Self-Supervised Monocular 3D Face Reconstruction by Occlusion-Aware Multi-view Geometry Consistency

Jiaxiang Shang[1 [0000−0001−7161−9765]], Tianwei Shen[1 [0000−0002−3290−2258]], Shiwei li[1 [0000−0003−0712−0059]], Lei Zhou[1 [0000−0003−4988−5084]], Mingmin Zhen[1 [0000−0002−8180−1023]], Tian Fang[2 [0000−0002−5871−3455]], and Long Quan[1 [00000001−8148−1771]]

[1] Hong Kong University of Science and Technology
{jshang,tshenaa,lzhouai,mzhen,quan}@cse.ust.hk
[2] Everest Innovation Technology
{sli,fangtian}@altizure.com

Abstract. Recent learning-based approaches, in which models are trained by single-view images have shown promising results for monocular 3D face reconstruction, but they suffer from the ill-posed face pose and depth ambiguity issue. In contrast to previous works that only enforce 2D feature constraints, we propose a self-supervised training architecture by leveraging the multi-view geometry consistency, which provides reliable constraints on face pose and depth estimation. We first propose an occlusion-aware view synthesis method to apply multi-view geometry consistency to self-supervised learning. Then we design three novel loss functions for multi-view consistency, including the pixel consistency loss, the depth consistency loss, and the facial landmark-based epipolar loss. Our method is accurate and robust, especially under large variations of expressions, poses, and illumination conditions. Comprehensive experiments on the face alignment and 3D face reconstruction benchmarks have demonstrated superiority over state-of-the-art methods. Our code and model are released in `https://github.com/jiaxiangshang/MGCNet`.

Keywords: 3D Face Reconstruction, Multi-view geometry consistency

1 Introduction

3D face reconstruction is extensively studied in the computer vision community. Traditional optimization-based methods [2,4,10,27,41–43] formulate the 3D Morphable Model (3DMM) [7] parameters into a cost minimization problem, which is usually solved by expensive iterative nonlinear optimization. The supervised CNN-based methods [16–18, 24, 31, 32, 39, 49, 53, 59] require abundant 3D face scans and corresponding RGB images, which are limited in amount and expensive to acquire. Methods that focus on face detail reconstruction [12, 21, 49, 53, 54, 62] need even high-quality 3D faces scans. To address the insufficiency of scanned 3D face datasets, some unsupervised or self-supervised methods are proposed

[15, 22, 40, 50–52, 54, 55, 61, 66], which employ the 2D facial *landmark loss* between inferred 2D landmarks projected from 3DMM and the ground truth 2D landmarks from images, as well as the *render loss* between the rendered images from 3DMM and original images. One critical drawback of existing unsupervised methods is that both *landmark loss* and *render loss* are measured in projected 2D image space and do not penalize incorrect face pose and depth value of 3DMM, resulting in the ambiguity issue of the 3DMM in the face pose and depth estimation.

To address this issue, we resort to the multi-view geometry consistency. Multi-view images not only contain 2D landmarks and pixel features but also they form the multi-view geometry constraints. Such training data is publicly available and efficient to acquire (*e.g.*,videos). Fortunately, a series of multi-view 3D reconstruction techniques named view synthesis [13, 14, 19] help to formulate self-supervised learning architecture based on multi-view geometry. View synthesis is a classic task that estimates proxy 3D geometry and establishes pixel correspondences among multi-view input images. Then they generate $N-1$ synthetic target view images by compositing image patches from the other $N-1$ input view images. View synthesis is commonly used in Monocular Depth Estimation (MDE) task [11, 34, 63, 65]. However, MDE only predicts depth map and relative poses between views without inferring camera intrinsics. The geometry of MDE is incomplete as MDE loses the relationship from 3D to 2D, and they can not reconstruct a full model in scene. MDE also suffers from erroneous penalization due to self-occlusion.

Inspired by multi-view geometry consistency, we propose a self-supervised Multi-view Geometry Consistency based 3D Face Reconstruction framework (MGCNet). The workflow of MGCNet is shown in Figure 1. I_t is always considered to be the **target** of multi-view data. To simplify the following formulation, we denote all $N-1$ views adjacent to the **target view** as the **source views**. To build up the multi-view consistency in the training process via view synthesis, we first design a covisible map that stores the mask of covisible pixels for each target-source view pair to solve self-occlusion, as the large and extreme face pose cases is common in the real world, and the self-occlusion always happens in such profile face pose cases. Secondly, we feed the 3DMM coefficients and face poses to the differentiable rendering module [22], producing the rendered image, depth map, and covisible map for each view. Thirdly, pixel consistency loss and depth consistency loss are formulated by input images and rendered depth maps in covisible regions, which ensures the consistency of 3DMM parameters in the multi-view training process. Finally, we introduce the facial epipolar loss, which formulates the epipolar error of 2D facial landmarks via the relative pose of two views, as facial landmarks is robust to illumination changes, scale ambiguity, and calibration errors. With these multi-view supervised losses, we are able to achieve accurate 3D face reconstruction and face alignment result on multiple datasets [3, 36, 60, 64, 67]. We conduct ablation experiments to validate the effectiveness of covisible map and multi-view supervised losses.

To summarize, this paper makes the following main contributions:

- We propose an end-to-end self-supervised architecture MGCNet for face alignment and monocular 3D face reconstruction tasks. To our best knowledge, we are the first to leverage multi-view geometry consistency to mitigate the ambiguity from monocular face pose estimation and depth reconstruction in the training process.
- We build a differentiable covisible map for general view synthesis, which can mitigate the self-occlusion crux of view synthesis. Based on view synthesis, three differentiable multi-view geometry consistency loss functions are proposed as pixel consistency loss, depth consistency loss, and facial epipolar loss.
- Our MGCNet result on the face alignment benchmark [67] shows that we achieve more than a **12**% improvement over other state-of-the-art methods, especially in large and extreme face pose cases. Comparison on the challenging 3D Face Reconstruction datasets [3, 36, 60, 64] shows that MGCNet outperforms the other methods with the largest margin of **17**%.

