Multi-view 3D Morphable Face Reconstruction via Canonical Volume Fusion

Jingqi Tian[1], Zhibo Wang[1], Ming Lu[2], and Feng Xu[1](✉)

[1] School of Software and BNRist, Tsinghua University, Beijing, China
xufeng2003@gmail.com
[2] Intel Labs, Beijing, China

Abstract. Due to the capability of easy animation and editing of faces, 3D Morphable Model (3DMM) is widely used in the task of face reconstruction. Recent methods recover 3DMM coefficients by fusing the information from a set of multi-view images via end-to-end Convolutional Neural Networks (CNNs), which alleviate the inherent depth ambiguity in the single-view setting. However, most of these methods fuse global features of all views to regress the 3D morphable face, without considering the dense correspondences of multi-view images. In this paper, we propose a novel approach to reconstruct high-quality 3D morphable faces. We first use a canonical feature volume to fuse multiple view features in 3D space, which establish dense correspondences between different views. Next, to bridge the gap between CNN regression and pixel-wise optimization and further leverage the muti-view information, we propose test-time optimization to improve the regressed results with negligible additional cost. Our method achieves the state-of-the-art performance on widely-used benchmarks, demonstrating the effectiveness of our approach. Code will be released.

Keywords: Multi-view 3D face reconstruction · 3D morphable model

1 Introduction

3D face reconstruction from images is a crucial problem in computer vision and has a wide range of applications such as face tracking [4,5], portrait relighting [41], gaze tracking [42], face reenactment [7,19,36] and so on. In order to address the difficulties in image-based face reconstruction, 3D Morphable Model (3DMM) is often adopted to provide a low-dimensional parametric representation of 3D face. Traditional methods recover the 3DMM coefficients by solving a costly nonlinear optimization problem and require a good initialization. In contrast, recent methods [9,13,15,18,30,34,35,39,43–45] adopt deep Convolutional

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-20500-2_45.

L. Fang et al. (Eds.): CICAI 2022, LNAI 13605, pp. 545–558, 2022.
https://doi.org/10.1007/978-3-031-20500-2_45

Neural Network (CNN) to directly learn the mapping between 2D images and 3DMM coefficients. Single-view face reconstruction [13,15,18,34,35,39,43,45] has been extensively studied in recent years, where an inherent difficulty is the ambiguity of depth estimation, especially in the forehead, nose and chin regions.

Compared with the single-view face reconstruction, multi-view face reconstruction [10,27,31,44] can effectively resolve the depth ambiguity. However, most of existing works [27,31,44] simply extend the techniques of single-view reconstruction to the multi-view setting. After carefully studying the pipeline of existing methods [10,27,31,44], we find that these methods mostly fuse the 2D global features extracted from different views to regress the 3D morphable face. However, the fusion of 2D global features cannot learn sufficient representation for 3D reconstruction.

In this paper, we propose a novel method for multi-view 3D morphable face reconstruction based on canonical volume fusion. Our method extracts the 3D feature volumes from multi-view images. As 3D volumes allow easy alignment of facial features in 3D space, we transform the volumes of multiple views to align with the canonical volume by the estimated head pose parameters. To fuse the transformed 3D feature volumes, our method adopt a confidence estimator to predict the confidences of the multi-view feature volumes. Therefore, the transformed feature volumes can be adaptively fused according to the estimated confidence volumes. This is essential for multi-view feature fusion since faces under different poses provide partial information of the 3D face. The fused canonical feature volume is used to regress the shape and texture coefficients. Compared with existing methods [31,44], our work can establish better dense correspondences between different views and generate more accurate 3D reconstruction.

