Deep Unsupervised 3D SfM Face Reconstruction Based on Massive Landmark Bundle Adjustment

Yuxing Wang
Rochester Institute of Technology
Rochester, NY, USA
yw2009@rit.edu

Yawen Lu
Rochester Institute of Technology
Rochester, NY, USA
yl4280@rit.edu

Zhihua Xie
Jiangxi Science and Technology Normal University
Nanchang, Jiangxi, China
xie_zhihua68@aliyun.com

Guoyu Lu
Rochester Institute of Technology
Rochester, NY, USA
luguoyu@cis.rit.edu

ABSTRACT

We address the problem of reconstructing 3D human face from multi-view facial images using Structure-from-Motion (SfM) based on deep neural networks. While recent learning-based monocular view methods have shown impressive results for 3D facial reconstruction, the single-view setting is easily affected by depth ambiguities and poor face pose issues. In this paper, we propose a novel unsupervised 3D face reconstruction architecture by leveraging the multi-view geometry constraints to train accurate face pose and depth maps. Facial images from multiple perspectives of each 3D face model are input to train the network. Multi-view geometry constraints are fused into unsupervised network by establishing loss constraints from spatial and spectral perspectives. To make the trained 3D face have more details, facial landmark detector is explored to acquire massive facial information to constrain face pose and depth estimation. Through minimizing massive landmark displacement distance by bundle adjustment, an accurate 3D face model can be reconstructed. Extensive experiments demonstrate the superiority of our proposed approach over other methods.

CCS CONCEPTS

• **Information systems** → **Multimedia content creation**.

KEYWORDS

Landmark Detection, Structure from Motion, 3D Face Reconstruction, Deep Learning

ACM Reference Format:
Yuxing Wang, Yawen Lu, Zhihua Xie, and Guoyu Lu. 2021. Deep Unsupervised 3D SfM Face Reconstruction Based on Massive Landmark Bundle Adjustment. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China* . ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475689

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/ 10…$15.00
https://doi.org/10.1145/3474085.3475689

1 INTRODUCTION

3D face reconstruction is widely applied to many fields such as virtual reality (VR) and augmented reality (AR) [28]. To obtain robust reconstruction against many factors such as age, gender and expression, current 3D face analysis methods majorly rely on the precise 3D Morphable Model (3DMM), which provide a parametric representation of 3D face models [27]. However, the research of 3D face reconstruction is obstructed by several inherent challenges. First, obtaining ground-truth 3D annotations for in-the-wild images is both expensive and laborious. Second, it is sensitive to the quantity and quality of training data. Third, 3D face reconstruction methods have limited capacity in representing details in face shapes and textures. Recently, some work has demonstrated that regressing 3DMM parameters using convolutional neural networks(CNN) achieves superior performance to traditional geometry methods[10].

In spite of the remarkable progress in this topic, the lack of reliable 3D constraints can cause unresolvable ambiguities: the height of nose and cheekbones. This paper mainly focuses on exploiting multi-view geometric constraints to reconstruct the faithful 3D shapes from 2D face images. The main motivation is to incorporate those constraints into our CNN model to estimate accurate face pose and depth maps. To enable the trained 3D face having more expression details, we address the problem of reconstructing 3D human face from multi-view facial images using Structure-from-Motion (SfM) based on the CNN network. Targeting at face reconstruction issues, we designed a learning-based SfM framework that can rely on the face characteristics to reconstruct an accurate face shape with arbitrary frames as input. To explore the face properties, we developed a face landmark detection network to identify extensive landmark points that cover the details of the entire face, which can detect much more landmarks (e.g., 500 to 800) than the existing methods that can only detect 68 or maximally 106 landmarks. The entire deep SfM network explores both spatial and spectral constraints relying on the concept of bundle adjustment. From the spatial perspective, displacement error is applied to constrain the 3D vertices and depth map, which enforces the 2D landmark points to be consistent with the corresponding landmarks across different frames on position. From spectral perspective, the RGB values of pixels corresponding to the same 3D vertex should be close. Though our network can leverage on massive landmarks, the SfM network based on even just 68 landmarks commonly appeared on existing