2 Related work

2.1 Single-view Method

Recent CNN methods [12, 17, 18, 21, 24, 28, 32, 53, 54, 56, 59, 62, 67] train the CNN network supervised by 3D face scan ground truth and achieve impressive results. [17, 24, 39] generate synthetic rendered face images with real 3D scans. [18, 28, 32, 53, 56, 59] propose their deep neural networks trained using fitted 3D shapes by traditional methods as substitute labels. Lack of realistic training data is still a great hindrance.

Recently, some self-supervised or weak-supervised methods are proposed [22, 40, 45, 51, 52, 54, 55, 61, 66] to solve the lack of high-quality 3D face scans with robust testing result. Tewari *et al.* [52] propose an differentiable rendering process to build unsupervised face autoencoder based on pixel loss. Genova *et al.* [22] train a regression network mainly focus on identity loss that compares the features of the predicted face and the input photograph. Nevertheless, face pose and depth ambiguity originated from only monocular images still a limitation.

2.2 Multi-view or Video Based Method

There are established toolchains of 3D reconstruction [20, 46, 47, 57], aimming at recovering 3D geometry from multi-view images. One related operation is view synthesis [13, 19, 48], and the goal is to synthesize the appearance of the scene from novel camera viewpoints.

Several unsupervised approaches [15, 45, 50, 58] are proposed recently to address the 3D face reconstruction from multiple images or videos. Deng *et al.* [15] perform multi-image face reconstruction from different images by shape aggregation. Sanyal *et al.* [45] take multiple images of the same and different person, then enforce shape consistency between the same subjects and shape inconsistency

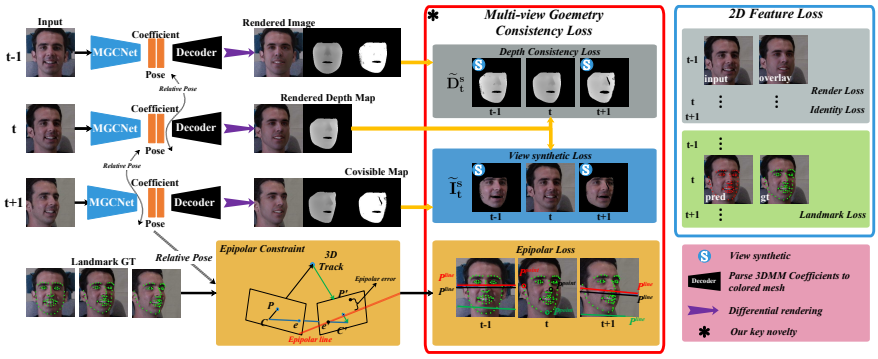


Fig. 1. The training flow of our MGCNet architecture, which is annotated in Section 3.1. The 2D feature loss part is our **baseline** in Section 3.3. Our novel multi-view geometry consistency loss functions are highlighted as $*$ in Section 3.

between the different subjects. Wu *et al.* [58] design an impressive multi-view framework (MVFNet), which is view-consistent by design, and photometric consistency is used to generate consistent texture across views. However, MVFNet is not able to generate results via a single input since it relies on multi-view aggregation during inference. Our MGCNet explicitly exploit multi-view consistency (both geometric and photometric) to constrain the network to produce view-consistent face geometry from **a single input**, which provides better supervision than 2D information only. Therefore, MGCNet improves the performance of face alignment and 3D face reconstruction as Section 4.

3 Method

3.1 Overview

The inference process of MGCNet takes a single image as input, while in the training process the input is N-view images (e.g. I_{t-1}, I_t, I_{t+1} for $N = 3$) of the same face and the corresponding ground-truth 2D landmarks $q_{t-1}^{gt}, q_t^{gt}, q_{t+1}^{gt}$. Then, the MGCNet estimates the 3DMM coefficients and face poses, whose notations are introduced in Section 3.2 in detail. $\widetilde{\mathcal{I}}_t^{(s)}, \widetilde{\mathcal{D}}_t^{(s)}$ represent the synthesized target images and the depth maps from the source views.

3.2 Model

Face model 3D Morphable Model (3DMM) proposed by [7] is the face prior model of the MGCNet. Specifically, the 3DMM encodes both face shape and texture as

$$\begin{aligned} S &= S(\alpha, \beta) = S_{mean} + A_{id}\alpha + B_{exp}\beta \\ T &= T(\gamma) = T_{mean} + T_{id}\gamma, \end{aligned} \tag{1}$$

where S_{mean} and T_{mean} denote the mean shape and the mean albedo respectively. A_{id}, B_{exp} and T_{id} are the PCA bases of identity, expression and texture. $\alpha, \beta \in R^{80}$ and $\gamma \in R^{64}$ are corresponding coefficient vectors to be estimated follow [15,51,52]. We use S_{mean}, T_{mean}, A_{id} and T_{id} provided by the Basel Face Model (BFM) [7,35], and B_{exp} from FaceWarehouse [9]. We exclude the ear and neck region as [15], and our final face model contains $\sim 36K$ vertices.

Camera model The pinhole camera model is employed to define the 3D-2D projection. We assume the camera is calibrated. The face pose P is represented by an euler angle rotation $R \in SO(3)$ and translation $t \in R^3$. The relative poses $P^{rel}_{t\to s} \in SE(3)$ from the **target view** to $N-1$ **source views** are defined as $P^{rel}_{t\to s} = \begin{bmatrix} R_t^{-1}R_s & R_t^{-1}(t_s - t_t) \\ 0 & 1 \end{bmatrix}$.