CNNs can directly and efficiently estimate the 3DMM coefficients, but it tends to predict reasonable but not pixel-wise accurate results, as it is trained to achieve the lowest average error over the entire dataset, not a particular sample. On the other hand, optimization fits the parametric model to multi-view images of a particular sample. However, it is sensitive to the initialization, and may fall into local minimums or take very long time without a good initialization. Therefore, the multi-view information of a particular sample may not be fully explored by the inference of the network. Directly involving multi-view constraints also in the testing rather than just in the training may further improve the results. We propose to introduce test-time optimization to CNN-based regression. Our test-time optimization can leverage the benefits of both paradigms. Specifically, we use the CNN regressed estimation to initialize the iterative optimization process, making the fitting stable and faster. We find this idea is simple but effective to bridge the gap between training and testing.

2 Related Work

2.1 3D Morphable Model (3DMM)

Since the seminal work [2], 3D morphable models have been widely used in face reconstruction over the past twenty years. [2] proposes to derive a morphable face model by transforming the shape and texture of the captured 3D faces into a latent space using Principal Component Analysis (PCA). 3D faces can be modeled by the linear combinations of PCA basis. [6] uses Kinect to capture 150 individuals aged 7–80 from various ethnic backgrounds. For each person, they capture the neutral expression and 19 other expressions. Bilinear face model is constructed by N-mode Singular Value Decomposition (SVD). [25] combines the linear shape space with an articulated jaw, neck, and eyeballs, pose-dependent corrective blendshapes, and additional global expression blendshapes. They can fit better to the static 3D scans and 4D sequences using the same optimization method compared with [2,6]. For a detailed survey of 3DMM over the past twenty years, we refer the readers to [12].

2.2 3D Face Reconstruction

With the help of 3DMM, the face reconstruction task can be formulated as a cost minimization problem [2]. Due to the nonlinearity of the optimization problem, it is time-consuming to optimize the coefficients of 3DMM. Therefore, numerous regression-based methods are proposed to employ convolutional neural network for face reconstruction. The biggest obstacle when applying deep learning to face reconstruction is the lack of training data. [45] proposes a face profiling technique which can generate synthetic images with the same identity but different face poses as the original images. They utilized their face profiling technique to create the 300W-LP database and trained a cascaded CNN to regress 3DMM coefficients. [11] utilizes publicly available 3D scans to render more realistic images. Recently, the self-supervision approaches are becoming prevailing. [15,34] enable the self-supervised training by introducing a differentiable rendering layer. This self-supervision scheme has been widely used in the following works [8,9,20,22,24,29,33,37,38].

Compared with single-view face reconstruction, multi-view face reconstruction can effectively resolve the depth ambiguity. Multi-view setting ensures that the faces in different views are geometrically consistent. There are several approaches [10,27,31,44] to study the multi-view face reconstruction. [10] proposes to address the problem using CNNs together with recurrent neural networks (RNNs). However, it is not reasonable to model the task with RNNs, and multi-view geometric constraints are not exploited in their approach. [44] adopts photometric loss and alignment loss to explicitly incorporate multi-view geometric constraints between different views. [30] further leverages multi-view geometry consistency to mitigate the ambiguity from monocular face pose estimation and depth reconstruction in the training process. However, the above methods [10,27,31,44] follow the network design of single view face reconstruction and fail to learn sufficient representation for 3D reconstruction.

3 Preliminaries

3.1 Face Model

With a 3DMM, the face shape S and texture T can be represented as a linear combination of shape and texture bases:

$$S = \overline{S} + B_{id}\alpha + B_{exp}\beta \tag{1}$$

$$T = \overline{T} + B_t\delta \tag{2}$$

where $\overline{S}$ and $\overline{T}$ are the mean shape and texture respectively. B_{id}, B_{exp} and B_t denote the PCA bases of identity, expression and texture. α, β and δ are corresponding coefficients to be estimated. All of bases are scaled with their standard deviations. In our method, $\overline{S}, B_{id}, \overline{T}, B_t$ are constructed from Basel Face Model (BFM) [26] and B_{exp} is constructed from FaceWareHouse [6]. We adopt the first 80 bases with the largest standard deviation for identity and texture, the first 64 bases for the expression bases.

