the NME respectively: `With GT Identity, Expression, Pose`. As shown in (b), `With GT Pose` yields the highest performance gain, and the gain is more significant as the face orientation degree increases. Our `Pose Guidance Network` takes advantage of this finding (Sec. 3.2), and greatly reduces the error caused by inaccurate pose parameter regression.

expression, and pose parameters with their corresponding ground truth parameters (denoted as `GT Identity`, `GT Expression`, and `GT Pose`), and recompute the 3D face reconstruction error shown in Fig. 1(b).

Surprisingly, we found that `GT Pose` yields almost 5 times more performance gain than its two counterparts. The improvement is even more significant when the face orientation degree increases. We posit that there are two reasons causing this result: (1) These three sets of parameters are heavily correlated, and predicting a bad pose will dominate the identity and expression estimation of the 3D face model; (2) 3D face annotations are scarce especially for those with unusual poses.

To address these issues, we propose a pose guidance network (PNG) to isolate the pose estimation from the original 3DMM parameters regression by estimating a UV position map [16] for 3D face landmark vertices. Utilizing the predicted 3D landmarks help to produce more accurate face poses compared to joint parameters regression (*i.e.*, `Baseline` in Fig. 1), and the predicted 3D landmarks also contain valuable identity and expression information that further refines the estimation of identity and expression. Moreover, this enables us to learn from both accurate but limited 3D annotations, and unlimited in-the-wild images with pseudo 2D landmarks (from off-the-shelf landmark extractor like [17]) to predict more accurate 3D landmarks. Consequently, with our proposed PGN, the performance degradation brought by inaccurate pose parameter regression is significantly mitigated as shown in Fig. 1(b).

To further overcome the scarcity of 3D face annotations, we leverage the readily available in-the-wild videos by introducing a novel set of self-consistency loss functions to boost the performance. Given 3D face shapes in multiple frames of

the same subject, we render a new image for each frame by replacing its texture with that of *commonly visible vertices* from other images. Then, by forcing the rendered image to be consistent with the original image in photometric space, optical flow space and semantic space, our network learns to avoid depth ambiguity and predicts better 3D shapes even without explicitly modeling albedo.

We summarize our key contributions as follows:

(1) We propose a PGN to solely predict the 3D landmarks for estimating the pose parameters based on a careful study (Fig. 1). The PGN effectively reduces the error compared to directly regressing the pose parameters and provides informative priors for 3D face reconstruction.

(2) The PGN allows us to utilize both fully annotated 3D landmarks and pseudo 2D landmarks from unlabeled in-the-wild data. This leads to a more accurate landmark estimator and thus helping better 3D face reconstruction.

(3) Built on a visible texture swapping module, our method explores multi-frame shape and texture consistency in a self-supervised manner, while carefully handling the occlusion and illumination change across frames.

(4) Our method shows superior qualitative and quantitative results on ALFW-2000-3D [15], Florence [18] and FaceWarehouse [19] datasets.

2 Related Work

Most recent 3D face shape models are derived from Blanz and Vetter 3D morphable models (3DMM) [7], which represents 3D faces with linear combination of PCA-faces from a collection of 3D face scans. To make 3DMM more representative, Basel Face Model (BFM) [2] improved shape and texture accuracy, and FaceWarehouse [19] constructed a set of individual-specific expression blend-shapes. Our approach is also built on 3DMM — we aim to predict 3DMM parameters to reconstruct 3D faces from monocular frames.

3D Face Landmark Detection and Reconstruction. 3D face landmark detection and 3D face reconstruction are closely related. On the one hand, if the 3DMM parameters can be estimated accurately, face landmark detection can be greatly improved, especially for the occluded landmarks [15]. Therefore, several approaches [15, 20, 21] aligned 3D face by fitting a 3DMM model. On the other hand, if 3D face landmarks are precisely estimated, it can provide strong guidance for 3D face reconstruction. Our method goes towards the second direction—we first estimates 3D face landmarks by regressing UV position map and then utilizes it to guide 3D face reconstruction.

3D Face Reconstruction from a Single Image. To reconstruct 3D faces from a single image, prior methods [5, 1, 22] usually conduct iterative optimization methods to fit 3DMM models by leveraging facial landmarks or local features *e.g.*, color or edges. However, the convergence of optimization is very sensitive to the initial parameters. Tremendous progress has been made by CNNs that directly regress 3DMM parameters [15, 23, 24]. Jackson *et al.* [25] directly regressed the full 3D facial structure via volumetric convolution. Feng *et al.* [16] predicted a UV position map to represent the full 3D shape. MMFace [12] jointly trained

a volumetric network and a parameter regression network, where the former one is employed to refine pose parameters with ICP as a post-processing. All these three methods need to be trained in a supervised manner, requiring full 3D face annotations, which are limited at scale [15]. To bypass the limitation of training data, Tewari *et al.* [26] and Genova *et al.* [27] proposed to fit 3DMM models with only unlabeled images. They show that it is possible to achieve great face reconstruction in an unsupervised manner by minimizing photometric consistency or facial identity loss. Later, Chang *et al.* [28] proposed to regress identity, expression and pose parameters with three independent networks. However, due to depth ambiguity, these unsupervised monocular methods fail to capture precise 3D facial structure. In this paper, we propose to mitigate the limitation of datasets by utilizing both labeled and unlabeled datasets, and to learn better facial geometry from multiple frames.