Illumination model To acquire realistic rendered face images, we model the scene illumination by Spherical Harmonics (SH) [37,38] as $SH(N_{re}, T|\theta) = T * \sum_{b=1}^{B^2} \theta_b H_b$, where N_{re} is the normal of the face mesh, $\theta \in R^{27}$ is the coefficient. The $H_b : R^3 \to R$ are SH basis functions and the $B^2 = 9$ (B = 3 bands) parameterizes the colored illumination in red, green and blue channels.

Finally, we concatenate all the parameters together into a $(\alpha, \beta, \gamma, R, t, \theta)$ 257-dimensional vector. All 257 parameters encode the 3DMM coefficients and the face pose, which are abbreviated as *coefficient* and *pose* in Figure 1.

3.3 2D Feature Loss

Inspired by recent related works, we leverage preliminary 2D feature loss functions in our framework.

Render loss The render loss aims to minimize the difference between the input face image and the rendered image as $\mathcal{L}_{render} = \frac{1}{M}\sum_{i=1}^{M} w^i_{skin}||I^i - I^i_{re}||$, where I_{re} is the rendered image, I is the input image, and M is the number of all 2D pixels in the projected 3D face region. w^i_{skin} is the skin confidence of i^{th} pixel as in [15]. Render loss mainly contributes to the albedo of 3DMM.

Landmark loss To improve the accuracy of face alignment, we employ the 2D landmark loss which defines the distance between predicted landmarks and ground truth landmarks as $\mathcal{L}_{lm} = \sum_{i=1}^{N} c^i_{lm}(q^i_{gt} - q^i)^2$, where N is the number of landmarks, and q the projection of the 3D landmarks picked from our face model. It is noted that the landmarks have different levels of importance, denoted as the confidence c_{lm} for each landmark. We set the confidence to 10 only for the nose and inner mouth landmarks, and to 1 else wise.

Identity loss The fidelity of the reconstructed face is an important criterion. We use the identity loss as in [22], which is the cosine distance between deep features of the input images and rendered images as $\mathcal{L}_{id} = \frac{\eta_1 \circ \eta_2}{|\eta_1||\eta_2|}$, where $\circ$ means the element-wise multiplication. η_1 and η_2 are deep features of input images and rendered images.

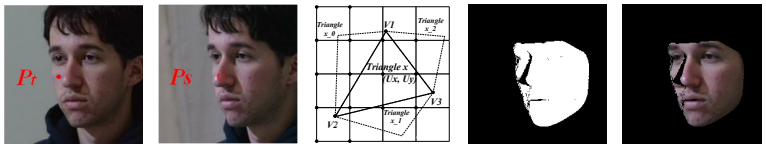
Regularization loss To prevent the face shape and texture parameters from diverging, regularization loss of 3DMM is used as $\mathcal{L}_{reg} = w_{id}\sum_{i=1}^{N_\alpha} \alpha^2 + w_{exp}\sum_{i=1}^{N_\beta} \beta^2 +$

$w_{tex}\sum_{i=1}^{N_\gamma}\gamma^2$ where $w_{id}, w_{exp}, w_{shape}$ are trade-off parameters for 3DMM coefficients regularization $(1.0, 0.8, 3\text{e}{-3}$ by default). $N_\alpha, N_\beta, N_\gamma$ are the length of 3DMM parameters α, β, γ.

Final 2D feature loss The combined 2D feature loss function $\mathcal{L}_{2D}$ is defined as $\mathcal{L}_{2D} = w_{render}\mathcal{L}_{render} + w_{lm}\mathcal{L}_{lm} + w_{id}\mathcal{L}_{id} + w_{reg}\mathcal{L}_{reg}$, where the trade-off parameters for 2D feature losses are set empirically $w_{render} = 1.9, w_{lm} = 1\text{e}{-3}, w_{id} = 0.2, w_{reg} = 1\text{e}{-4}$. We regard the **baseline** approach in the later experiement as the model trained by only 2D feature losses. In the followings, we present key ingredients of our contributions.

3.4 Occlusion-Aware View Synthesis

The key of our idea is to enforce multi-view geometry consistency, so as to achieve the self-supervised training. This could be done via view synthesis, which establishes dense pixel correspondences across multi-view input images. However, the view synthesis for face reconstruction is very easily affected by self-occlusion, as large and extreme face pose cases are common in read world applications. As shown in Figure 2, assuming a pixel p_t is visible in the left cheek as shown in Figure 2(a), the correspondence pixel p_s could not be found in Figure 2(b) due to nose occlusion. Self-occlusion leads to redundant pixel consistency loss and depth consistency loss. Furthermore the related gradient of self-occlusion pixels will be highly affected by the salient redundant error as red part in Figure 3 (Pixel Consistency Loss subfigure), which makes the training more difficult. For more practical and useful navigation in real scenarios, self-occlusion is worth to solve.



(a) (b) (c) (d) (e)

Fig. 2. (a) and (b) are the target view and source view pair; (c) is the covisible points and triangles; (d) is the covisible map; and (e) is the synthetic target view

We introduce the covisible maps C_s to account for the self-occlusion. Covisible map is a binary mask indicating the pixels which are visible in both source and target views. During the rendering process of the MGCNet, rasterization builds the correspondence between vertices of a triangle and image pixels $(V_{1,2,3} \sim U_x, U_y)$, as shown in Figure 2(c). The common vertices visible in two views (i.e., vertices that contribute to pixel rendering) are called covisible points. Then we define all triangles adjacent to covisible points as covisible triangles. Finally, we project covisible triangles of the 3D face from the target view to

image space, as shown in Figure 2(d), where the white region is covisible region. The improvement brings from covisible maps is elaborated in Figure 3, pixels are not covisible in the left of the nose in target view (red in Figure 3), which result in redundant error. The quantitative improvements are discussed in Section 4.5.