3.2 Camera Model

We employ the perspective camera model to define the 3D-2D projection. The focal length of the perspective camera is selected empirically. The face pose P is represented by an Euler angle rotation $R \in SO(3)$ and translation $t \in \mathbb{R}^3$.

3.3 Illumination Model

We model the lighting by Spherical Harmonics(SH) and assume a Lambertian surface for face. Given the surface normal n_i and face texture t_i, the color can be computed as $C(n_i, t_i|\gamma) = t_i \cdot \sum_{b=1}^{B^2} \gamma_b \Phi_b(n_i)$. $\Phi_b : \mathbb{R}^3 \rightarrow \mathbb{R}$ is SH basis function and we choose the first $B^2 = 9$ functions following [34,35]. $\gamma \in \mathbb{R}^{27}$ represents the colored illumination in red, green and blue channels.

Our method can take any number of multi-view images of the same person $\{I_i\}_{i=1}^n$ as input and output the corresponding coefficients $\{x_i\}_{i=1}^n$ of these images, where $x_i = \{\alpha, \beta, \delta, P_i, \gamma_i\}$. It should be noticed that α, β, δ are shared by all images and P_i, γ_i are variant across the input multi-view images.

4 Method

Our method aims to regress 3DMM coefficients by leveraging the dense correspondences of the multi-view facial images of one subject. Therefore, we propose a Canonical Volume Fusion Network whose architectures are designed to integrate the dense information from different views. As shown in Figure 1 (a), our network first extracts 3D feature volumes from input images. Then, the dense

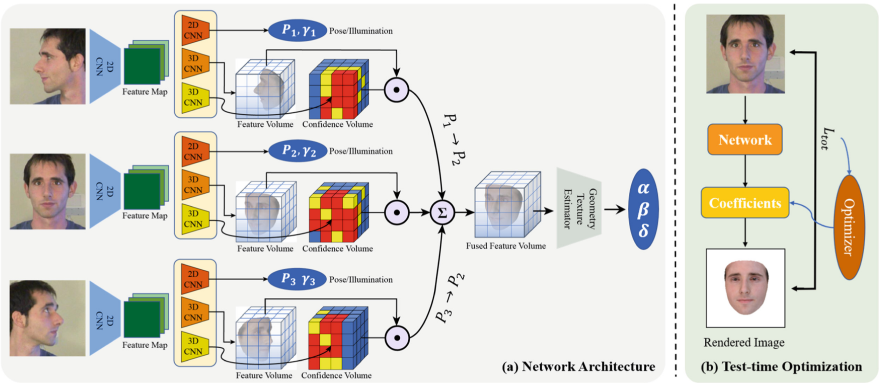

Fig. 1. Overview of our approach. (a) The network architecture of our method. (b) The test-time optimization mechanism.

feature volumes are transformed to a canonical coordinate system through feature volume alignment. Next, the aligned feature volumes are fused together in a confidence-aware manner. From the fused feature volumes, a shape/texture estimator is trained to output 3DMM coefficients. During testing, we apply test-time optimization to further improve performance, as shown in Figure 1 (b).

4.1 Canonical Volume Fusion Network

Feature Extraction. Previous methods mostly use 2D CNN backbone such as VGG-Face [32] or ResNet [16] to regress 3DMM coefficients. However, as human faces are 3D objects, it is more intuitive to model the facial correspondences in 3D space. We employ a 2D-3D feature extraction network to map a 2D face image to a 3D feature volume. Several 2D downsampling convolutional blocks extract a 2D feature map f_{2D} from the input image. Then, we utilize a "reshape" operation to project 2D feature maps to 3D feature volumes. The following 3D CNN finally extracts the 3D feature volume f_{3D}.