3D Face Reconstruction from Multiple Images. Multiple images of the same person contain rich information for learning better 3D face reconstruction. Piotraschke *et al.* [29] introduced an automated algorithm that selects and combines reconstructions of different facial regions from multiple images into a single 3D face. RingNet [11] considered shape consistency across different images of the same person, while we focus on face reconstruction from videos, where photometric consistency can be well employed. MVF [14] regressed 3DMM parameters from multi-view images. However, MVF assumes that the expressions in different views are the same, therefore its application is restricted to multi-view images. Our method does not have such constraint and can be applied to both single-view and multi-view 3D face reconstruction.

The approach that is closest to ours is FML [30], which learns face reconstruction from monocular videos by ensuring consistent shape and appearance across frames. However, it only adds multi-frame identity consistency constraints, which does not fully utilize geometric constraints among different images. Unlike FML, we do not model albedo to estimate texture parameters, but directly sample textures from images, swap commonly visible texture and project them onto different image planes while enforcing photometric and semantic consistency. Additionally, we introduce a PGN, which removes the need of pose parameter estimation and enables our model to produce more accurate identity and expression estimation.

3 Method

We illustrate our framework overview in Fig. 2. First, we utilize a shared encoder to extract semantic feature representations from multiple frames of the same person. Then, an identity regression branch and an expression regression branch are employed to regress 3DMM face identity and expression parameters (Sec. 3.1) with the help of our PGN that predicts 3D face landmarks (Sec. 3.2). Finally, we explore self-consistency (Sec. 3.3) with our newly designed consistency losses (Sec. 3.4).

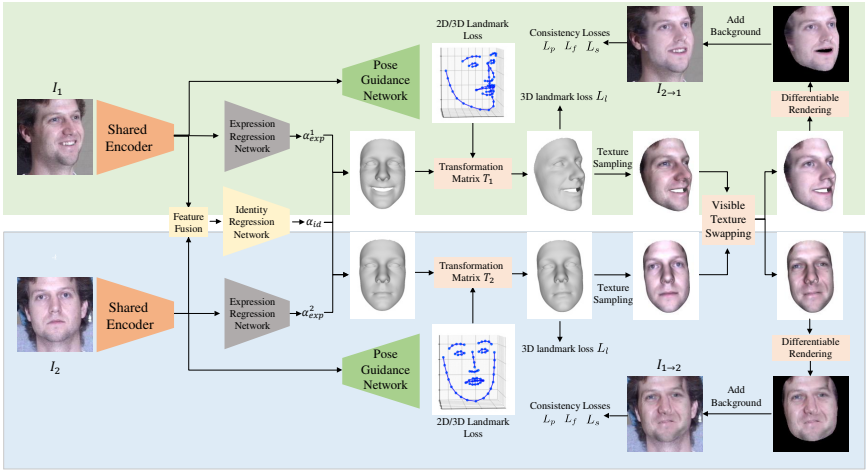


Fig. 2. Framework overview. Our shared encoder extracts semantic feature representation from multiple images of the same person. Then, our identity and expression regression networks regress 3DMM face identity and expression parameters (Sec. 3.1) with accurate guidance of our PGN that predicts 3D face landmarks (Sec. 3.2). Finally, We utilize multiple frames (Sec. 3.3) to train our proposed network with a set of self-consistency loss functions (Sec. 3.4).

3.1 Preliminaries

Let $\mathbf{S} \in \mathbb{R}^{3N}$ be a 3D face with N vertices, $\overline{\mathbf{S}} \in \mathbb{R}^{3N}$ be the mean face geometry, $\mathbf{B}_{id} \in \mathbb{R}^{3N\times199}$ and $\mathbf{B}_{exp} \in \mathbb{R}^{3N\times29}$ be PCA basis of identity and expression, $\boldsymbol{\alpha}_{id} \in \mathbb{R}^{199}$ and $\boldsymbol{\alpha}_{exp} \in \mathbb{R}^{29}$ be the identity and expression parameters. The classical 3DMM face model [7] can be defined as follows:

$$\mathbf{S}(\boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{exp}) = \overline{\mathbf{S}} + \mathbf{B}_{id}\boldsymbol{\alpha}_{id} + \mathbf{B}_{exp}\boldsymbol{\alpha}_{exp}. \tag{1}$$

Here, we adopt BFM [2] to obtain $\overline{\mathbf{S}}$ and $\mathbf{B}_{id}$, and expression basis $\mathbf{B}_{exp}$ is extracted from FaceWareHouse [19]. Then, we employ a perspective projection model to project a 3D face point $\mathbf{s}$ onto an image plane:

$$\mathbf{v}(\boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{exp}) = \begin{bmatrix}1&0&0\\0&1&0\end{bmatrix} \cdot (f \cdot \mathbf{R} \cdot \mathbf{s} + \mathbf{t}) = \begin{bmatrix}1&0&0\\0&1&0\end{bmatrix} \cdot \begin{bmatrix}f \cdot \mathbf{R} & \mathbf{t}\end{bmatrix} \cdot \begin{bmatrix}\mathbf{s}\\1\end{bmatrix}, \tag{2}$$

where $\mathbf{v}$ is the projected point on the image plane, f is a scaling factor, $\mathbf{R} \in \mathbb{R}^{3\times3}$ indicates a rotation matrix, $\mathbf{t} \in \mathbb{R}^3$ is a translation vector.

However, it is challenging for neural networks to regress identity parameter $\boldsymbol{\alpha}_{id}$, expression parameter $\boldsymbol{\alpha}_{exp}$ and pose parameter $\{f, \mathbf{R}, \mathbf{t}\}$ together, because these parameters cannot be easily disentangled and pose parameters turn to dominate the optimization, making it more difficult to estimate accurate identity and expression (as discussed in Sec. 1 and illustrated in Fig. 1).