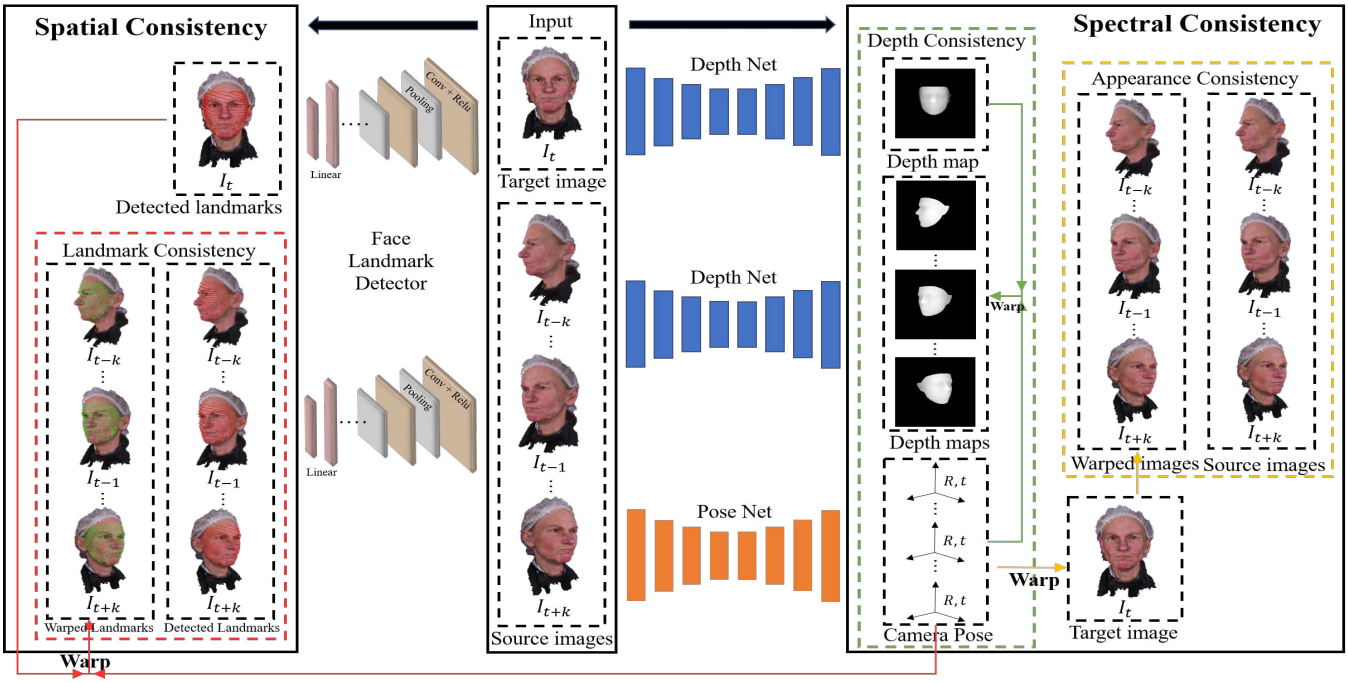

Figure 1: Overview of our deep 3D SfM face reconstruction architecture. Our network takes multiple frames as input. Geometry bundle adjustment landmark consistency loss functions, spectral appearance consistency loss and global depth consistency loss are highlighted as in Section 4 based on the concept of bundle adjustment. With the designed depth estimation and camera motion network, the massive detected facial landmarks from target image can be warped to all other frames to serve as a spatial geometry constraint together with spectral consistency. During the inference, our network is able to accept arbitrary number of frames as input to generate a complete 3D face shape.

facial landmark detection methods can also achieve a superior 3D face reconstruction effect.

To summarize, the contribution of our network is as follows: 1) a deep unsupervised SfM network targeting at face reconstruction is designed to explore the facial properties; 2) a landmark detection framework is developed to detect massive and stable landmark points under various head poses; 3) the concept of bundle adjustment in neural network is utilized to optimize the SfM learning process, which explores the multiple-view geometric constraints; 4) both spatial and spectral cues are applied to enhance the learning effect. The training framework is shown in Fig. 1.

2 RELATED WORK

Single-view based reconstruction method Recent learning-based methods mainly trained convolutional neural networks (CNN) to recover 3D shapes from a single image. 3D scanning face objects are served as ground truths to guide the network training [5] [23] [12] [30]. Specially, those methods design networks to regress 3D morphable face models (3DMM) [27] and fit the facial shape during the testing. Such methods usually rely on a pre-existing 3DMM and lack enough labelled training data.

Few recent works use self-supervised methods to deal with the limited capacity of high-quality 3D face models for training [34] [18]. Sanyal et al. [18] leveraged multiple images of a person to fit a FLAME model. Zhou et al. [34] proposed a non-linear 3D morphable face model to jointly learn shape and texture within a geometric convolutional network. Tewari et al. [22] presented an unsupervised model-based face auto-encoder based on pixel loss to learn parameters like pose, shape, expression and illumination. However, most of the aforementioned approaches still heavily rely on 3DMM model parameters, and reconstruction only from one single-view image exists pose and depth ambiguity.

Multiple-view based reconstruction method There are several classic pipelines for 3D reconstruction with multi-view images [20] [25], The majority of methods based on Structure-from-Motion (SfM) or Simultaneous Localization and Mapping (SLAM) can generate 3D objects from 2D images by using the principles of multiple view geometry. However, these classic geometry-based methods are subject to a number of restrictions, especially precise feature matching across images of different perspectives. The effect of feature matching could be extremely poor when there is a large baseline between the viewpoints [16]. In addition, the correct feature correspondences are also difficult due to surface reflections and low/repetitive textures on objects [19][24][14]. Liang et al. [14] proposed a 3D reconstruction method based on factorization SfM. 68 facial landmarks are extracted via learning-based method to factorize matching landmarks. Although it addresses the landmark self-occlusion issue caused by yaw rotation, the rotation invariance is only limited to a relatively small angle. In addition, the detected landmarks are not distinguished to be visible or occluded towards the camera, which adds false positive correspondences in their conventional factorization SfM method.

A better way to reconstruct faithful 3D faces is to exploit multi-view geometric constraints based on deep neural networks. Dou

et al. [4] combine deep convolutional neural networks (CNNs) together with recurrent neural networks (RNNs) to produce more discriminative reconstructions. There are several unsupervised networks to address 3D face reconstruction from multiple images [21] [26]. Tewari et al. [21] proposed a video-based unsupervised training network to learn a cross-frame consistent face shape based on the shape and appearance across multiple frames of the same face collected from the Internet. Wu et al. [26] also designed an unsupervised multi-view framework to explore view-consistency photometric loss to generate consistent texture information across multiple views. However, those methods require extensive view's input to aggregate the 3D face. Different from such methods, our method explores both spatial and spectral consistency to realize a better representation of face shapes. During inference, we can reconstruct an accurate face shape with arbitrary number of frames.