To generate the synthetic target RGB images from source RGB images, we first formulate the pixel correspondences between view pairs (I_s, I_t). Given a pair correspondence pixel coordinate p_t, p_s in I_t, I_s, the pixel value p_s is computed by bilinear-sampling [29], and the pixel coordinate p_s is defined as

$$p_s \sim \mathbf{K}_s[\mathbf{P}^{rel}_{t\to s}]\mathbf{D}_t(p_t)\mathbf{K}_t^{-1}p_t, \tag{2}$$

where $\sim$ represents the equality in the homogeneous coordinates, $\mathbf{K}_s$ and $\mathbf{K}_t$ are the intrinsics for the input image pairs, $\mathcal{D}_t$ is the rendered depth map of the target view, and $\mathbf{D}_t(p_t)$ is the depth for this particular pixel p_t in $\mathcal{D}_t$.

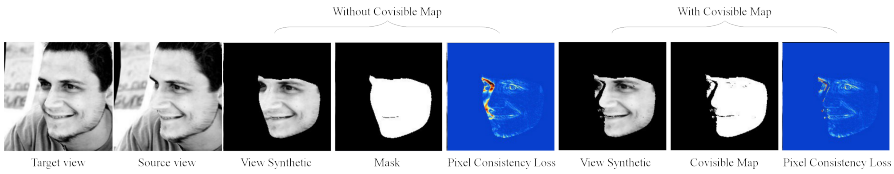


Fig. 3. The view synthesis results with and without covisible map. Without covisiable map, the pixel consistency loss is highly affected by self-occlusion.

3.5 Pixel Consistency Loss

We generate the synthesized target images by view synthesis, then we minimize the pixel error between the target view and the synthesized target views from the source views as

$$\mathcal{L}_{pixel} = \frac{1}{|C_s|}\sum_{i=1}^{|C_s|} C_s^i * \left|\widetilde{\mathcal{I}}_t^s(i) - \mathcal{I}_t(i)\right|, \tag{3}$$

where $\widetilde{\mathcal{I}}_t^s$ represents the synthesized target views from the source views. $\mathcal{I}_t(i)$ is the $i-th$ pixel value . Concretely, the first term $\widetilde{\mathcal{I}}_t^s$ is the bilinear-sampling operation, which computes the corresponding pixel coordinates using the relative pose $P^{rel}_{t\to s}$ and target depth map D_t^{re}. C_s is covisible map and $|C_s|$ denotes the total number of covisible pixels.

3.6 Dense Depth Consistency Loss

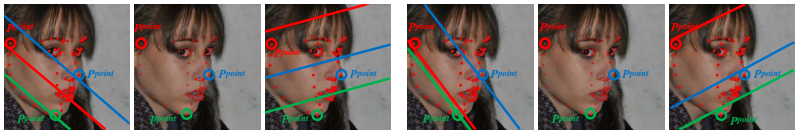
Compared to RGB images, depth maps are less adversely affected by the gradient locality issue [5]. Thus, we propose a dense depth consistency loss function which

contributes to solving depth ambiguity more explicitly, enforcing the multi-view consistency upon depth maps. Similarly, we synthesize the target depth maps $\widetilde{\mathbf{D}}_t^s$ from the source views via bilinear interpolation, and compute the consistency against the target depth map $\mathbf{D}_t$.

One critical issue is that the face region is cropped in the face detection data-preprocessing stage, making the depth value up to scale. To tackle this issue, we compute a ratio of two depth maps $\mathcal{S}_{depth}$ and rectify the scale of depth maps. Therefore, we define the dense depth consistency loss as

$$\begin{aligned} \mathcal{S}_{depth} &= \frac{\sum_{i=1}^{|C_s|} \mathbf{D}_t(i) C_s(i)}{\sum_{i=1}^{|C_s|} \widetilde{\mathbf{D}}_t^s(i) C_s(i)} \\ \mathcal{L}_{depth} &= \frac{1}{|C_s|} \sum_{i=1}^{|C_s|} \left| \mathcal{S}_{depth} \cdot \widetilde{\mathbf{D}}_t^s(i) - \mathbf{D}_t(i) \right|, \end{aligned} \tag{4}$$

where $\mathcal{C}_s(i), \mathcal{D}_t(i)$ are the $i-th$ covisible and depth value. $\mathcal{S}_{depth}$ is the depth scale ratio. Our experiment shows that the multi-view geometry supervisory signals significantly improve the accuracy of the 3D face shape.



(a) Baseline (b) Multi-view

Fig. 4. Eipipolar error from target to source views, (a) is the epipolar visualization from baseline network trained by 2D feature loss; and (b) is epipolar visualization of the MGCNet. The red, green and blue point means left ear bound, lower jaw and nose tip landmarks. The epipolar error of baseline is significantly larger than MGCNet.

3.7 Facial epipolar Loss

We use facial landmarks to build the epipolar consistency, as our epipolar loss is based on sparse ground-truth 2D facial landmarks, which is less likely to be affected by radiometric or illumination changes compared to pixel consistency or depth consistency losses, The epipolar loss in the *symmetric epipolar distance* [25] form between source and target 2D landmark $q_{t\leftrightarrow s} = \{\mathbf{p} \leftrightarrow \mathbf{p}'\}$ is defined as

$$\mathcal{L}_{epi}(q|\mathbf{R},\mathbf{t}) = \sum_{\forall(\mathbf{p},\mathbf{p}')\in q} \frac{\mathbf{p}'^{\mathbf{T}}\mathbf{E}\mathbf{p}}{\sqrt{(\mathbf{E}\mathbf{p})_{(\mathbf{1})}^{\mathbf{2}} + (\mathbf{E}\mathbf{p})_{(\mathbf{2})}^{\mathbf{2}}}} \tag{5}$$

where $\mathbf{E}$ being the essential matrix computed by $\mathbf{E} = [\mathbf{t}]_\times \mathbf{R}$, $[\cdot]_\times$ is the matrix representation of the cross product with $\mathbf{t}$. We simply omit the subindices for conciseness (q for $q_{t\leftrightarrow s}$, $\mathbf{R}$ for $\mathbf{R}_{t\to s}$, $\mathbf{t}$ for $\mathbf{t}_{t\to s}$).

