Learning Detailed 3D Face via CLIP Model from Monocular Image

Pengfei Zhou*
Shandong University of Science and Technology.

Yongtang Bao†
Shandong University of Science and Technology.

Yue Qi‡
Beihang University
Beihang University of
Qingdao Research Institute

2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) | 979-8-3503-4839-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/VRW58643.2023.00228

Abstract

3D morphable face models (3DMMs) methods cannot accurately estimate facial expressions and geometric details. We propose a framework for regressing 3D facial expressions and geometric details to address this problem. First, we propose a parameter refinement module to learn rich feature representations. Second, a novel feature consistency loss during training is designed, which exploits the powerful representation ability of CLIP (Contrastive-Language-Image-Pretraining) to capture facial expressions and geometric details. Finally, we leverage text-guided expression-specific transfer for 3D face reconstruction. Our method achieves significant performance in terms of reconstructed expressions and geometric details.

Index Terms: Human-centered computing—Computer vision and graphics—3D face reconstruction

1 Introduction

It is challenging to recover realistic 3D face shapes from single 2D images. Since the paired 3D data is not readily available, some unsupervised or weakly supervised learning approaches [1, 2] for the training process of 3D face reconstruction have obtained acceptable results. However, these methods can only reconstruct coarse geometry and texture information and cannot capture the geometric details. Some methods [3, 4] can recover the detailed facial shape via displacement map, but these methods still cannot accurately represent facial geometric details.

This paper proposes a method to recover realistic facial expressions and geometric details. Unlike previous work, we utilize the CLIP model as a supervision signal to encourage the similarity of geometric details between input and rendered images during training. Meanwhile, we propose a parameter refinement module to accelerate the convergence speed of the model during the training. We utilize a parallel transformer encoders and depthwise separable residual blocks to learn global semantic and local detail features. In addition, we are the first to implement expression transfer for 3D faces using expressive text. In contrast to DECA, which utilizes reference images for expression transfer, we employ the text-guided CLIP model to reconstruct 3D faces with specific expressions and maintain the consistency of face identities.

2 Method

2.1 Parameter Refinement Module

Given a 2D image I as input. First, we use a backbone to extract the coarse feature representation, which can be defined as $c = R(I)$. Then we employ parallel depthwise separable residual blocks and transformer encoders to learn local high-frequency and global semantic features. Local detail features and global semantic features are fused into rich feature representations. It can be defined as $v = cat(T(c') + R(c'))$. The depthwise separable residual block can effectively learn the local details. The transformer encoders can learn global semantic features from coarse feature representations. Enrich feature representation v using a classifier to obtain low-dimensional parametric codes. The parameter code consists of shape code $\alpha \in \mathbb{R}^{80}$, expression code $\beta \in \mathbb{R}^{64}$, texture code $t \in \mathbb{R}^{80}$, pose code $\rho \in \mathbb{R}^{6}$, and lighting code $l \in SO(2)$.

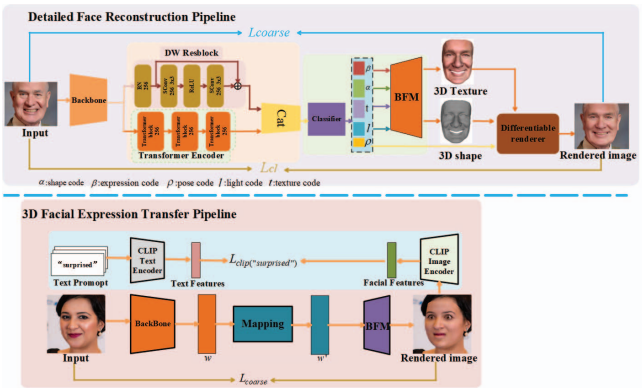

Figure 1: The framework of our method consists of a parameter refinement model, a Feature Consistency Loss, and 3D Facial Expression Transfer Module. The parameter refinement model refiners to obtain rich features. The feature consistency loss utilizes the CLIP model to recover detailed 3D face shapes. 3D Facial Expression Transfer module adopt the text-driven of CLIP model to 3D face expression transfer.

2.2 3D Facial Expression Transfer

We aim to learn an additional mapper to achieve expression transfer while training 3D faces. Given a source image I, a text st describing the expression is used as input. First, we use a pretrained backbone network to extract feature representation w from the source image. Then we employ a mapper to predict the feature representation of expression changes Δw. $\mathrm{w}' = \mathrm{w} + \Delta \mathrm{w}$. The mapper is composed of 3 transformer block, and the operation of attributes is realized by mapping the 3 transformer block. The expression described by the text uses the text encoder of the CLIP model to obtain a 512-dimensional expression feature vector. The rendered image uses the image encoder of the CLIP model to obtain 512-dimensional face features. Finally, the text manipulation loss is used to bias the expression of the facial features towards the expression feature. Specifically, we use the expression feature representing the textual expression to control the mapper to manipulate the expression of the source image.