3 FACE LANDMARK DETECTION FRAMEWORK

In this section, we elaborate the learning-based face landmark detection framework. The landmark detection framework is to train massive face landmark points which can facilitate the 3D SfM reconstruction tasks. Existing methods mainly detect 68 face landmark points, which distribute around the entire face and organ contours. To further enhance the landmark constraints so as to reconstruct more accurate face structures, we investigate to detect more face landmark points. This process involves 3D face model landmark identification and 2D face image landmark generation. Once we obtain the trained labels for 2D image face landmark from the 3D models, a designated neural network can be developed to detect massive face landmarks that can effectively navigate the 3D SfM reconstruction neural network training.

3.1 3D Model Landmark Generation

In order to provide training samples for the 2D landmark network, we use 3DMM [28] to generate a 3D landmark model, as shown in Fig. 2. Those 3D vertices cover the whole 3D face including visible and invisible part from 2D view. However, because multi-view face images are fed into 3D sfm reconstruction network, they need to find the corresponding landmark matches between each view pair based on sfm principles. The matching pairs can be formed from two images crossing arbitrary pose variations (e.g.,0 degree and 30 degree view pair). We keep the 3D vertices visible from the camera's perspectives and filter out the invisible 3D vertices when labeling the 3D vertices, which helps to eliminate the interference of the invisible landmark from the 2D view in the process of detecting 2D landmark. As shown in Fig. 3, when the face is rotated to different angles, only the landmarks visible to the camera are displayed, which is critical to establish correct correspondences. To align landmarks between 2D images of different perspectives, each 3D vertex is labeled with a unique ID.

3.2 2D Image Landmark Generation

To train landmark detection network, 2D images with labeled landmarks are used as training samples. Once we have the 3D landmark models with labeled vertices (with a unique ID for each vertex), the 3D models and labeled vertices can be projected to 2D space to obtain images with identified landmarks. Assuming there is a virtual camera, the projection matrix is composed by camera extrinsics

Figure 2: 3D landmark model using 3DMM. Each 3D vertex is labeled with a unique ID.

Figure 3: Selected 800 landmarks on 2D images projected from 3D vertices.

(rotation and translation) and intrinsics (focal length and principal point), which projects 3D vertices in world coordinate to 2D image pixels in image coordinate, as equation below:

$$p = \begin{pmatrix} x^{im} \\ y^{im} \\ 1 \end{pmatrix} = \begin{bmatrix} f & 0 & x_0 & 0 \\ 0 & f & y_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X^w \\ Y^w \\ Z^w \\ 1 \end{bmatrix} \quad (1)$$

where p is the pixel position in 2D images, composed by its x coordinate x^{im} and y coordinate y^{im}. f represents focal length. x_0 and y_0 are the principal point's coordinates. R is a $3 * 3$ rotation matrix and t is a $3 * 1$ translation vector. X^w, Y^w, and Z^w represent the 3D point's world coordinates. To increase the robustness of face landmark detection, we rotate the virtual camera around the 3D face model to generate a sequence of 2D face images and the 2D landmark points projected by 3D vertices with unique IDs in world coordinate, as shown in Fig. 3. Since each 3D vertex has a unique ID in a 3D face model, when 3D vertices are projected to 2D images from different perspectives, all the projected 2D landmark points are ensured to be true positive correspondences in the sequence of images from different perspectives.

3.3 Landmark Detection Framework

After using 3DMM to generate 3D vertices visible to camera and project 3D vertices to 2D landmark points, we train the face landmark detection network with labeled 2D landmark points. Our supervised landmark detection network takes facial images as input with 800 (can be changed to other numbers) landmark labels and outputs a feature vector of 1600 dimensions. Therefore, to learn the 800 landmark points in 2D face images, the CNN network is trained to make predicted feature vector close to ground truth label. The network is composed of seventeen convolutional layers followed by Rectified Linear Units (ReLU) activation function. A single fully connected layer is applied to output a 1600 dimension vectors which can reshape to 800×2 dimensional landmarks. During 3D face reconstruction, the trained landmark detector detects the landmarks for each face image to build detailed and tight constraints for later depth and camera pose estimation network training, and

then the landmarks from target image can be projected to all other source images to achieve landmark bundle adjustment constraint.

4 DEEP SFM FRAMEWORK

The entire structure-from-motion neural network explores the multiple-view geometry constraints to reconstruct the 3D face based on a sequence of image inputs. With the images captured from different perspectives of the 3D face model, the SfM network can reconstruct an accurate 3D face. In the training process, bundle adjustment is applied to enforce the 2D pixels projected by reconstructed 3D face points to be consistent in different camera poses, which involves both spatial and spectral constraints. For spatial loss, landmark points are applied to decrease the displacement distance between the originally detected landmark points from pre-trained landmark detector and the 2D pixels projected by the reconstructed 3D face points. For spectral constraints, the 2D pixels' RGB values of the target image warped by the face depth map and camera motion are enforced to be close to RGB values of the source image. Based on our neural network structure and the effective loss constraints, our deep SfM network can estimate the face depth of each image accurately. Once the landmarks are detected, the above spatial and spectral consistency constraints are applied to optimize the 3D reconstructed face and camera poses, which is entirely unsupervised due to the self-supervision between 3D reconstructed faces and camera poses based on our spatial and spectral constraints.

4.1 Network Structure

The 3D SfM reconstruction framework is to predict the depth map and camera pose through a sequence of images inputs with different perspectives. Figure 1 demonstrates the basic structure of our deep unsupervised 3D SfM face reconstruction network, which is to jointly learn the depth map and the corresponding camera pose by both spatial and spectral constraints between target image and source image. The network structure can be divided into two parts. One marked by blue color in Fig.1 is to generate an accurate depth map with an encoder-decoder network structure. The encoder network extracts significant features from the input images, composed of seventeen convolutional layers and a single fully connected layer, and then the decoder network uses skip connections [15] to further interprets those feature representations to generate depth map. Similar to the structure of the depth estimation network, instead of generating depth maps, the camera pose estimation network outputs relative 6 DoF parameters which can construct a rotation matrix R (3×3) and a translation vector t (3×1).