To address this issue, we design a robust landmark-based PGN to obtain the transformation matrix $\mathbf{T} = \begin{bmatrix} f \cdot \mathbf{R} & \mathbf{t} \end{bmatrix}$ instead of directly regressing its parameters. Next, we describe our PGN in detail.

3.2 Pose Guidance Network

To decouple the optimization of pose parameter $\{f, \mathbf{R}, \mathbf{t}\}$ with identity parameter $\boldsymbol{\alpha}_{id}$ and expression parameter $\boldsymbol{\alpha}_{exp}$, we design a multi-task architecture with two output branches (shown in Fig. 2). One branch optimizes the traditional 3DMM identity and expression parameters $\boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{exp}$. The other branch is trained to estimate a UV position map [16] for 3D face landmarks, which provide key guidance for pose estimation.

Specifically, Let $\mathbf{X}$ be the 3D landmark positions in the face geometry $\mathbf{S}$, and $\mathbf{X}_{UV}$ be the 3D landmarks estimated from our UV position map decoder, we estimate a transformation matrix $\mathbf{T}$ by,

$$\min_{\mathbf{T}} ||\mathbf{T} \cdot \begin{bmatrix} \mathbf{X} \\ \mathbf{1} \end{bmatrix} - \mathbf{X}_{UV}||_2. \tag{3}$$

Here, $\mathbf{T}$ has a closed-form solution:

$$\mathbf{T} = \mathbf{X}_{UV} \cdot \begin{bmatrix} \mathbf{X} \\ \mathbf{1} \end{bmatrix}^T \cdot \Big(\begin{bmatrix} \mathbf{X} \\ \mathbf{1} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{X} \\ \mathbf{1} \end{bmatrix}^T \Big)^{-1}. \tag{4}$$

As a result, we convert the estimation of $\mathbf{T}$ into the estimation of a UV position map for 3D face landmarks rather than regressing $\mathbf{T}$'s parameters. This disentangles the pose estimation and results in better performance than joint regression of $\boldsymbol{\alpha}_{id}, \boldsymbol{\alpha}_{exp}$and $\{f, \mathbf{R}, \mathbf{t}\}$. Another merit of this design is enabling us to train our network with two types of images: images with 3D landmark annotations and in-the-wild unlabeled images with 2D facial landmarks extracted by off-the-shelf detectors. During training, we sample one image batch with 3D landmark labels and another image batch from unlabeled datasets. 3D landmark loss and 2D landmark loss are minimized for them, respectively. For 3D landmarks, we calculate the loss across all x, y and z channels of the UV position map, while for 2D landmark loss, only x and y channels are considered. More abundant training data leads to more accurate pose estimation, and hence better face reconstruction.

Note our work is different from PRN [16], which utilizes a CNN to regress dense UV position maps for all 3D face points. PRN requires dense 3D face labels which are extremely difficult to obtain. Our network learns directly from sparse landmark annotations, which are much easier to obtain and more accurate than the synthetic data derived from facial landmarks.

3.3 Learning from Multiple Frames

The PGN combined with identity and expression parameters regression can achieve quite accurate 3D face reconstruction, but the estimated 3D mesh lacks

facial details. This is because 3D landmarks can only provide a coarse prediction of identity and expression. To generate meshes with finer details, we leverage multi-frame images from monocular videos as input and explore their inherent complementary information. In contrast to the common perspective that first estimates albedo maps and then enforces photometric consistency [30], we propose a self-consistency framework based on a visible texture swapping scheme.

Every vertex in a 3DMM model has a specific semantic meaning. Given multiple images with the same identity, we can generate one 3D mesh for every image. Every corresponding vertex of different meshes share the same semantic meaning, even though these images are captured with different poses, expressions, lightings, *etc*. If we sample texture from one image and project it onto the second image that has different pose and expression, the rendered image should have the same identity, expression and pose as the second image despite the illumination change. Our multi-image 3D reconstruction is built on this intuition.

More specifically, our method takes multiple frames of the same subject as input, and estimates the same set of identity parameters for all images, and different expressions and poses (obtained from 3D face landmarks output by our PGN) for each image. To generate the same identity parameters, we adopt a similar strategy as [30], which fuses feature representations extracted from the shared encoders of different images via average pooling (Feature Fusion in Fig. 2). In this way, we can achieve both single-image and multi-image face reconstruction.

For simplicity, we assume there are two images of the same person as input (the framework can easily extend to more than two images), denoted as I_1 and I_2 respectively. Then, as illustrated on the left side of Fig. 2, we can generate two 3D meshes with the same identity parameter $\boldsymbol{\alpha}_{id}$, two different expression parameters $\boldsymbol{\alpha}^1_{exp}, \boldsymbol{\alpha}^2_{exp}$, and pose transformation matrices $\mathbf{T}_1, \mathbf{T}_2$ obtained by our PGN. After that, we sample two texture maps C_1, C_2 with Equation 2, and project the first texture C_1 onto the second image I_2 with its expression parameter $\boldsymbol{\alpha}^2_{exp}$ and pose transformation matrix $\mathbf{T}_2$ to obtain rendered image $I_{1\to 2}$. Similarly, we can project C_2 to I_1 to obtain the rendered image $I_{2\to 1}$. Ideally, if there is no illumination change, I_2 shall be the same as $I_{1\to 2}$ over their non-occluded facial regions. However, there exists occlusion and illumination usually changes a lot for different images in real-world scenarios. To this end, we introduce several strategies to overcome these issues.