Volume Feature Alignment. Pose and illumination coefficients are private for each multi-view image. We regress these coefficients from f_{2D} separately. The f_{2D} is pooled to a 512-dimensional feature vector and sent through several linear layers. The 3D feature volumes extracted from multi-view images are semantically misaligned. It is unreasonable to fuse them directly and this is also the main drawback of previous work [44]. We align the 3D feature volumes extracted from multi-view images according to the estimated pose via the following equation:

$$p_d \sim T_{m\rightarrow NDC}(R_d R_s^{-1}(T_{NDC\rightarrow m}(p_s) - t_s) + t_d) \tag{3}$$

where subscript s and t represent source image and target image respectively, p is a coordinate in the feature volume, R, t are the face pose rotation and translation in the image, $T_{NDC\to m}(\cdot)$ is the coordinate transformation from the normalized device coordinate (NDC) system to model coordinate system. The f_{3D} extracted from images is assumed to be aligned with the NDC system. Therefore, we first convert the coordinate system to the model coordinate system. For any coordinate p_s in the feature volume of source image, we can compute the corresponding coordinate p_t in the feature volume of target images by Eq. (3). In practice, we align other feature volumes to the feature volume of the pre-selected frontal view image.

Confidence-Aware Feature Fusion. The input images taken from different views have different confidence and quality in the different face region. For example, the left view image has the low confidence and quality in the right face region. Therefore, we use a confidence estimator to learn the measurement of confidence and quality of the feature volume. The estimator is similar to the 3D CNN used for feature extraction but more lightweight. It outputs a 3D volume $c \in R^{h\times w\times d}$ with positive elements. c_i has the same height, width and depth as the f_{3D}. The feature can be fused via the following equation:

$$f_{3D,fuse} = \sum_i c_i \odot f_{3D,i} / \sum_i c_i \tag{4}$$

where $f_{3D,i}$ donates the 3D feature extracted from image I_i and c_i is the confidence of $f_{3D,i}$.

Coefficients Estimator. The method of estimating pose and illumination coefficients from f_{2D} has been introduced in the previous section. The shape and texture coefficients will be estimated from $f_{3D,fuse}$. Inspired by [40], we implement a similar keypoints detector, which extracts K 3D keypoints $\{x_i\}_{i=1}^{K}$ in feature volume. These keypoints are unsupervisedly learned and different from the common facial landmarks. The feature at the keypoint location is considered to have main contribution to shape and texture estimation. We conduct bilinear sampling operation at the keypoints locations of $f_{3D,fuse}$ to obtain the local feature f_{loc} and apply a 3D average pooling operation over the $f_{3D,fuse}$ to obtain the global feature f_{glo}. The f_{loc} and the f_{glo} are concatenated to regress the shape and texture coefficients by several linear layers.

4.2 Loss Function

Single-view face reconstruction has been widely studied. Therefore, We transfer the loss function used in single-view face reconstruction method to the multi-view setting.

Photometric Loss. The photometric loss aims to minimize the pixel difference between the input images and the rendered images, defined as $L_{photo} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{M}_i|}\sum_{\mathcal{M}_i}||I_i'(x_i) - I_i||_2$, where the $I_i'(x_i)$ is the image rendered using

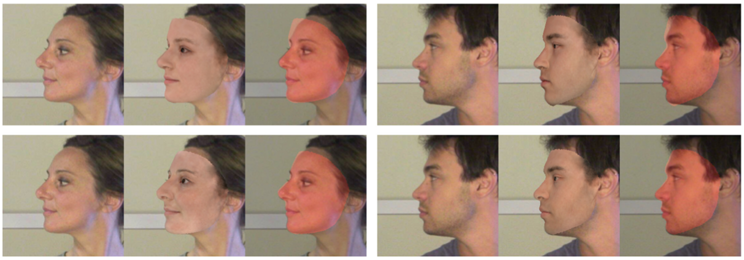

Fig. 2. Comparison of the results without (top row) and with (bottom row) using silhouette loss for training. We use the red region to mark the face region of rendered images on the input images. (Color figure online)

the face model coefficients x_i, $\mathcal{M}_i$ is the face region of $I'_i(x_i)$ and N is the number of different view images.