3.8 Combined Loss

The final loss function $\mathcal{L}$ for our MGCNet is the combination of 2D feature loss and multi-view geometry loss. Training the network by only 2D feature losses leads to face pose and depth ambiguity, which is reflected in geometry inconsistency as the large epipolar error shown in Figure 4(a). Our MGCNet trained with pixel consistency loss, dense depth consistency, and facial epipolar loss shows remarkable improvement in Figure 4(b), which outstands our novel multi-view geometry consistency based self-supervised training pipeline. Finally, the combined loss function is defined as

$$\mathcal{L} = w_{2D} * \mathcal{L}_{2D} + w_{mul} * [w_{pixel} * \mathcal{L}_{pixel} + w_{depth} * \mathcal{L}_{depth} + w_{epi} * \mathcal{L}_{epi}], \quad (6)$$

where $w_{2D} = 1.0$ and $w_{mul} = 1.0$ balance the weights between the 2D feature loss for each view and the multi-view geometry consistency loss. The trade-off parameters to take into account are $w_{pixel} = 0.15, w_{depth} = 1\text{e}{-4}, w_{epi} = 1\text{e}{-3}$.

4 Experiment

We evaluate the performance of our MGCNet on the face alignment and 3D face reconstruction tasks which compared with the most recent state-of-the-art methods [6, 8, 15, 18, 22, 45, 50–52, 56, 66, 67] on diverse test datasets including AFLW20003D [67], MICC Florence [3], Binghamton University 3D Facial Expression (BU-3DFE) [60, 64], and FRGC v2.0 [36].

4.1 Implementation Details

Data 300W-LP [67] has multi-view face images with fitted 3DMM model, the model is widely used as ground truth in [18, 28, 56, 59], such multi-view images provide better supervision than only 2D features. Multi-PIE [23] are introduced to provide multi-view face images that help solve face pose and depth ambiguity. As multi-view face datasets are always captured indoor, and thus cannot provide diversified illumination and background for training, CelebA [33] and LS3D [8] are used as part of training data, which only contribute to 2D feature losses. Detail data process can be found in the *suppl.* material.

Network We use the ResNet50 [26] network as the backbone of our MGCNet, we only convert the last fully-connected layer to 257 neurons to match the dimension of 3DMM coefficients. The pre-trained model from ImageNet [44] is used as an initialization. We only use $N = 3$ views in practice, as $N = 5$ views lead to a large pose gap between the first view and the last view. We implement our approach by Tensorflow [1]. The training process is based on Adam optimizer [30] with a batch size of 5. The learning rate is set to 1e−4, and there are 400K total iterations for the whole training process.

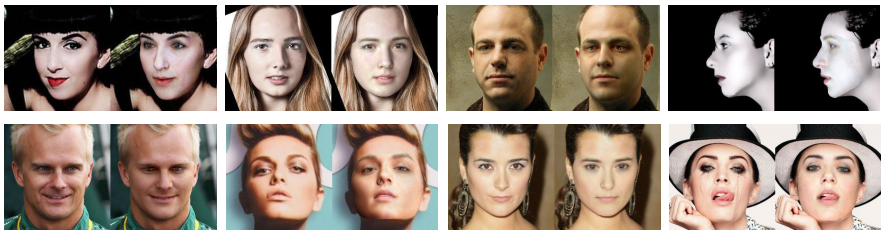


Fig. 5. A few results on a full range of lighting, pose, including large expressions. Each image pair is input image (left) and reconstruction result overlay (right). Further detail result (shape, albedo, and lighting) can be found in the *suppl.* material.

4.2 Qualitative Result

Result in different situations Our MGCNet allows for high-quality reconstruction of facial geometry, reflectance and incident illumination as Figure 5, under full range of lighting, pose, and expressions situations.
Geometry We evaluate the qualitative results of our MGCNet on AFLW20003D [67]. First, we compare our MGCNet with 3DDFA [67], RingNet [45], PRN [18], and Deng *et al.* [15] on front view samples, as Row 1 and Row 2 in Figure 6. Our predicted 3DMM coefficients produce more accurate results than the most methods, and we get comparable results with Deng *et al.* [15].

For these large and extreme pose cases as Row 3-6 in Figure 6, our MGCNet has better face alignment and face geometry than other methods. We have more vivid emotion in Row 4 of Figure 6, and the mouths of our result in Row 3,5 have obviously better shape than 3DDFA [67], RingNet [45], PRN [18], and Deng *et al.* [15]. Besides, the face alignment results from Row 3 to Row 6 support that we achieve better face pose estimation, especially in large and extreme pose cases.
Texture, illumination shadings We also visualize our result under geometry, texture, illumination shadings, and notice that our approach performs better than Tewari18 *et al.* [51] and Tewari19 *et al.* [50], where the overlay result is very similar to the input image as Figure 7(a). Further result and analysis about the result can be found in the *suppl.* material.

MGCNet does not focus on the appearance of 3DMM as [12, 21, 49–51, 53, 54, 62], which is only constrained by render loss. However, our multi-view geometry supervision can help render loss maximize the potential during training by accurate face alignment and depth value estimation. This makes MGCNet able to handle 3DMM texture and illumination robustly.

4.3 2D Face Alignment

The quantitative comparison of 6Dof pose is not conducted due to different camera intrinsic assumptions of different methods. Therefore, to validate that our MGCNet can mitigate the ambiguity of monocular face pose estimation, we evaluate our method on AFLW2000-3D, and compare our result with Zhu

Fig. 6. Comparisons with 3DDFA [67], RingNet [45], PRN [18], and Deng *et al.* [15] on ALFW20003D.

Table 1. (a)Performance comparison on AFLW2000-3D (68 landmarks). The normalized mean error (NME) for 2D landmarks with different yaw angles is reported. The first best result is highlighted in bold. (b) Average and standard deviation root mean squared error (RMSE) with *mm* in three environments of MICC Florence.