Occlusion Handling. We adopt a simple strategy to effectively determine if a pixel is occluded or non-occluded based on triangle face normal direction. Given a triangle with three vertices, we can compute its normal $\mathbf{n} = (n_x, n_y, n_z)$. If the normal direction towards outside of the face mesh (*i.e.*, $n_z > 0$), we regard these three vertices as non-occluded; otherwise they are occluded. According to this principle, we can compute two visibility maps M_1 and M_2, where value 1 indicates the vertex is non-occluded and 0 otherwise. A common visibility map M_{12} is then defined as:

$$M_{12} = M_1 \odot M_2, \tag{5}$$

where value 1 means that the vertex is non-occluded for both 3D meshes.

Considering the occlusion, when projecting C_1 onto the second image, we combine C_1 and C_2 by

$$C_{1\to 2} = C_1 \odot M_{12} + C_2 \odot (1 - M_{12}). \tag{6}$$

That is, we alleviate the influence of the occlusion by only projecting the *commonly visible texture* from I_1 to I_2 to generate $C_{1\to 2}$, while *keeping the original pixels for the occluded part*. In this way, the rendered image $I_{1\to 2}$ shall have the same identity, pose and expression information as I_2. The projection from I_2 to I_1 can be derived in the same manner.

Illumination Change. The sampled texture is not disentangled to albedo, lighting, etc. Due to lighting and exposure changes, even if we can estimate accurate 3D geometry, the rendered texture $I_{1\to 2}$ is usually different from I_2. To cope with these issues, we propose three schemes. First, we adopt the Census Transform [31] from optical flow estimation, which has been shown to be very robust to illumination change when computing photometric difference (Eq. 9). Specifically, we apply a 7×7 census transform and then compute the Hamming distance between the reference image I_2 and the rendered image $I_{1\to 2}$. Second, we employ an optical flow estimator [32] to compute the flow between I_2 and the rendered image $I_{1\to 2}$. Since optical flow provides a 2D dense correspondence constraint, if the face is perfectly aligned, the optical flow between I_2 and $I_{1\to 2}$ should be zeros for all pixels, so we try to minimize difference, *i.e.*, minimize the magnitude of optical flow between them (Eq. 10). Third, even though illumination changes, the identity, expression and pose shall be the same for I_2 and $I_{1\to 2}$. Therefore, they must share similar semantic feature representation. Since our shared encoder can extract useful information to predict facial landmarks, identity and expression parameters, we use it as a semantic feature extractor and compare the feature difference between I_2 and $I_{1\to 2}$ (Eq. 11).

3.4 Training Loss

To train our network for accurate 3D face reconstruction, we define a set of self-consistency loss functions, and minimize the following combination:

$$L = L_l + L_p + L_f + L_s + L_r. \tag{7}$$

Each loss term is defined in detail as follows. Note that for simplicity, we only describe these loss terms regarding projecting I_1 to I_2 (*i.e.*, $I_{1\to 2}$) and the other way around ($I_{2\to 1}$) can be defined similarly.

Sparse Landmark Loss. Our landmark loss measures the difference between the landmarks of transformed face geometry $\mathbf{T} \cdot \mathbf{X}$ and the prediction of PGN $\mathbf{X}_{UV}$:

$$L_l = \lambda_l \sum |\mathbf{T} \cdot \mathbf{X} - \mathbf{X}_{UV}| \tag{8}$$

This is the core guidance loss, which is trained with both 3D and 2D landmarks.

Photometric Consistency Loss. Photometric loss measures the difference between the target image and the rendered image over those visible regions.

We can obtain the visible mask M^{2d} on the image plane with differentiable mesh render [27]. Note that M^{2d} is different from the vertex visibility map M, where the former denotes whether the pixel is occluded on the image plane, and the latter denotes whether the vertex in 3D mesh is occluded. Besides, considering that most of the face regions have very similar color, we apply a weighted mask W to the loss function, where we emphasize eye, nose, and mouth regions with a larger weight of 5, while the weight is 1 for other face regions [16]. The photometric loss then writes:

$$L_p = \lambda_p \frac{\sum \text{Hamming}|\text{Census}(I_2) - \text{Census}(I_{1\to 2})| \odot M_2^{2d} \odot W}{\sum M_2^{2d} \odot W}, \tag{9}$$

where `Census` represents the census transform, `Hamming` denotes Hamming distance, and M_2^{2d} is the corresponding visibility mask.

Flow Consistency Loss. We use optical flow to describe the dense correspondence between the target image and the rendered image, then the magnitude of optical flow is minimized to ensure the visual consistency between two images:

$$L_f = \lambda_f \sum |\mathbf{w}(I_2, I_{1\to 2})| \odot W / \sum W, \tag{10}$$

where $\mathbf{w}$ is the optical flow computed from [32] and the same weighted mask W is applied as in the photometric consistency loss.

Semantic Consistency Loss. Photometric loss and 2D correspondence loss may break when the illumination between two images changes drastically. However, despite the illumination changes, I_2 and $I_{1\to 2}$ should share the same semantic feature representation, as the target image and the rendered image share the same identity, expression and pose. To this end, we minimize the cosine distance between our semantic feature embeddings:

$$L_s = \lambda_s - \lambda_s < \frac{F(I_2)}{||F(I_2)||_2}, \frac{F(I_{1\to 2})}{||F(I_{1\to 2})||_2} >, \tag{11}$$

where F denotes our shared feature encoder. Unlike existing approaches (*e.g.*, [27]) which align semantic features in a pre-trained face recognition network, we simply minimize the feature distance from our learned shared encoder. We find that this speeds up our training process and empirically works better.