Landmark Loss. The landmark loss mainly contributes to the geometry of reconstructed face. We use a state-of-the-art landmark detector [3] to detect the 68 landmarks $\{q_i^k\}_{k=1}^{68}$ of input image I_i. We also can obtain the landmarks $\{{q'}_i^k(x_i)\}_{k=1}^{68}$ by projecting the 3D vertices on the reconstructed mesh to image plane. The landmark loss can be represented as: $L_{lmk} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{68}\sum_{k=1}^{68}\omega_k||q_i^k - {q'}_i^k(x_i)||_2$, where ω_k is the landmark weight. We set the weight to 20 for nose and inner month and others to 1.

Perceptual Loss. We adopt the perceptual loss L_{per} as in [9] to improve the fidelity of the reconstructed face texture. The perceptual loss measures the cosine distance between the deep feature of the input images and rendered images. With the perceptual loss, the textures are sharper and the shapes are more faithful.

Silhouette Loss. Inspired by the silhouette loss which used in human body reconstruction [14,17,21], we apply it in multi-view face reconstruction task. We use a face parsing network [23] to segment the face region from the input image. Then we detect the side view silhouette (left silhouette for left view image and right for right) of the face region. The silhouette is represented as a 2D point set $\mathcal{S}_i$ in the image plane, where i is the index of input image I_i. We can also extract silhouette from the rendered image $I'_i(x_i)$ to get another point set $\mathcal{S}'_i$. The silhouette loss is defined as the chamfer distance between the two point sets: $L_{sil} = \frac{1}{N}\sum_{i=1}^{N} chamfer(\mathcal{S}_i, \mathcal{S}'_i)$. It should be noticed that the silhouette loss will not be applied in the frontal view images. In the experiment, the face parsing network may fail due to the occlusion of the face region. Therefore, we discard the silhouette loss when its value is greater than a presetting threshold to make the training process more stable. Figure 2 illustrates the benefit of using our silhouette loss.

Regularization Loss. To ensure the face geometry and texture are reasonable, regularization loss of 3DMM is used as $L_{reg} = \omega_{id}||\alpha||_2 + \omega_{exp}||\beta||_2 + \omega_{tex}||\delta||_2$. $\omega_{id}, \omega_{exp}, \omega_{tex}$ are balancing weights of different 3DMM coefficients and are set to 1.0, 0.8, 2e-3 respectively.

To sum, the total loss function is:

$$L_{tot} = \omega_{pho}L_{pho} + \omega_{lmk}L_{lmk} + \omega_{per}L_{per} + \omega_{sil}L_{sil} + \omega_{reg}L_{reg} \tag{5}$$

4.3 Test-Time Optimization

The CNN-based approaches predict the face model coefficients x from image I by learning a mapping function $x = f_\theta(I)$, where θ is the parameters of CNNs. Assuming that the CNNs is trained on a dataset $\mathcal{D}_{train}$, the training process aims to find the optimal parameters θ^* which satisfies:

$$\theta^* = \arg\min_\theta \sum_{I \in \mathcal{D}_{train}} L_{tot}(I, f_\theta(I)) \tag{6}$$

However, when testing a particular sample I, we want to find the face model coefficients x^* which satisfies:

$$x^* = \arg\min_x L_{tot}(I, x) \tag{7}$$

There are two main gaps between Eq. (6) and Eq. (7). The first one is the test image may not be sampled from $\mathcal{D}_{train}$. This is a crucial but difficult problem caused by domain gap between datasets and is still a hotspot issue in deep learning. The second gap is neural network minimizes the loss over the whole dataset. Although we test a sample $I \in \mathcal{D}_{train}$, the neural network still can't produce a optimal result for this particular sample. Thus, we propose the test-time optimization mechanism to fill the two gaps. We take the output of neural network $x = f_\theta(I)$ as the initialization and try to find the optimal x^* by Eq. (7). Our test-time mechanism can be easily implemented in the existing reconstruction methods based on neural network, which only need to calculate derivative of $L_{tot}(I, x)$ with respect to x and conduct gradient descent algorithm.