(a)

Method	0 to 30	30 to 60	60 to 90	Mean
3DDFA [67]	3.78	4.54	7.93	5.42
3D-FAN [8]	3.61	4.34	6.87	4.94
3DSTN [6]	3.15	4.33	5.98	4.49
CMD [66]	-	-	-	3.98
PRN [18]	2.75	3.55	5.11	3.62
Ours+BL	**2.75**	**3.28**	**4.31**	**3.45**
Ours+MGCNet	**2.72**	**3.12**	**3.76**	**3.20**

(b)

Method	Cooperative	Indoor	Outdoor
Zhu *et al.* [67]	2.69 ± 0.64	2.23 ± 0.49	2.22 ± 0.56
Sanyal *et al.* [45]	2.33 ± 0.43	2.19 ± 0.43	2.07 ± 0.45
Feng *et al.* [18]	2.30 ± 0.54	2.02 ± 0.50	2.10 ± 0.60
Tran *et al.* [56]	2.00 ± 0.55	2.05 ± 0.51	1.95 ± 0.51
Genova *et al.* [22]	1.87 ± 0.61	1.86 ± 0.60	1.87 ± 0.57
Deng *et al.* [15]	1.83 ± 0.59	1.78 ± 0.53	1.78 ± 0.59
Ours	**1.73 ± 0.48**	**1.78 ± 0.47**	**1.75 ± 0.47**

et al. [67] (3DDFA), Bulat and Tzimiropoulos [8] (3D-FAN), Bhagavatula *et al.* [6] (3DSTN), Zhou *et al.* [66] (CMD), and Feng *et al.* [18] (PRN). Normalized mean error (NME) is used as the evaluation metric, and the bounding box size of ground truth landmarks is deemed as the normalization factor. As shown in Table 1(a) Column 5, our result outperforms the best method with a large margin of **12**% improvement. Qualitative results can be found in the *suppl.* material.

Learning face pose from 2D features of monocular images leads to face pose ambiguity, the results of large and extreme face pose test samples suffer from this heavily. As the supervision of large and extreme face pose case is even less, which is not enough for training monocular face pose regressor. Our MGCNet provides further robust and dense supervision by multi-view geometry for face alignment in both frontal and profile face pose situations. The comparison in Table 1(a) corroborates our point that the compared methods [6, 8, 18, 66, 67] obviously degrades when the yaw angles increase from $(30, 60)$ to $(60, 90)$ in Column 4 of Table 1(a). We also conduct an ablation study that our MGCNet outperforms the baseline, especially on large and extreme pose case.

4.4 3D Face Reconstruction

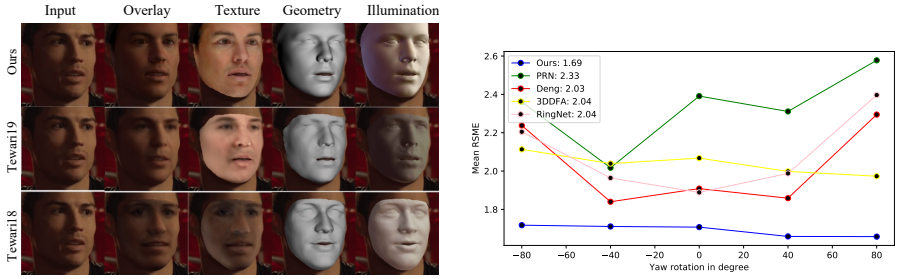
MICC Florence with Video MICC Florence provides videos of each subject in cooperative, indoor and outdoor scenarios. For a fair comparison with Genova *et al.* [22], Trans *et al.* [56] and Deng *et al.* [15], we calculate error with the average shape for each video in different scenarios. Following [22], we crop the ground truth mesh to $95mm$ around the nose tip and run iterative closest point (ICP) algorithm for rigid alignment. The results of [56] only contain part of the forehead region. For a fair comparison, we process the ground-truth meshes similarly. We use the point-to-plane root mean squared error(RMSE) as the evaluation metric. We compare with the methods of Zhu *et al.* [67] (3DDFA), Sanyal *et al.* [45] (RingNet), Feng *et al.* [18] (PRN), Genova *et al.* [22], Trans *et al.* [56] and Deng *et al.* [15]. Table 1(b) shows that our method outperforms state-of-the-art methods [15, 18, 22, 45, 56, 67] on all three scenarios.

MICC Florence with Rendered Images Several current methods [15, 18, 28, 59] also generate rendered images as test input. Following [15, 18, 28, 59], we render face images of each subject with 20 poses: a pitch of -15, 20 and 25 degrees, yaw angles of -80, -60, 0, 60 and 80 degrees, and 5 random poses. We use the point-to-plane RMSE as the evaluation metric, and we process the ground truth mesh as above. Figure 7(b) shows that our method achieves a significant improvement of **17**% higher than the state-of-the-art methods.

The plot also shows that our MGCNet performs obvious improvement on the extreme pose setting $x - axis[-80, 80]$ in Figure 7(b). As we mitigate both pose and depth ambiguity by multi-view geometry consistency in the training process. Extreme pose sample benefits from this more significantly, since the extreme pose input images have even less 2D features. Profile face case contains more pronounced depth info (eg. bridge of the nose), where large error happen.

FRGC v2.0 Dataset FRGC v2.0 is a large-scale benchmark includes 4007 scans. We random pick 1335 scans as test samples, then we crop the ground truth

mesh to 95mm around the nose tip. We first use 3D landmark as correspondence to align the predict and ground truth result, then ICP algorithm is used as fine alignment. Finally, point-to-point mean average error (MAE) is used as the evaluation metric. We compare with the methods of Galteri *et al.* [21] (D3R), 3DDFA [67], RingNet [45], PRN [18], and Deng *et al.* [15]. Table 2(a) shows that our method outperforms these state-of-the-art methods. Our MGCNet performs higher fidelity and accurate result on both frontal and profile face pose view.