4.2 Landmark Bundle Adjustment

With any consecutive frames as input, our deep SfM network is able to estimate depth maps D, and recover them into a complete point cloud. The relative 6 DoF poses P_{rel} between them can be estimated from the pose estimation network. As we have already trained a massive facial landmark detection network as introduced in Sec.3.3, with the predicted depth information for each frame and the estimated relative camera motions between any two frames, we can build an unsupervised constraint using the detected landmarks in the source view images and the warped landmarks of the target view image. Different from other camera motion estimation algorithms [33] [31], we extend the local frames optimization (3-snippet) to a full sequence for Bundle Adjustment (BA) optimization. The proposed bundle adjustment-based landmark consistency loss is defined as:

$$L_{Landmark\_BA} = \sum_{M}\sum_{N}\sum_{i} \|u_{Mi}(\pi(\mathbf{P_{ab}},\ D_M)) - u_{src-Ni}\|_2 \quad (2)$$

where π represents a mapping relationship from 3D points to 2D pixels. $\mathbf{P_{ab}}$ is the relative camera motion from the target image to the source images. D_M corresponds to the depth value at the M_{th} target image. i is the number of the detected landmarks, and N represents the number of all other views except the current view. $u \in R^2$ means the 2D coordinates of the face landmarks. The L2 loss guidance here is to measure the distance difference between each projected landmarks in the warped target image and the detected landmarks in the source image and minimize it.

4.3 Spectral Constraints

Except for the designated bundle-adjustment based landmark consistency loss, we further enforce the spectral appearance of target image to be consistent with source images after warping based on face depth and camera motion. This can be achieved by imposing the following spectral photometric consistency loss:

$$L_{spectral} = \sum_{i}\sum_{k}\sum_{j} \left\|I_i(\pi(\mathbf{P_{ab}},\ D_j \cdot \boldsymbol{p_j})) - I_{src-k}(\boldsymbol{p_j})\right\|_1 \quad (3)$$

where $\mathbf{P_{ab}}$ is the relative camera motion from the target image to the source images. D_j corresponds to the depth value of a pixel p_j at the i_{th} target image I_i. I_{src-k} is the k_{th} source image. The L1 loss is to guide the optimization by reducing the pixel RGB value difference between the warped target image and the source image.

As L1 loss alone is not robust enough to the light illumination and contrast variation, we extend it with the image structural similarity index to jointly evaluate two images in illuminance, contrast, and structure. The improved spectral appearance consistency loss is a comprehensive expression of SSIM and L1 loss as:

$$L_{spectral} = \sum_{i}\sum_{k}\sum_{j} \lambda_1 \left\|I_i(\pi(\mathbf{P_{ab}},\ D_j \cdot \boldsymbol{p_j})) - I_{src-k}(\boldsymbol{p_j})\right\|_1 + \lambda_2 \frac{1-SSIM(I_i, I_{src-k})}{2} \quad (4)$$

where $SSIM(I_i, I_{src-k})$ compute the element-wise similarity between the warped target image I_i and the source view image I_{src}. We set λ_1=0.15 and λ_2 = 0.85 following [8] [7].

Depth maps are less sensitive to the gradient locality [1] compared with normal color images. Therefore, we further introduce depth consistency across multiple-view frames to solve possible depth ambiguity. We synthesize the source depth maps $\tilde{D_s}$ from the target depth maps D_t, and then force the synthesized source depth to be close to the original source depth. We first compute a scale ratio of these two depths, and then define a depth consistency loss as follows:

$$L_{depth} = \frac{1}{|c|}\sum_{i}^{c} \left|\eta \cdot \tilde{D}_s(i) - D_t(i)\right|, \eta = \frac{\sum_i D_t(i)}{\sum_i \tilde{D}_s(i)} \quad (5)$$

where the η is the depth scale ratio between the synthesized depth and the original depth. The designed loss is able to achieve a scale-consistent estimation and provide an additional global geometry supervisory to improve the reconstruction performance.

5 EXPERIMENTS

5.1 Experiment setting

For facial landmark detection network, the input image is a grayscale image with the size of 384 × 384. The feature extraction stage is composed of four convolutional layers, four pooling layers, and three fully connected layer in the end. Each convolutional layer contains a filter bank producing multiple feature maps. The Rectified Linear Units (ReLu) [17] is selected as the activation function. For the pooling layers, We conduct max-pooling on non-overlap regions in the feature map. The fully connected layers are able to output a 1600 feature vectors which reshapes 800 landmarks × 2 (x and y coordinates) dimensions.

After extracting the 800 facial landmarks from each input image, the SfM-based learning network takes a video sequence from multiple views as input. The network simultaneously estimates the camera motion from each concatenated image pair and estimate a 3D point cloud from the depth prediction network. The learning-based 3D face reconstruction network is trained in an unsupervised manner and do not need any 3D supervision to guide the training process. It is implemented with PyTorch library and trained from scratch using Adam optimizer [13] with $\beta 1 = 0.9$ and $\beta 2 = 0.99$. Rectified Linear Units (ReLu) [17] is applied as activation functions for all convolutional layers. The weights of the depth estimation network and pose estimation network are initialized with Kaiming initialization [11] method with 2 batch size to achieve a trade-off between the efficiency and the memory usage. The whole framework is trained for 50 epochs.