5 Experiments

In this section, we compare the qualitative and quantitative result with both the state-of-the-art single-view and multi-view approaches. We also demonstrate the effectiveness of our approach with extensive ablation studies in the Supplementary Material. Besides, the implementation details including training strategy, datasets and hyperparameters setting will also be showed in the Supplementary Material.

Table 1. Average and standard deviation of the symmetric point-to-plane L2 errors on the MICC dataset (in *mm*).

Method	Cooperative		Indoor	
	Mean	Std	Mean	Std
3DDFA [45]	2.65	0.63	2.26	0.50
RingNet [28]	2.35	0.49	2.21	0.46
PRN [13]	2.30	0.54	2.02	0.50
Tran *et al.* [39]	2.05	0.54	2.07	0.51
MVFNet [44]	1.73	0.49	1.76	0.52
MGCNet [30]	1.71	0.47	1.73	0.48
Deng *et al.* [9]	1.69	0.53	1.70	0.51
Ours	**1.59**	**0.47**	**1.61**	**0.46**

5.1 Quantitative Comparisons

We evaluation our approaches on the MICC Florence dataset [1] which is a benchmark test dataset of the multi-view face reconstruction task. It consists of 53 identities and the corresponding 3D scans which can be regarded as the ground-truth. Each identity has two videos of "indoor-cooperative" and "indoor" respectively. We compare our methods with both multi-view methods and single-view methods. For multi-view methods, we manually select three images in each video as a triplet, where the camera viewpoints are largely different and expressions are kept neutral very well. For comparing with the single-view methods on the image triplets, we follow the method from [30,44]. We follow the data preprocessing methods and evaluation metrics from [15,44]. Then the symmetric point-to-plane L2 errors (in millimeters) between the predict 3D models and the groundtruth scan will be computed as the evaluation metrics.

We compare our method with Zhu *et al.* [45] (3DDFA), Sanyal *et al.* [28], Feng *et al.* [13] (PRN), Tran *et al.* [39], Wu *et al.* [44] (MVFNet), Shang*et al.* [30] (MGCNet), Deng *et al.* [9]. Notice that for each comparison, we use exactly the same input to test all the comapred methods. As shown in Table 1, our method outperforms all the state-of-the-art single-view and multi-view methods. Several examples of the comparison of the error maps are shown in Fig. 3. Since our method better explore the multi-view 3D information by a 3D volume-based feature fusion and a test-time optimization, it achieves lower error than the compared methods especially in the regions of forehead and chin where the z direction ambiguity is more severe.

5.2 Qualitative Comparisons

We present some visual examples from the MICC dataset. We compared our methods with RingNet [28], Deng *et al.* [9], MVFNet [44] and MGCNet [30].

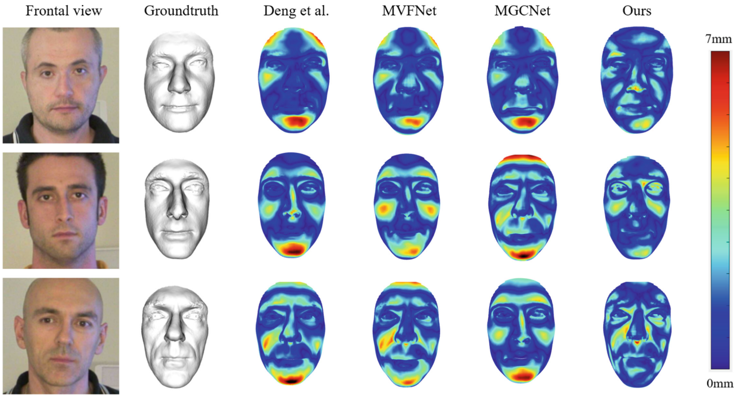

Fig. 3. The error map comparisons with Deng et al. [9], MVFNet [44], MGCNet [30] on the MICC dataset.