(a) (b)

Fig. 7. (a)Comparison to Tewari18 *et al.* [51] and Tewari19 *et al.* [50]. Our MGCNet trained by multi-view consistency loss outperforms Tewari's results in face pose, illumination and geometry. Further result can be found in the *suppl.* material. (b) Comparison with 3DDFA [67], RingNet [45], PRN [18], and Deng *et al.* [15] on MICC Florence rendered images.

Table 2. (a) Comparison with D3R [21], 3DDFA [67], RingNet [45], PRN [18], and Deng *et al.* [15] with MAE of *mm* on FRGC v2.0 dataset. (b) Mean and standard deviation point-to-point RMSE with *mm* on the BU-3DFE dataset [60, 64] compared with Tewari17 *et al.* [52], Tewari18 *et al.* [51], Tewari19 *et al.* [50], Deng *et al.* [15].

(a)

Method	[21]	[18]	[45]	[67]	[15]	**Ours**
MAE	3.63	2.33	2.22	2.21	2.18	**1.93**
Time	-	9.8ms	2.7ms	75.7ms	20ms	20ms

(b)

Method	[52]	[51] Fine	[51] Coarse	[50]	[15]	**Ours**
Mean	3.22	1.83	1.81	1.79	1.63	**1.55**
Std	0.77	0.39	0.47	0.45	0.33	**0.31**

BU-3DFE Dataset We evaluate our method on the BU-3DFE dataset following [50]. Following [50], a pre-computed dense correspondence map is used to calculate a similarity transformation from predict mesh to the original ground-truth 3D mesh, and help to calculate the point-to-point RMSE. From Table 2(b), the reconstruction error of our method is lower than the current state-of-art methods [15, 50–52]. Our MGCNet achieves better performance by using the multi-view geometry consistency loss functions in the training phase. Qualitative results can be found in the *suppl.* material.

Table 3. Evaluation of different training loss configurations. The ablation study performance of MGCNet is evaluated on MICC Florence 3D Face dataset [3] by RMSE.

Loss Configuration					MICC Florence video		
2D feature	Pixel Consistency	Depth Consistency	Epipolar	Covisible map	Cooperative	Indoor	Outdoor
✓	-	-	-	-	1.83	1.82	1.81
✓	✓	-	-	✓	1.80	1.80	1.80
✓	-	✓	-	✓	1.77	1.79	1.80
✓	-	-	✓	-	1.79	1.81	1.77
✓	✓	✓	-	✓	1.76	1.81	1.81
✓	✓	✓	✓	-	1.80	1.81	1.82
✓	✓	✓	✓	✓	**1.73**	**1.78**	**1.75**

4.5 Ablation study

To validate the efficiency of our multi-view geometry consistency loss functions. We conduct ablation studies for each component on the MICC Florence dataset [3], as shown in Table 3. The ablation study mainly focuses on the proposed multi-view geometry consistency loss functions. Firstly, we deem the baseline method as the model trained with only 2D feature losses, as in Row 1. Secondly, we add our pixel consistency loss, dense depth consistency loss, and epipolar loss to the baseline in Row 2, Row 3 and Row 4. It shows that these losses help produce lower reconstruction errors than the baseline, even when they are used separately. Thirdly, comparing from Row 5 to Row 7, we combine multiple multi-view geometry loss functions and achieve state-of-the-art results, which demonstrates the effectiveness of the proposed self-supervised learning pipeline. Finally, comparing from Row 6 to Row 7, we prove that our novel covisible map to solve self-occlusion in view synthesis algorithm can help training a more accurate model. The qualitative ablation study is in the *suppl.* material.

5 Conclusion

We have presented a self-supervised pipeline MGCNet for monocular 3D Face reconstruction and demonstrated the advantages of exploiting multi-view geometry consistency to provide more reliable constraint on face pose and depth estimation. We emphasize on the occlusion-aware view synthesis and multi-view losses to make the result more robust and reliable. Our MGCNet profoundly reveals the capability of multi-view geometry consistency self-supervised learning in capturing both high-level cues and feature correspondences with geometry reasoning. The results compared to other methods indicate that our MGCNet can achieve the outstanding result without costly labeled data. Our further investigations will focus on multi-view or video-based 3D face reconstruction.

6 Acknowledgements

This work is supported by Hong Kong RGC GRF 16206819 & 16203518 and T22-603/15N.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). pp. 265–283 (2016)
2. Aldrian, O., Smith, W.A.: Inverse rendering of faces with a 3d morphable model. IEEE transactions on pattern analysis and machine intelligence **35**(5), 1080–1093 (2012)
3. 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. p. 7980. J-HGBU 11, ACM, New York, NY, USA (2011). https://doi.org/10.1145/2072572.2072597, `http://doi.acm.org/10.1145/2072572.2072597`
4. Bas, A., Smith, W.A., Bolkart, T., Wuhrer, S.: Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences. In: Asian Conference on Computer Vision. pp. 377–391. Springer (2016)
5. Bergen, J.R., Anandan, P., Hanna, K.J., Hingorani, R.: Hierarchical model-based motion estimation. In: European conference on computer vision. pp. 237–252. Springer (1992)
6. Bhagavatula, C., Zhu, C., Luu, K., Savvides, M.: Faster than real-time facial alignment: A 3d spatial transformer network approach in unconstrained poses. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3980–3989 (2017)
7. Blanz, V., Vetter, T., et al.: A morphable model for the synthesis of 3d faces. In: Siggraph. vol. 99, pp. 187–194 (1999)
8. 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: Proceedings of the IEEE International Conference on Computer Vision. pp. 1021–1030 (2017)
9. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics **20**(3), 413–425 (2013)
10. Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K.: Real-time facial animation with image-based dynamic avatars. ACM Transactions on Graphics **35**(4) (2016)
11. Casser, V., Pirk, S., Mahjourian, R., Angelova, A.: Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8001–8008 (2019)
12. Chen, A., Chen, Z., Zhang, G., Mitchell, K., Yu, J.: Photo-realistic facial details synthesis from single image. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 9429–9439 (2019)
13. Chen, S.E., Williams, L.: View interpolation for image synthesis. In: Proceedings of the 20th annual conference on Computer graphics and interactive techniques. pp. 279–288. ACM (1993)
14. Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and rendering architecture from photographs. University of California, Berkeley (1996)
15. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: 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. pp. 0–0 (2019)