Regularization Loss. Finally, we add a regularization loss to identity and expression parameters to avoid over-fitting:

$$L_r = \lambda_r \sum_{i=1}^{199} |\frac{\boldsymbol{\alpha}_{id}(i)}{\sigma_{id}(i)}| + \frac{\lambda_r}{2} \sum_{i=1}^{29} |\frac{\boldsymbol{\alpha}_{exp}(i)}{\sigma_{exp}(i)}|, \tag{12}$$

where σ_{id} and σ_{exp} represent the standard deviation of $\boldsymbol{\alpha}_{id}$ and $\boldsymbol{\alpha}_{exp}$.

4 Experimental Evaluation

Training Datasets. To train the shared encoder and PGN, we utilize two types of datasets: synthetic dataset with pseudo 3D annotations and in-the-wild

(a) 2D NME on AFLW2000-3D dataset

Method	NME_{2d}^{68}			
	0 to 30	30 to 60	60 to 90	Mean
SDM[33]	3.67	4.94	9.67	6.12
3DDFA [15]	3.78	4.54	7.93	5.42
3DDFA + SDM [15]	3.43	4.24	7.17	4.94
Yu et al. [34]	3.62	6.06	9.56	-
3DSTN[35]	3.15	4.33	5.98	4.49
DeFA[20]	-	-	-	4.50
Face2Face [5]	3.22	8.79	19.7	10.5
3DFAN [17]	2.77	3.48	4.61	3.62
PRN [16]	2.75	3.51	4.61	3.62
ExpNet [13]	4.01	5.46	6.23	5.23
MMFace-PMN [12]	5.05	6.23	7.05	6.11
MMFace-ICP-128 [12]	2.61	3.65	4.43	3.56
Ours (PGN)	**2.49**	**3.30**	4.24	**3.34**
Ours (3DMM)	2.53	3.32	**4.21**	3.36

(b) 3D NME on AFLW2000-3D dataset

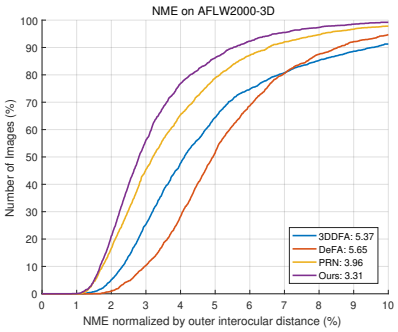


Fig. 3. Performance comparison on AFLW2000-3D. (a) **2D landmarks**. The NME (%) for 68 2D landmarks with different face orientation along the Y-axis are reported. (b) **3D face reconstruction**. X-axis denotes the NME normalized by outer interocular distance, the Y-axis denotes the percentage of images. Following [16], around 45k points are used for evaluation.

datasets. For synthetic dataset, we choose 300W-LP [15], which contains 60k synthetic images with fitted 3DMM parameters. These images are synthesized from around 4k face images with face profiling synthetic method [36]. To enable more robust 3D face landmark detection, we choose a corpus of in-the-wild datasets, including Menpo [37], CelebA [38], 300-VW [39] and Multi-PIE [40] with their 68 2D landmarks automatically extracted by [17].

To train identity and expression regression networks with our proposed self-consistency losses, we utilize 300-VW [39] and Multi-PIE [40], where the former contains monocular videos, and the latter contains faces images of the same identity under different lightings, poses, expressions and scenes.

Evaluation Datasets and Metrics. We evaluate our model on AFLW-2000-3D [15], Florence [18] and FaceWarehouse [19] datasets. AFLW-2000-3D contains the first 2000 images from AFLW [41], which is annotated with fitted 3DMM parameters and 68 3D landmarks in the same way as 300W-LP. We evaluate face landmark detection performance and 3D face reconstruction performance on this dataset, which is measured by Normalized Mean Error (NME). Florence dataset contains 53 subjects with ground truth 3D scans, where each subject contains three corresponding videos: "Indoor-Cooperative", "PTZ-Indoor" and "PIZ-Outdoor". We report Point-to-Plane Distance to evaluate 3D shape reconstruction performance. The Florence dataset only contains 3D scans with the neutral expression, which can only be used to evaluate the performance of shape reconstruction. To evaluate the expression part, we further evaluate our method on the FaceWarehouse dataset. Following previous work [26, 42, 43, 30], we use a subset with 180 meshes (9 identities and 20 expressions each) and report per-vertex error. Florence and FaceWarehouse are also employed to verify the effectiveness of our proposed multi-frame consistency scheme.

Training Details. The face regions are cropped according to either pseudo 3D face landmarks or detected 2D facial landmarks [17]. Then the cropped images are resized to 256×256 as input. The shared encoder and PGN structures are the same as PRN [16]. For PGN, another option is using fully connected layers to regress sparse 3D landmarks, which can reduce a lot of computation with slightly decreased performance. The identity and expression regression networks take the encoder output as input, followed by one convolutional layer, one average pooling layer and three fully-connected layers.

Table 1. Comparison of mean point-to-plane error on the Florence dataset.