5.2 Dataset Introduction

Stirling ESRC 3D face dataset [6] is utilized to manually generate a group of rendered multi-view images. The original format of the 3D objects is in a wavefront file containing 101 subjects (male = 47, female = 54) of 3D facial scans in a neutral expression. We utilize Trimesh [3] to continuously rotate the virtual camera every 2 degrees to project 3D objects into an image to generate rendered image sequences with a fixed focal length of 500 pixels. The resolution of the generated images are 384 × 384.

Facescape [29] is a large-scale dataset that contains a large number of high-quality 3D face subjects, parametric models and multi-view images. The age and gender of each object are also included in the original dataset. 847 subjects with 20 expressions (totally 16490 models) are provided for training, which is roughly 90 percent of the complete whole dataset. In this work, we randomly choose 6640 together with the selected 3D models in the Stirling ESRC 3D face dataset for training, and the other 2000 objects for testing. The 2D image generation process shares the same process as in the Stirling ESRC 3D face dataset.

5.3 Landmark Detection

Through training a set of rendered multi-view face images generated by the Stirling ESRC 3D face datasets and the Facescape 3D face datasets, the landmark detection network has superior performance with massive landmarks. As shown in Fig. 4, we compare ground truth landmarks with detected landmarks on the same image to demonstrate a good generalization and robustness when taking different views as input. Meanwhile, the mean squared error (MSE) between ground truth landmarks and corresponding detected landmarks is 4.52. In addition, to verify whether the landmarks are aligned between different perspectives, we visually plot corresponding landmarks detected from the face landmark detector between image pairs having different perspectives. As shown in Fig. 5, all the corresponding landmarks are matched correctly. Our landmark model is not affected by occlusion when rotating any degrees, indicating the stable landmark detection performance.

Figure 4: Green: ground truth landmarks; Red: detected landmarks by the landmark detection network.

Figure 5: Aligned 2D landmarks across multiple views

5.4 Qualitative Result

We provide visual results of the reconstructed 3D point clouds from input RGB images in Fig. 6 and Fig. 7. For each input image, the corresponding 3D point cloud is produced accurately and shown at two different viewpoints. With the help of the proposed spectral

Figure 6: Visual results on Stirling ESRC 3D face dataset. For each sample, we show the input image at the first column, and the 3D reconstruction results at different viewpoints at the second and third columns.

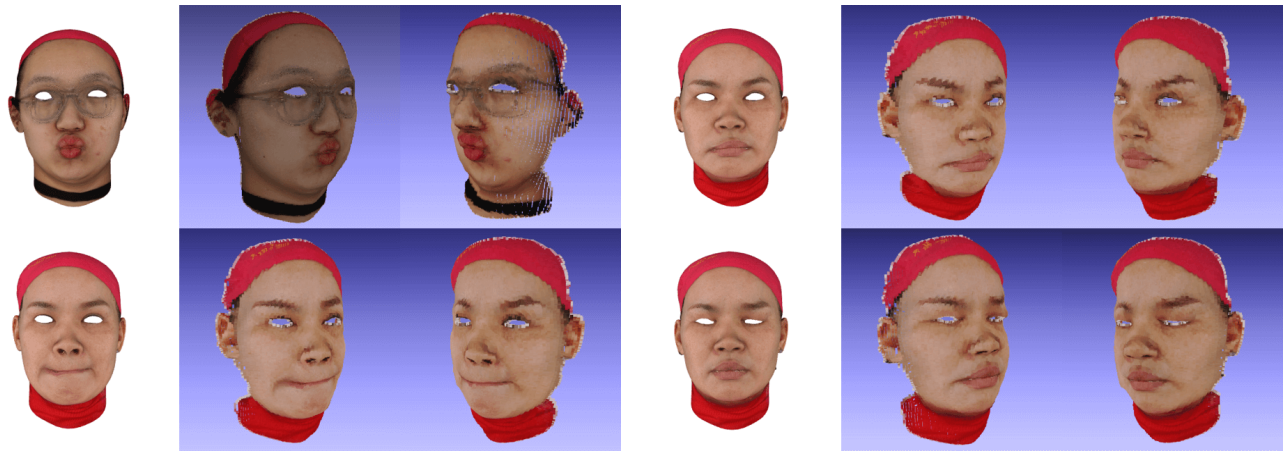

Figure 7: Visual results on Facescape face dataset. For each sample, we show the input image at the first column, and the 3D reconstruction results at different viewpoints at the second and third columns.

and spatial loss, the 3D shape is able to be well recovered and the details are able to be reproduced, such as in nose and mouth. we also validate the accuracy of our proposed method using Facescape dataset,. As shown in Fig. 7, the four input images are similar but the reconstructed 3D point clouds can clearly be distinguished by different expression feature from their eyebrow, eye and mouth.

5.5 Quantitative Result

We first demonstrate the effect of the number of the facial landmarks to the accuracy of the depth estimation result. We report the MSE between the ground truth depth map that is directly projected from the 3D face model and the estimated depth map. As shown in Table 1. We observe that from 68 to 568 landmarks, the MSE achieves a significant decrease from 20.15 to 4.26 on Stirling ESRC 3D dataset and from 19.72 to 7.09 on facescape dataset, which demonstrates the effectiveness of the proposed pipeline.