2.3 Loss Function

2.3.1 Feature Consistency Loss

we present the feature consistency loss to reconstruct accurate 3D face shapes. The feature consistency loss includes both geometric

*e-mail: 202083060075@sdust.edu.cn
†e-mail: baozi0221@sdust.edu.cn, Corresponding Author
‡e-mail: qy@buaa.edu.cn

979-8-3503-4839-2/23/$31.00 ©2023 IEEE
DOI 10.1109/VRW58643.2023.00228

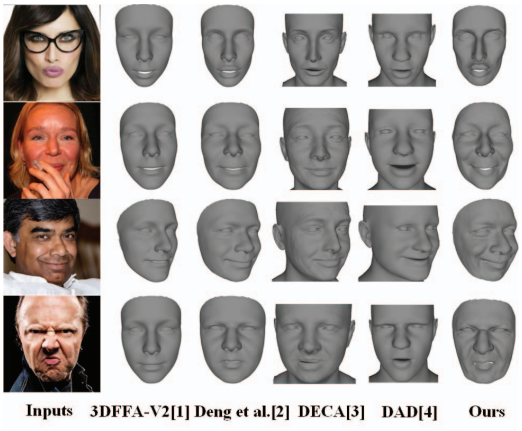

Figure 2: Comparison with recent methods. Our method exprecsses the sentiment of the input image better than other methods.

feature consistency and semantic feature consistency. We utilize CLIP's pretrained image encoder model to extract features from input and rendered images without needing further training. We adopt geometric feature consistency to measure the geometric similarity of the input image and the rendered image. We define the geometric feature consistency loss as:

$$L_{geometric} = \sum_{l=2}^{3} w_l \left\| CLIP_l(I) - CLIP_l(I_r) \right\|_2, \tag{1}$$

where $CLIP_l$ is layers 2 and 3 of the RN50x4 CLIP model. w_l is the weight of $CLIP_l$, $w_l = \{1, 1/2\}$. I, I_r are the input image and the rendered image, respectively. $\|\cdot\|_2$ is L_2 paradigm.

The semantic feature consistency computes the cosine similarity distance between the input and the rendered images' features. It can be defined as:

$$L_{semantic} = 1 - \cos(CLIP(I), CLIP(I_r)), \tag{2}$$

where $CLIP$ is the layer of the last layer of the ViT-B/32 CLIP model, I, I_r are the input image and the rendered image, respectively. $\cos(\cdot)$is the cosine distance.

Finally, the feature consistency loss can be defined as:

$$L_{cl} = L_{geometric} + L_{semantic}, \tag{3}$$

2.3.2 Text Manipulation Loss

To perform corresponding expression operations according to the text prompts of expressions, we use CLIP to design the following text manipulation loss:

$$L_{clip(st)} = 1 - \cos(E_i(I_r), E_t(st)), \tag{4}$$

where $\cos(.)$ means cosine similarity, E_i represents the image encoder of CLIP, E_t represents the text encoder of CLIP, the embedding of a given expression description text st.

Other evaluation indicators can be seen in [2]

2.4 Implementation Details

Our method is implemented in Pytorch and uses Adam to optimize the objective function. We train 200 batches on RTX 3060 GPU with batch size 16. The initial learning rate is set to 2×10^{-4}, and the learning rate decay is performed every 25 batches with a decay rate of 0.75. For the loss weight of the objective function, we set it to $\lambda_{photo} = 1, \lambda_{id} = 2, \lambda_{lm} = 1.7 \times 10^{-3}, \lambda_{reg} = 1 \times 10^{-4}, \lambda_{cl} = 2$.

Figure 3: Our method utilizes text-driven to generate rendered images of different expressions.

3 Results

To qualitatively evaluate the effectiveness of our method in reconstructing expression details, we compare our approach with recent 3D face reconstructions. As shown in Fig.2, The 3DDFA-V2 method cannot learn accurate expressions, resulting in some expression confusion. The results of Deng et al. and DAD methods cannot recover the facial expression details. Although the results of the DECA method can recover the facial expression details, the conveyed expressions are not realistic enough. The results generated by our method can recover the geometric details of the expressions, and the generated expressions are sufficiently realistic.

In Fig.3, we leverage text to drive expression transfer for 3D faces. It can be seen from the generated results that Our method realizes the transfer of expressions. We did not use feature consistency loss to learn details in the training process of expression transfer, which makes our results not outstanding in detail. Moreover, we use the textured skin tone of the model, which makes our results slightly different from natural images. However, these are not the most important because we aim to learn the transfer of expressions.

4 Conclusions

We present a method that learns the geometric details from many unconstrained face images and reconstructs 3D face models with slightly different expressions using the text. First, we propose a parameter refinement module to learn rich feature representations. Second, a novel feature consistency loss is designed, which utilizes the powerful representational CLIP model to capture facial expressions and geometric details. The feature consistency loss can effectively recover the local geometric details. Finally, we use the CLIP model's ability to prompt text to compare and learn the prompt text and the rendered image and reconstruct a 3D face model of the expression indicated by the text.

Acknowledgments

This work was supported by the Shandong Provincial Natural Science Foundation (ZR2020MF132); the National Natural Science Foundation of China (62072020).

References

[1] J. Guo, X. Zhu, Y. Yang, F. Yang, Z. Lei, and SZ Li. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *ECCV*, pp. 152-168, 2020.

[2] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to imageset. In *CVPR Workshops*, pp. 0-0, 2019. 2, 3, 4, 6

[3] Y. Feng, H. Feng, M. J. Black, and T. Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *TOG*, 40(4)1-13, 2021.6

[4] T. Martyniuk, O. Kupyn, Y. Kurlyak, I. Krashenyi. DAD-3DHeads: A Large-scale Dense, Accurate and Diverse Dataset for 3D Head Alignment from a Single Image. In *CVPR*, pp. 20942-20952, 2022.