Method	Indoor-Cooperative		PTZ-Indoor	
	Mean	Std	Mean	Std
Tran *et al.*[24]	1.443	0.292	1.471	0.290
Tran *et al.*+ pool	1.397	0.290	1.381	0.322
Tran *et al.*+ [29]	1.382	0.272	1.430	0.306
MoFA [26]	1.405	0.306	1.306	0.261
MoFA + pool	1.370	0.321	1.286	0.266
MoFA + [29]	1.363	0.326	1.293	0.276
Genova *et al.*[27]	1.405	0.339	1.271	0.293
Genova *et al.*+ pool	1.372	0.353	1.260	0.310
Genova *et al.*+ [29]	1.360	0.346	1.246	0.302
MVF [14] - pretrain	1.266	0.297	1.252	0.285
MVF [14]	1.220	0.247	1.228	0.236
Ours	**1.122**	**0.219**	**1.161**	**0.224**

Our whole training procedure contains 3 steps: (1) We first train the shared encoder and PGN. We randomly sample one batch images from 300W-LP and another batch from in-the-wild datasets, then employ 3D landmark and 2D landmark supervision respectively. We set batch size to 16 and train the network for 600k iterations. After that, both the shared encoder and PGN parameters are fixed. (2) For identity and expression regression networks, we first pre-train them with only one image for each identity as input using L_l and L_r for 400k iterations. This results in a coarse estimation and speeds up the convergence for training with multiple images. (3) Finally, we sequentially choose 2 and 4 images for each identity as input and train for another 400k iterations by minimizing Eq. (7). The balance weights for loss terms are set to $\lambda_l = 1$, $\lambda_p = 0.2$, $\lambda_f = 0.2$, $\lambda_s = 10$, $\lambda_r = 1$. Due to the memory consumption brought by rendering and optical flow estimation, we reduce the batch size to 4 for multi-image input. All 3 steps are trained using Adam [44] optimizer with an initial learning rate of 10^{-4}. Learning rate decays half after 100k iterations.

3D Face Alignment Results. Fig. 3(a) shows the 68 facial landmark detection performance on AFLW2000-3D dataset [15]. By training with a large corpus of unlabeled in-the-wild data, our model greatly improves over previous state-of-the-art 3D face alignment methods (*e.g.*PRN [16], MMFace [12]) that heavily rely on 3D annotations. Our method achieves the best performance without any post-processing such as the ICP used in MMFace. Moreover, our PGN is robust. We can fix it and directly use its output as ground truth of 3D landmarks to guide the learning of 3D face reconstruction.

Quantitative 3D Face Reconstruction Results. We evaluate 3D face reconstruction performance with NME on AFLW2000-3D, Point-to-Plane error on Florence and Per-vertex error on FaceWarehouse. Thanks to the robustness of our PGN, we can directly fix it and obtain accurate pose estimation without further learning. Then, our model can focus more on shape and expression

Table 2. Per-vertex geometric error (measured in mm) on FaceWarehouse dataset. PGN denotes PGN. Our approach obtains the lowest error, outperforming the best prior art [30] by 7.5%.

Method	MoFA [26]	Inversefacenet [45]	Tewari *et al.* [42]	FML [30]	Ours Single-Frame without PGN	Ours Single-Frame with PGN	Ours Mult-Frame without PGN	Ours Multi-Frame with PGN
Error	2.19	2.11	2.03	2.01	2.18	2.09	1.98	**1.86**

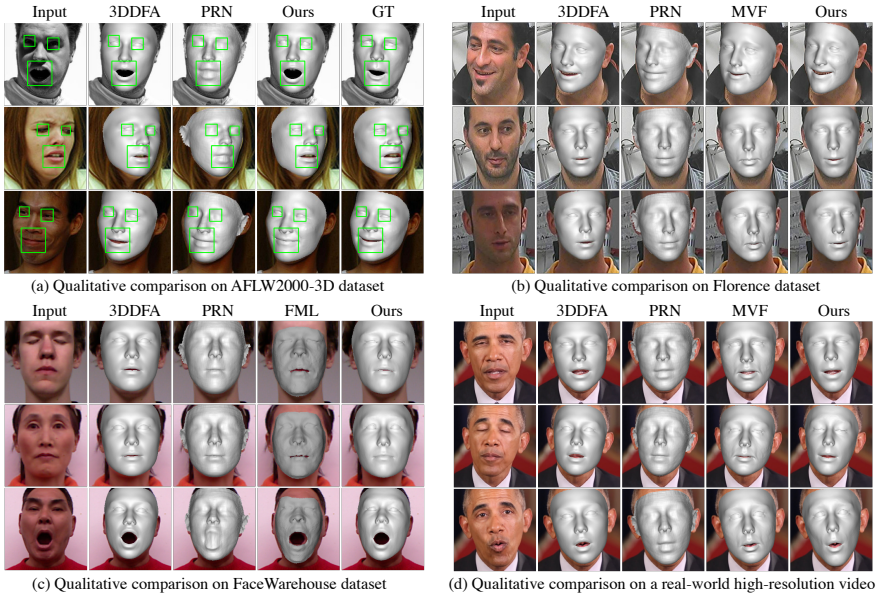
estimation. As shown in Fig. 3(b), we achieve the best results on the AFLW2000-3D dataset, reducing NME_{3d} of previous state-of-the-art from 3.96 to 3.31, with 16.4% relative improvement.

Table 1 shows the results on the Florence dataset. In contrast to MVF that concatenates encoder features as input to estimate a share identity parameter, we employ average pooling for encoder features, enabling us to perform both single-image and multi-image face reconstruction. In the evaluation setting, it does not make much difference using single-frame or multi-frame as input, because we'll finally average all the video frame output. Notably, our method is more general than the previous state-of-the-art MVF that assumes expressions are the same among multiple images (*i.e.*, multi-view images), while our method can directly train on monocular videos.