To evaluate the effects of each component and the proposed loss constraint, we conduct an ablation analysis on both Stirling ESRC 3D face dataset and Facescape dataset, as shown in Fig. 10. We visually compare the depth maps with only the proposed spectral appearance consistency loss, only bundle-adjustment consistency loss based on 68 landmarks, and full pipeline with 568 landmarks. It can be observed that depth information becomes more consistent and smooth from left to right. As shown in first column, due to the similarity of face's pixel values, only using spectral appearance consistency is unable to recover depth accurately. If only using bundle-adjustment consistency loss on 68 landmarks, extensive face depth details (e.g., regions around nose) are unable to be recovered since there are few or even no landmarks in specific regions, as shown in second column. Therefore, to improve the accuracy, both spectral and spatial constraints relying on the massive detected landmarks are introduced to learn facial depth information. As shown in fourth column, the depth maps are complete, consistent and smooth. The largest accuracy improvement benefits from the proposed landmark consistency loss across multiple-view frames. The full pipeline with all the constraints achieves the best performance on both datasets.

5.6 Comparison with State-of-the-art Methods

We provide a visual comparison on Stirling ESRC 3D face and Facescape dataset with other recent state-of-the-art face reconstruction methods. Each method is fed with the input image to generate a 3D face model to get a fair comparison. We compare our network with 3DDFA [35], Pix2Face [2], VRN [12], 3DDFA_v2 [9]

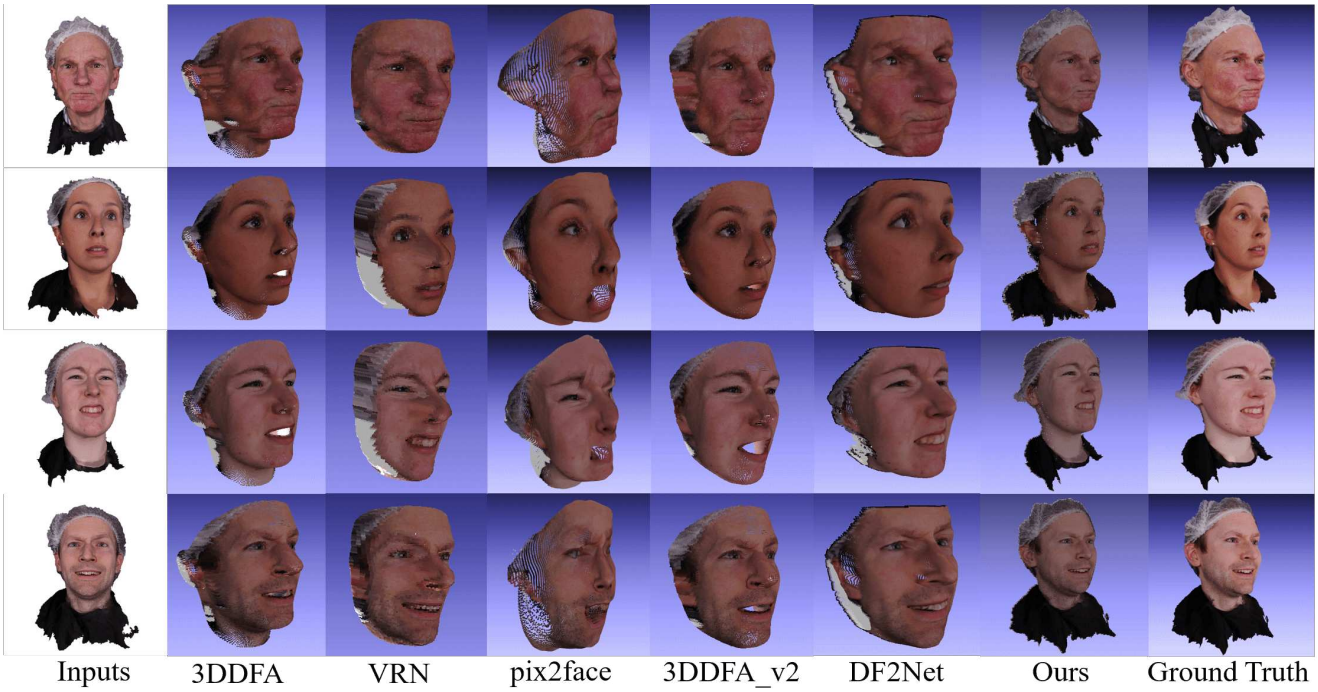

Figure 8: Visual comparisons on reconstructed face shape from Stirling ESRC 3D face dataset between our result and other recent methods. First column: raw input image; second column: result from 3DDFA method; third column:result from VRN method; fourth column: result from pix2face method; fifth column: result from 3DDFA_v2; sixth column: result from DF2Net method; Seventh column: result from our pipeline; Eighth column: ground truth.