16. Dou, P., Kakadiaris, I.A.: Multi-view 3d face reconstruction with deep recurrent neural networks. Image and Vision Computing **80**, 80–91 (2018)
17. Dou, P., Shah, S.K., Kakadiaris, I.A.: End-to-end 3d face reconstruction with deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5908–5917 (2017)
18. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 534–551 (2018)
19. Fitzgibbon, A., Wexler, Y., Zisserman, A.: Image-based rendering using image-based priors. International Journal of Computer Vision **63**(2), 141–151 (2005)
20. Furukawa, Y., Curless, B., Seitz, S.M., Szeliski, R.: Towards internet-scale multi-view stereo. In: 2010 IEEE computer society conference on computer vision and pattern recognition. pp. 1434–1441. IEEE (2010)
21. Galteri, L., Ferrari, C., Lisanti, G., Berretti, S., Del Bimbo, A.: Deep 3d morphable model refinement via progressive growing of conditional generative adversarial networks. Computer Vision and Image Understanding **185**, 31–42 (2019)
22. Genova, K., Cole, F., Maschinot, A., Sarna, A., Vlasic, D., Freeman, W.T.: Unsupervised training for 3d morphable model regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8377–8386 (2018)
23. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28**(5), 807–813 (2010)
24. Guo, Y., Cai, J., Jiang, B., Zheng, J., et al.: Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. IEEE transactions on pattern analysis and machine intelligence **41**(6), 1294–1307 (2018)
25. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
26. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
27. 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. ACM Transactions on Graphics (TOG) **36**(6), 195 (2017)
28. Jackson, A.S., Bulat, A., Argyriou, V., Tzimiropoulos, G.: Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1031–1039 (2017)
29. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. In: Advances in neural information processing systems. pp. 2017–2025 (2015)
30. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
31. Liu, F., Zeng, D., Zhao, Q., Liu, X.: Joint face alignment and 3d face reconstruction. In: European Conference on Computer Vision. pp. 545–560. Springer (2016)
32. Liu, F., Zhu, R., Zeng, D., Zhao, Q., Liu, X.: Disentangling features in 3d face shapes for joint face reconstruction and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5216–5225 (2018)
33. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015)
34. Mahjourian, R., Wicke, M., Angelova, A.: Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In: Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5667–5675 (2018)
35. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. pp. 296–301. Ieee (2009)
36. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). vol. 1, pp. 947–954. IEEE (2005)
37. Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 497–500. ACM (2001)
38. Ramamoorthi, R., Hanrahan, P.: A signal-processing framework for inverse rendering. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 117–128. ACM (2001)
39. Richardson, E., Sela, M., Kimmel, R.: 3d face reconstruction by learning from synthetic data. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 460–469. IEEE (2016)
40. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1259–1268 (2017)
41. Romdhani, S., Vetter, T.: Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 2, pp. 986–993. IEEE (2005)
42. Roth, J., Tong, Y., Liu, X.: Unconstrained 3d face reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2606–2615 (2015)
43. Roth, J., Tong, Y., Liu, X.: Adaptive 3d face reconstruction from unconstrained photo collections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4197–4206 (2016)
44. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
45. Sanyal, S., Bolkart, T., Feng, H., Black, M.J.: Learning to regress 3d face shape and expression from an image without 3d supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7763–7772 (2019)
46. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
47. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
48. Seitz, S.M., Dyer, C.R.: View morphing. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 21–30. ACM (1996)
49. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1576–1585 (2017)
50. 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: Proceed-

ings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 10812–10822 (2019)
51. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2549–2559 (2018)
52. 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: Proceedings of the IEEE International Conference on Computer Vision. pp. 1274–1283 (2017)
53. Tran, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.G.: Extreme 3d face reconstruction: Seeing through occlusions. In: CVPR. pp. 3935–3944 (2018)
54. Tran, L., Liu, F., Liu, X.: Towards high-fidelity nonlinear 3d face morphable model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1126–1135 (2019)
55. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7346–7355 (2018)
56. Tuan Tran, A., Hassner, T., Masi, I., Medioni, G.: Regressing robust and discriminative 3d morphable models with a very deep neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5163–5172 (2017)
57. Wu, C., et al.: Visualsfm: A visual structure from motion system (2011)
58. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 959–968 (2019)
59. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7663–7672 (2019)
60. Yin, L., Wei, X., Sun, Y., Wang, J., Rosato, M.J.: A 3d facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06). pp. 211–216. IEEE (2006)
61. Yoon, J.S., Shiratori, T., Yu, S.I., Park, H.S.: Self-supervised adaptation of high-fidelity face models for monocular performance tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4601–4609 (2019)
62. Zeng, X., Peng, X., Qiao, Y.: Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2315–2324 (2019)
63. Zhan, H., Garg, R., Saroj Weerasekera, C., Li, K., Agarwal, H., Reid, I.: Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 340–349 (2018)
64. Zhang, X., Yin, L., Cohn, J.F., Canavan, S., Reale, M., Horowitz, A., Liu, P.: A high-resolution spontaneous 3d dynamic facial expression database. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). pp. 1–6. IEEE (2013)
65. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1851–1858 (2017)

66. Zhou, Y., Deng, J., Kotsia, I., Zafeiriou, S.: Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1097–1106 (2019)
67. Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S.Z.: Face alignment across large poses: A 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 146–155 (2016)