Table 2 shows the results on FaceWarehouse dataset. For single frame setting, without modeling albedo, we still achieve comparable performance with MoFA [26], Inversefacenet [45] and Tewari *et al.* [42]. For multi-frame settings, we achieve better results than FML [30]. For both single-frame and multi-frame settings, we achieve improved performance with PGN. All these show the effectiveness of PGN and self-consistency losses.

Qualitative 3D Face Reconstruction Results. Fig. 4(a) shows the qualitative comparisons with 3DDFA [15], PRNet [16] and the pseudo ground truth. 3DDFA regresses identity, expression and pose parameters together and is only trained with synthetic datasets 300W-LP, leading to performance degradation. The estimated shape and expression of 3DDFA is close to mean face geometry and looks generally similar. PRNet directly regresses all vertices stored in UV position map, which cannot capture the geometric constraints well; thus, it does not look smooth and lacks geometric details, *e.g.*, eye and mouth regions. In contrast, our estimated shape and expression looks visually convincing. Even when compared with the pseudo ground truth generated with traditional matching methods, our estimation is more accurate in many cases. Fig. 4(b) shows the comparison on the Florence dataset, which further demonstrates the effectiveness of our method. Compared with FML on FaceWarehouse dataset, our results can generate more accurate expressions with visibly pleasing face reconstruction results (Fig. 4 (c)).

Ablation Study. The effectiveness of PGN has been shown in Fig 3 (a) (for face alignment) and Table 2 (for face reconstruction). To better elaborate the contributions of different components in our self-consistency scheme, we perform detailed ablation study in Fig 5.



(a) Qualitative comparison on AFLW2000-3D dataset

(b) Qualitative comparison on Florence dataset

(c) Qualitative comparison on FaceWarehouse dataset

(d) Qualitative comparison on a real-world high-resolution video

Fig. 4. Qualitative Comparison on various datasets. Our model generates more accurate shapes and expressions, especially around the mouth and eye region, as we leverage unlimited 2D face data and cross image consistency. The estimated shape of 3DDFA is close to mean face geometry and the results of PRN lack geometric details. (a) On **AFLW2000-3D**, our results look even more visually convincing than ground truth in many cases. (b) **Florence**. (c) **FaceWarehouse**. Compared with FML, our results are more smooth and visibly pleasing. (d) **Video results**. Our consistency losses work especially well for high resolution images with few steps of fine-tuning. We generate accurate shape and expression, *e.g.*, challenging expression of complete eye-closing. **Zoom in for details.**

Our baseline model is single-image face reconstruction trained only with L_l and L_r. However, it doesn't lead to accurate shape estimation, because our PGN with sparse landmarks can only provide a coarse shape estimation. To better estimate the shape, we employ multi-frame images as input. As shown in Fig. 5(a), even without census transform, the photometric consistency (L_{p-}) improves the performance. However, photometric loss does not work well when illumination changes among video frames. Therefore, we enhance the photometric loss with census transform to make the model more robust to illumination change. This improves the performance quantitatively (Fig. 5(a)), and qualitatively (Fig. 5(b-f)). Applying semantic consistency (L_s) and flow consistency (L_f) enforces the rendered image and the target image to look semantically similar and generates better face geometry.

Video Results. Our proposed multi-image face reconstruction method is based on texture sampling, then it shall obtain better face reconstruction results with higher texture quality (higher video resolution). To verify it, we fine-tune

(a) **Ablation study on Florence.**

L_{p-}	L_p	L_s	L_f	Indoor-Cooperative Mean	Indoor-Cooperative Std	PTZ-Indoor Mean	PTZ-Indoor Std
✗	✗	✗	✗	1.364	0.352	1.379	0.326
✓	✗	✗	✗	1.263	0.312	1.323	0.251
✗	✓	✗	✗	1.219	0.261	1.255	0.256
✗	✓	✗	✓	1.193	0.230	1.221	0.247
✗	✓	✓	✗	1.161	0.268	1.269	0.276
✗	✓	✓	✓	**1.122**	**0.219**	**1.161**	**0.224**

(b) Input (c) Pretrain (d) L_{p-} (e) L_p (f) Full

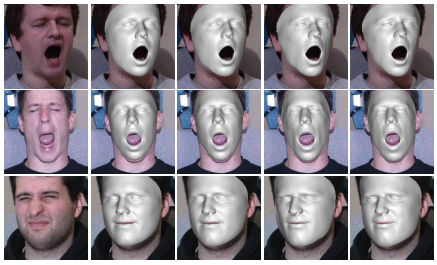


Fig. 5. (a) Ablation Study on Florence dataset. L_{p-} indicates that census transform is not applied when computing photometric differences. We find that census transform is robust for illumination variations. **(b-f) Visual results of our ablations on Multi-PIE dataset:** (b) Input image. (c) Pre-trained model with only landmark loss and regularizer loss. (d) Employ photometric loss. (e) Employ census transform when computing photometric consistency. (f) Full loss. We can find that key components of our model improve the accuracy of shape and expression. **Zoom in for details.**

our model on a high-quality video from the Internet, *i.e.*, the fine-tuned model is specialized for the video. No 3D ground truth is used here. As shown in Fig. 4(d), our estimated shape and expression look surprisingly accurate after several thousand iterations. Specifically, our model captures the detailed expression (*e.g.*, totally closed eyes) and face shape very well. This can be an interesting application when we need to obtain accurate 3D face reconstruction for one specific person.

5 Conclusion

We have presented a pose guidance network which yields superior performance on 3D face reconstruction from a single image or multiple frames. Our approach effectively makes use of in-the-wild unlabeled images and provides accurate 3D landmarks as an intermediate supervision to help reconstruct 3D faces. Furthermore, we have demonstrated that swapping textures of multiple images and exploring their photometric and semantic consistency greatly improve the final performance. We hope that our work can inspire future research to develop new techniques that leverage informative intermediate representations (*e.g.*, 3D landmarks in this paper) and learn from unlabeled images or videos.