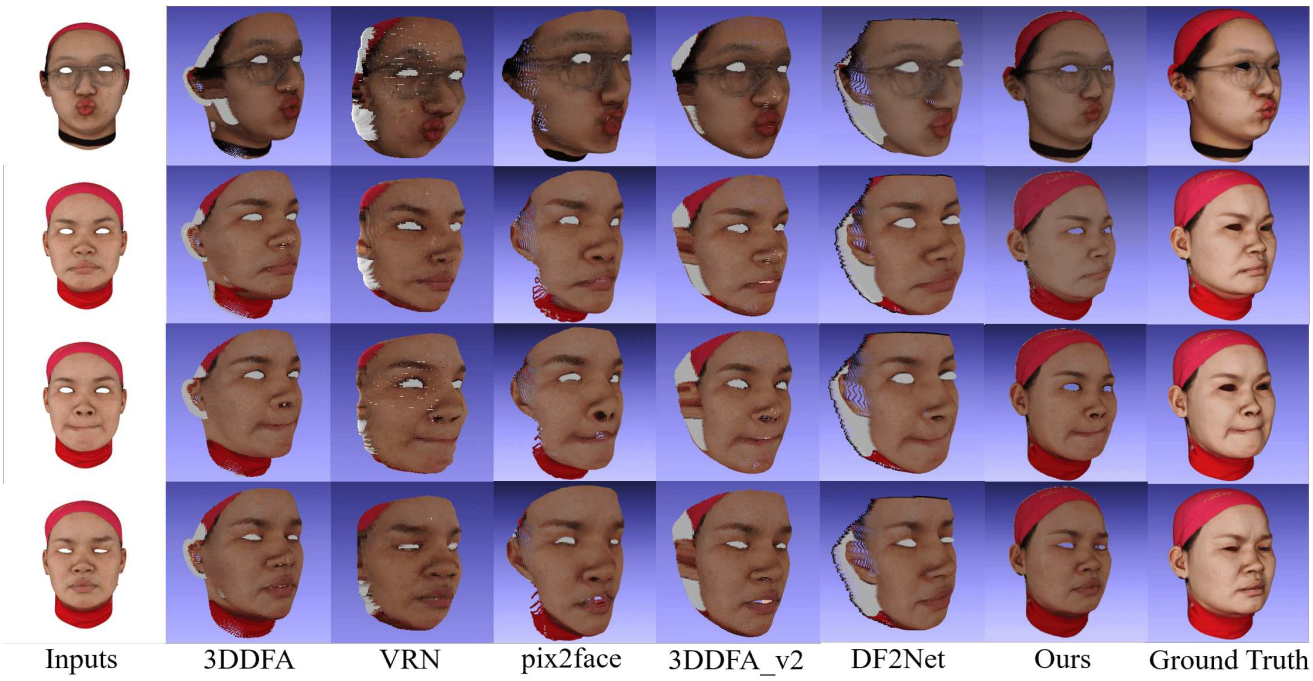

Figure 9: Visual comparisons on reconstructed face shape from facescape dataset between our result and other recent methods. First column: raw input image; second column: result from 3DDFA method; third column:result from VRN method; fourth column: result from pix2face method; fifth column: result from 3DDFA_v2; sixth column: result from DF2Net method; Seventh column: result from our pipeline; Eighth column: ground truth.

	MSE					
	68 landmarks	200 landmarks	300 landmarks	400 landmarks	500 landmarks	568 landmarks
Stirling	17.15	16.83	12.31	9.65	6.89	4.26
Facescape	19.72	18.81	16.30	12.48	10.78	7.09

Table 1: Mean Squared Error using different number of landmarks to train network on Stirling ESRC 3D face and Facescape datasets(in mm). We use ICP for alignment and compute point-to-point distance between our results and ground truth.

	MSE	
	Stirling	Facescape
VRN [12]	17.32	20.89
pix2face[2]	11.74	25.84
3DDFA [35]	8.82	15.45
3DDFA_v2 [9]	5.08	13.52
DF2Net [32]	4.58	8.14
Ours	**4.26**	**7.09**

Table 2: Mean Squared Error subjects on Stirling ESRC 3D face dataset and Facescape dataset(in mm).

and DF2Net [32]. It can be observed in Fig. 8 and Fig. 9, our method achieves more realistic and accurate reconstruction results than most methods in shape. In Stirling ESRC 3D face dataset, it is especially obvious that our 3D reconstructed shape outperforms VRN and pix2face methods. VRN method produce a lot of invalid information when generating 3D shape and their nose scale is larger than ground truth. The 3D reconstructed shape generated by Pix2face method exists distortion and deformation, such as nose, mouth and eyebrow. Compared to 3DDFA and 3DDFA_v2 outcomes, our reconstructed 3D face model fully retains the input face image information. For example, 3DDFA and 3DDFA_v2 respectively ignore the teeth regions in the input image, with significant distortion and deformation on the ear parts, even missing in some reconstruction examples. In addition, our 3D shapes are more real overall. For DF2Net method, although their 3D shaping effects are close to ours, the 3D reconstruction of this method in the ear is very poor, which is generally the most difficult components to reconstruct due to the significant depth changes.

In the Facescape dataset, our method presents more detailed information than other methods. For example, the ground truth shapes of the second and the fourth rows are very similar. The main differences are the different degrees of eye and mouth openness. The method we propose can better present these differences. Other methods have produced more blur and ambiguity because they could not establish effective constraints in details to recover the subtle changes. In addition, the 3D shapes generated by VRN, pix2face and 3DDFA_v2 are quite different from the ground truth. It is difficult to correspond the 3D shapes generated by their methods to respective ground truth face models. The DF2Net method loses significant information of the ear, and the depth values of the eye are also significantly wrong (white convex). However, our method can correctly estimate the depth of each organ based on the precise spectral and spatial constraints. Overall, our method preserves the complete input image information in the 3D reconstruction face models compared with other state-of-the-art methods.

To obtain quantitative evaluation, we compare the reconstructed 3D point clouds with ground truth 3D models and report the averaged point-to-point errors. The results are shown in Table 2. On both Stirling ESRC 3D face and Facescape datasets, our proposed method achieves the lowest average Mean Squared Error (MSE) across all test objects. Especially, our method achieves error of 4.26 lower than other methods (separately DF2Net : 4.58; 3DFFA : 8.82; 3DFFA_v2 : 5.08; pix2face : 11.74; VRN : 17.32) on Stirling ESRC 3D face and error of 7.09 much lower than other methods (separately DF2Net : 8.14; 3DFFA: 15.45;3DFFA_v2: 13.52;pix2face: 25.84;VRN: 20.89) on Facescape dataset.