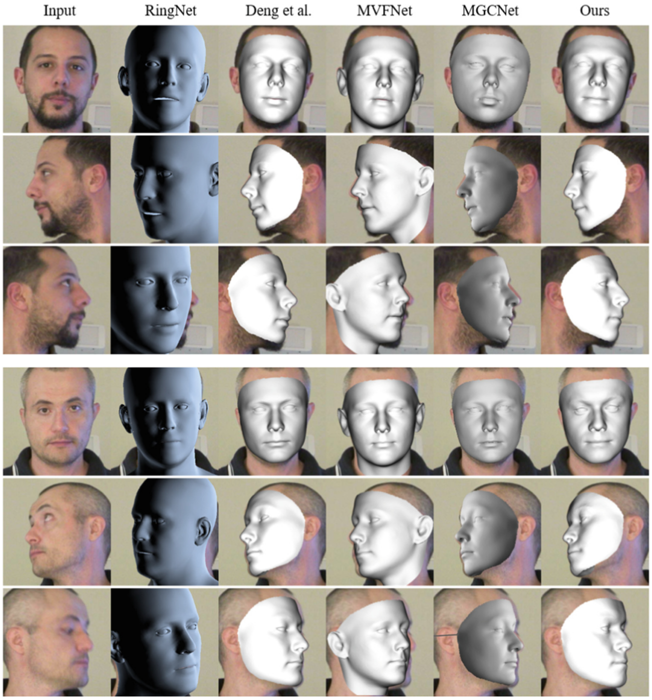

Fig. 4. Geometry comparisons with RingNet [28], Dent et al. [9], MVFNet [44], MGCNet [30] on the MICC dataset.

From the front view images, Fig. 4 shows that the overall face shapes reconstructed by our method and Deng *et al.* [9] are more fidelity than the other methods. For the side views images, although MGCNet [30] and Deng *el al.*

[9] have achieved better pose estimation than MVFNet [44] and RingNet [28], there still exits obvious misalignment at the forehead region. While our method achieves better face alignment than the other methods by better exploring of multi-view information. More Visual comparisons in different facial expressions will be showed in Supplementary Material.

6 Conclusion

We have proposed a novel multi-view 3D morphable face reconstruction method via canonical volume fusion and demonstrated the advantages of explicitly establishing dense feature correspondences to solve the depth ambiguity in the multi-view reconstruction task. Besides, we introduced an easy-implemented and effective mechanism called test-time optimization, which refines the outputs of CNNs and obtain more accurate results. Our methods outperforms the state-of-art methods in both quantitative and qualitative.

Acknowledgement. This work was supported by Beijing Natural Science Foundation (JQ19015), the NSFC (No. 62021002, 61727808), the National Key R&D Program of China (2018YFA0704000), and the Key Research and Development Project of Tibet Autonomous Region (XZ202101ZY0019G). This work was also supported by THUIBCS, Tsinghua University, and BLBCI, Beijing Municipal Education Commission.

References

1. 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, pp. 79–80 (2011)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–194 (1999)
3. 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)
4. Cao, C., Hou, Q., Zhou, K.: Displaced dynamic expression regression for real-time facial tracking and animation. ACM Trans. Graph. (TOG) **33**(4), 1–10 (2014)
5. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3d shape regression for real-time facial animation. ACM Trans. Graph. (TOG) **32**(4), 1–10 (2013)
6. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: a 3d facial expression database for visual computing. IEEE Trans. Vis. Comput. Graph. **20**(3), 413–425 (2013)
7. Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K.: Real-time facial animation with image-based dynamic avatars. ACM Trans. Graph. **35**(4) (2016)
8. Chaudhuri, B., Vesdapunt, N., Shapiro, L., Wang, B.: Personalized face modeling for improved face reconstruction and motion retargeting. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12350, pp. 142–160. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_9