Acknowledgement

This work was partially supported by the RRC of the Hong Kong Special Administrative Region (No. CUHK 14210717 of the General Research Fund) and National Key Research and Development Program of China (No. 2018AAA0100204). We also thank Yao Feng, Feng Liu and Ayush Tewari for kind help.

References

1. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. TPAMI (2003)
2. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: AVSS. (2009)
3. Nagano, K., Seo, J., Xing, J., Wei, L., Li, Z., Saito, S., Agarwal, A., Fursund, J., et al.: pagan: real-time avatars using dynamic textures. In: SIGGRAPH Asia. (2018)
4. Hu, L., Saito, S., Wei, L., Nagano, K., Seo, J., Fursund, J., Sadeghi, I., Sun, C., Chen, Y.C., Li, H.: Avatar digitization from a single image for real-time rendering. TOG (2017)
5. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: CVPR. (2016)
6. Kim, H., Carrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. TOG (2018)
7. Blanz, V., Vetter, T., et al.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH. (1999)
8. Saito, S., Li, T., Li, H.: Real-time facial segmentation and performance capture from rgb input. In: ECCV. (2016)
9. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. TOG (2014)
10. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3d shape regression for real-time facial animation. TOG (2013)
11. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: CVPR. (2019)
12. Yi, H., Li, C., Cao, Q., Shen, X., Li, S., Wang, G., Tai, Y.W.: Mmface: A multi-metric regression network for unconstrained face reconstruction. In: CVPR. (2019)
13. Chang, F.J., Tran, A.T., Hassner, T., Masi, I., Nevatia, R., Medioni, G.: Expnet: Landmark-free, deep, 3d facial expressions. In: FG. (2018)
14. Wu, F., Bao, L., Chen, Y., Ling, Y., Song, Y., Li, S., Ngan, K.N., Liu, W.: Mvf-net: Multi-view 3d face morphable model regression. In: CVPR. (2019)
15. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: CVPR. (2016)
16. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: ECCV. (2018)
17. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: ICCV. (2017)
18. Bagdanov, A.D., Del Bimbo, A., Masi, I.: The florence 2d/3d hybrid face dataset. In: Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding. (2011)
19. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. TVCG (2013)
20. Liu, Y., Jourabloo, A., Ren, W., Liu, X.: Dense face alignment. In: ICCV. (2017)
21. Gou, C., Wu, Y., Wang, F.Y., Ji, Q.: Shape augmented regression for 3d face alignment. In: ECCV. (2016)
22. Romdhani, S., Vetter, T.: Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In: CVPR. (2005)

23. Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3d face reconstruction with deep neural networks. In: CVPR. (2017)
24. Tuan Tran, A., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3d morphable models with a very deep neural network. In: CVPR. (2017)
25. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In: ICCV. (2017)
26. Tewari, A., Zollhofer, M., Kim, H., Garrido, P., Bernard, F., Perez, P., Theobalt, C.: Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: ICCV. (2017)
27. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression. In: CVPR. (2018)
28. Chang, F.J., Tran, A.T., Hassner, T., Masi, I., Nevatia, R., Medioni, G.: Deep, landmark-free fame: Face alignment, modeling, and expression estimation. IJCV (2019)
29. Piotraschke, M., Blanz, V.: Automated 3d face reconstruction from multiple images using quality measures. In: CVPR. (2016)
30. Tewari, A., Bernard, F., Garrido, P., Bharaj, G., Elgharib, M., Seidel, H.P., Pérez, P., Zollhofer, M., Theobalt, C.: Fml: face model learning from videos. In: CVPR. (2019)
31. Hafner, D., Demetz, O., Weickert, J.: Why is the census transform good for robust optic flow computation? In: SSVM. (2013)
32. Liu, P., King, I., Lyu, M.R., Xu, J.: Ddflow: Learning optical flow with unlabeled data distillation. AAAI (2019)
33. Xiong, X., De la Torre, F.: Global supervised descent method. In: CVPR. (2015)
34. Yu, R., Saito, S., Li, H., Ceylan, D., Li, H.: Learning dense facial correspondences in unconstrained images. In: ICCV. (2017)
35. Bhagavatula, C., Zhu, C., Luu, K., Savvides, M.: Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In: ICCV. (2017)
36. Zollhöfer, M., Thies, J., Garrido, P., Bradley, D., Beeler, T., Pérez, P., Stamminger, M., Nießner, M., Theobalt, C.: State of the art on monocular 3d face reconstruction, tracking, and applications. In: Computer Graphics Forum, Wiley Online Library (2018)
37. Deng, J., Roussos, A., Chrysos, G., Ververas, E., Kotsia, I., Shen, J., Zafeiriou, S.: The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. IJCV (2019)
38. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV. (2015)
39. Shen, J., Zafeiriou, S., Chrysos, G.G., Kossaifi, J., Tzimiropoulos, G., Pantic, M.: The first facial landmark tracking in-the-wild challenge: Benchmark and results. In: ICCVW. (2015)
40. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing (2010)
41. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: ICCVW. (2011)
42. Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In: CVPR. (2018)

43. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: CVPR. (2018)
44. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR. (2015)
45. Kim, H., Zollhöfer, M., Tewari, A., Thies, J., Richardt, C., Theobalt, C.: Inverse-facenet: Deep monocular inverse face rendering. In: CVPR. (2018)