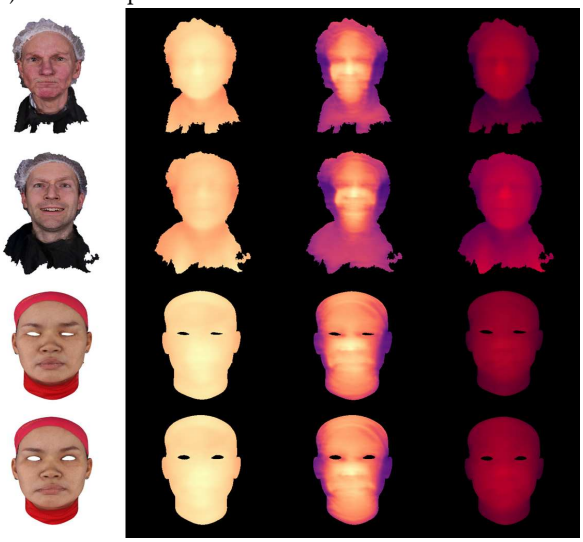

Figure 10: Ablation study from depth maps on the Stirling ESRC 3D face dataset (top two) and Facescape dataset (bottom two) respectively . Left to right: raw input image; result with only the designed spectral consistency loss;result with only the designed bundle-adjustment 68 landmark consistency loss; result in full loss with 568 landmark.

6 CONCLUSION

This paper develops a deep SfM framework targeting at 3D face reconstruction. Bundle adjustment is incorporated in the CNN training scheme to explore multiple view geometry relationship across frames captured from different perspectives of the faces. To establish reliable spatial constraints, a massive face landmark detection method is developed that can detect 500 or even 800 landmarks with associated unique ID for each landmark. Spectral constraints are further introduced to enhance the network training effect, which reduces the color difference of pixels corresponding to the same 3D point. Once the landmark detection is trained, the SfM network training is unsupervised, which mitigates the labeling efforts. Even with the widely used 68 landmarks, our SfM network still achieves extraordinary 3D reconstruction accuracy.

7 ACKNOWLEDGE

This material is based upon work supported by the National Science Foundation under Award No. 2105257.

REFERENCES

[1] P Anandan, JR Bergen, KJ Hanna, and Rajesh Hingorani. 1993. Hierarchical model-based motion estimation. In *Motion analysis and image sequence processing*. Springer, 1–22.

[2] Daniel Crispell and Maxim Bazik. 2017. Pix2face: Direct 3D face model estimation. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2512–2518.

[3] M Dawson-Haggerty. 2017. Trimesh.

[4] Pengfei Dou and Ioannis A Kakadiaris. 2011. Multi-view 3d face reconstruction with deep recurrent neural networks. (2011).

[5] Pengfei Dou, Shishir K Shah, and Ioannis A Kakadiaris. 2017. End-to-end 3D face reconstruction with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5908–5917.

[6] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. 2018. Evaluation of dense 3D reconstruction from 2D face images in the wild. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 780–786.

[7] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 270–279.

[8] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. 2019. Digging into self-supervised monocular depth estimation. *ICCV* (2019).

[9] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. *arXiv preprint arXiv:2009.09960* (2020).

[10] Y. Guo, j. zhang, J. Cai, B. Jiang, and J. Zheng. 2019. CNN-Based Real-Time Dense Face Reconstruction with Inverse-Rendered Photo-Realistic Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 6 (2019), 1294–1307.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

[12] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. 2017. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proceedings of the IEEE International Conference on Computer Vision*. 1031–1039.

[13] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[14] Geng Liang, Zhan Shu, and Jiang Jianguo. 2016. Fusing deep convolutional network with SFM for 3D face reconstruction. In *2016 IEEE 13th International Conference on Signal Processing (ICSP)*. IEEE, 873–878.

[15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[16] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 2 (2004), 91–110.

[17] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *ICML*.

[18] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. 2019. 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*. 7763–7772.

[19] Silvio Savarese and Pietro Perona. 2001. Local analysis for 3D reconstruction of specular surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[20] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4104–4113.

[21] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2019. Fml: Face model learning from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10812–10822.

[22] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 1274–1283.

[23] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. 2017. 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*. 5163–5172.

[24] Ryan White and David A Forsyth. 2006. Combining cues: Shape from shading and texture. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

[25] Changchang Wu et al. 2011. VisualSFM: A visual structure from motion system. (2011).

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

[27] X. Shao Y. Wang Y. Feng, F. Wu and X. Zhou. 2017. A 3d morphable model of craniofacial shape and texture variation. In *Proceedings of the IEEE International Conference on Computer Vision*.

[28] X. Shao Y. Wang Y. Feng, F. Wu and X. Zhou. 2018. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV2018*.

[29] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 601–610.

[30] Hongwei Yi, Chen Li, Qiong Cao, Xiaoyong Shen, Sheng Li, Guoping Wang, and Yu-Wing Tai. 2019. Mmface: A multi-metric regression network for unconstrained face reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7663–7672.

[31] Zhichao Yin and Jianping Shi. 2018. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1983–1992.

[32] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. 2019. DF2Net: A Dense-Fine-Finer Network for Detailed 3D Face Reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*. 2315–2324.

[33] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. 2017. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1851–1858.

[34] Yuxiang Zhou, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. 2019. Dense 3d face decoding over 2500fps: Joint texture & shape convolutional mesh decoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1097–1106.

[35] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z Li. 2017. Face alignment in full pose range: A 3d total solution. *IEEE transactions on pattern analysis and machine intelligence* 41, 1 (2017), 78–92.