9. 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/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
10. Dou, P., Kakadiaris, I.A.: Multi-view 3d face reconstruction with deep recurrent neural networks. Image Vis. Comput. **80**, 80–91 (2018)
11. 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)
12. Egger, B., et al.: 3d morphable face models-past, present, and future. ACM Trans. Graph. (TOG) **39**(5), 1–38 (2020)
13. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3D face reconstruction and dense alignment with position map regression network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 557–574. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_33
14. Gavrila, D.M., Davis, L.S.: 3-d model-based tracking of humans in action: a multiview approach. In: Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 73–80. IEEE (1996)
15. 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)
16. 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)
17. Huang, Y., et al.: Towards accurate marker-less human shape and pose estimation over time. In: 2017 International Conference on 3D Vision (3DV), pp. 421–430. IEEE (2017)
18. Jourabloo, A., Liu, X.: Large-pose face alignment via CNN-based dense 3d model fitting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4188–4196 (2016)
19. Kim, H., et al.: Deep video portraits. ACM Trans. Graph. (TOG) **37**(4), 1–14 (2018)
20. Koizumi, T., Smith, W.A.P.: Look Ma, No Landmarks!– unsupervised, model-based dense face alignment. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 690–706. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_41
21. Lassner, C., Romero, J., Kiefel, M., Bogo, F., Black, M.J., Gehler, P.V.: Unite the people: Closing the loop between 3d and 2d human representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6050–6059 (2017)
22. Lattas, A., Moschoglou, S., Gecer, B., Ploumpis, S., Triantafyllou, V., Ghosh, A., Zafeiriou, S.: Avatarme: realistically renderable 3d facial reconstruction" in-the-wild". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 760–769 (2020)
23. Lee, C.H., Liu, Z., Wu, L., Luo, P.: Maskgan: towards diverse and interactive facial image manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5549–5558 (2020)
24. Lee, G.H., Lee, S.W.: Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6100–6109 (2020)

25. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4d scans. ACM Trans. Graph. **36**(6) (2017)
26. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 6th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 296–301. IEEE (2009)
27. Ramon, E., Escur, J., Giro-i Nieto, X.: Multi-view 3d face reconstruction in the wild using siamese networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
28. 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/CVF Conference on Computer Vision and Pattern Recognition, pp. 7763–7772 (2019)
29. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.W.: Sfsnet: learning shape, reflectance and illuminance of facesin the wild'. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6296–6305 (2018)
30. Shang, J., et al.: Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12360, pp. 53–70. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58555-6_4
31. Shao, X., et al.: 3d face shape regression from 2d videos with multi-reconstruction and mesh retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
32. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint. arXiv:1409.1556 (2014)
33. Tewari, A., et al.: Fml: face model learning from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10812–10822 (2019)
34. Tewari, A., et al: Mofa: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 1274–1283 (2017)
35. Tewari, A., et al.: 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)
36. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: real-time face capture and reenactment of RGB videos. In: Proceedings of the IEEE Conference on Computer Vision and pattern Recognition, pp. 2387–2395 (2016)
37. Tran, L., Liu, F., Liu, X.: Towards high-fidelity nonlinear 3d face morphable model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1126–1135 (2019)
38. 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)
39. 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)
40. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10039–10049 (2021)

41. Wang, Z., Yu, X., Lu, M., Wang, Q., Qian, C., Xu, F.: Single image portrait relighting via explicit multiple reflectance channel modeling. ACM Trans. Graph. (TOG) **39**(6), 1–13 (2020)
42. Wen, Q., et al.: Accurate real-time 3d gaze tracking using a lightweight eyeball calibration. In: Computer Graphics Forum, vol. 39, pp. 475–485. Wiley Online Library (2020)
43. Wen, Y., Liu, W., Raj, B., Singh, R.: Self-supervised 3d face reconstruction via conditional estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13289–13298 (2021)
44. Wu, F., et al.: Mvf-net: multi-view 3d face morphable model regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 959–968 (2019)
45. Zhu, X., Liu, X., Lei, Z., Li, S.Z.: Face alignment in full pose range: a 3d total solution. IEEE Trans. Pattern Anal. Mach. Intell. **41**(1), 78–92 (2017)