Learning Animatable 3D Face Models from In-the-Wild Images

Authors: Jiang Cuilian

Date: 2024-01-06T00:00:00+00:00

Abstract

Although current single-image-based 3D face reconstruction methods can recover fine geometric details, these methods have limitations. Some methods generate faces that cannot be realistically animated because they do not model how wrinkles change with expression. Other methods are trained on high-quality facial scans and do not generalize well to in-the-wild images. The method employed in this work can regress details of 3D facial shape and animation that are person-specific yet vary with expression. The model of this method is trained to robustly generate UV displacement maps from a low-dimensional latent representation composed of person-specific detail parameters and generic expression parameters, while the regressor is trained to predict detail, shape, expression, pose, and lighting parameters from a single image. To achieve this, the method introduces a novel detail consistency loss that disentangles person-specific details from expression-dependent wrinkles. This disentanglement enables the synthesis of realistic person-specific wrinkles by controlling expression parameters while keeping person-specific details unchanged. This method is learned from in-the-wild images without paired 3D data supervision.

Full Text

Abstract

Although current single-image-based 3D face reconstruction methods can recover fine geometric details, they suffer from significant limitations. Some methods produce faces that cannot be realistically animated because they fail to model how wrinkles evolve with facial expressions. Others are trained on high-quality facial scans and do not generalize well to in-the-wild images. The method discussed in this report regresses 3D facial shape and animation details that are both person-specific and expression-dependent. The model is trained to robustly generate UV displacement maps from a low-dimensional latent representation

composed of person-specific detail parameters and generic expression parameters, while the regressor is trained to predict detail, shape, expression, pose, and illumination parameters from a single image. To achieve this, the method introduces a novel detail consistency loss that disentangles person-specific details from expression-dependent wrinkles. This disentanglement enables the synthesis of realistic person-specific wrinkles by controlling expression parameters while preserving person-specific details. The method learns from natural scene images without paired 3D data supervision.

Keywords: 3D face reconstruction, deep learning, facial details

Table of Contents

Chapter 1: Introduction

1.1 Research Background

The human face, as our most distinctive biological feature, conveys substantial information in daily life, including identity, emotion, and age—characteristics that are both intuitive and unique. Consequently, facial research holds considerable value and has become a prominent direction in computer vision. Unlike 2D facial images, 3D faces can display the human face from multiple angles in space, thereby enriching the representation of shape, pose, and texture information. Moreover, viewpoint transformation and occlusion do not affect facial representation in 3D space, significantly enhancing model robustness. However, because faces carry extensive information such as identity, gender, ethnicity, age, and emotion, fully and accurately representing this information in 3D space requires reconstructing high-quality 3D face models. How to reconstruct such high-quality 3D face models remains a challenging problem in 3D face reconstruction technology.

3D face reconstruction technology finds extensive application across multiple domains:

(1) Smart Healthcare: 3D face reconstruction technology is widely used in facial plastic surgery. Before surgery, doctors can use 3D face reconstruc-

tion software to model a patient's face, deepening their understanding of the patient's condition. Additionally, by performing cosmetic, makeup, and editing operations on the reconstructed 3D face model, post-surgical results can be demonstrated to patients, facilitating communication between doctors and patients.

(2) Film, Television, and Entertainment: As audiences demand higher-quality visual experiences and gamers seek more diverse interactive experiences, traditional 2D facial image processing technologies struggle to meet these growing needs. Consequently, 3D face reconstruction technology has gained widespread application in film, television, and gaming. Numerous films have presented 3D-reconstructed face models to audiences, such as the domestic animations *Monkey King: Hero is Back* and *Ne Zha*, which successfully combined traditional Chinese cultural figures with 3D face reconstruction technology to achieve both critical acclaim and commercial success. In gaming, 3D face reconstruction technology was adopted even earlier—China's first 3D online game, *Heavenly Walker*, launched in 2004 and received high praise from players. Since then, an increasing number of game companies have developed 3D games that deliver better user experiences and realism, incorporating 3D face reconstruction technology into their development processes. Today, some games not only provide facial customization services but even enable personalized facial modeling from player-provided photos, enhancing playability and user engagement.

(3) Face Recognition: Each person's facial features are unique; even among look-alikes, their facial characteristics are not identical. Therefore, faces can serve as a unique biometric identifier, much like fingerprints. Researchers have developed numerous face recognition algorithms accordingly. With the rise of deep learning methods, face recognition technology has achieved widespread application in finance, security screening, and attendance systems. However, most current face recognition technologies use 2D facial images, with only a handful employing 3D techniques. This leads to ineffective recognition under large poses and extreme lighting conditions, failing to meet practical application requirements. Some even use 2D facial images to spoof 3D face recognition devices. Unlike 2D images, 3D face models contain richer depth and texture information and are unaffected by viewing angle. Therefore, integrating 3D face reconstruction technology with face recognition will significantly improve recognition accuracy.

As one of the most popular research directions in computer vision, 3D face reconstruction technology has numerous practical applications that continuously improve human life experiences. Currently, industry still relies on structured-light cameras or 3D laser scanners to acquire facial shape and texture information. While the resulting 3D face models are highly accurate and realistic, the expensive equipment costs and complex, time-consuming processes create high barriers that severely limit practical application. In contrast, 2D facial images are much easier to obtain, as they can be captured effortlessly using mobile phones or cameras in daily life. Consequently, 3D face reconstruction from sin-

gle 2D images has become a focal research direction for scholars. With the rapid advancement of deep learning technology, the field of 3D face reconstruction has experienced swift development, holding significant research value and practical implications for future human life.

1.2 Research Status

3D face reconstruction technology aims to obtain accurate 3D facial data from 2D images and reconstruct 3D face models from this data, representing a hot topic among researchers and industry both domestically and internationally. Early researchers employed simple modeling methods and template deformation for reconstruction. Later, after the proposal of the 3DMM (3D Morphable Model) face model, 3D face reconstruction technology continuously evolved around this model. Presently, with the rapid development of deep learning, end-to-end 3D face reconstruction using deep learning methods has become the industry mainstream.

This report focuses on deep learning-based 3D face reconstruction from single facial images. In recent years, deep learning technology has propelled rapid development in computer vision-related fields. Traditional 3D face reconstruction methods, which fail to capture deep image features, suffer from limited expressive power and can no longer meet the growing demand for fine-grained 3D face models. Additionally, traditional methods involve numerous intermediate steps and require cumbersome operations to obtain reconstructed models. Deep learning models, however, can select appropriate loss functions for constraints based on learning tasks, enabling adaptive deep feature extraction from input images. This effectively compensates for the insufficient expressive power of traditional methods. Moreover, the end-to-end design of deep learning methods simplifies the intermediate modeling process, addressing the shortcomings and deficiencies of traditional approaches. Due to these advantages, deep learning-based 3D face reconstruction has become a new research hotspot in this field.

Deep learning-based 3D face reconstruction methods can be categorized into supervised and unsupervised approaches based on whether they require 3D face data. Supervised methods need corresponding 3D face data to optimize the reconstruction network. Some methods [2][3] introduce Convolutional Neural Networks (CNN) into the 3DMM model, using CNNs to directly predict 3DMM parameters. DOU [3] et al. proposed a method for predicting 3DMM model parameters at different layers of a CNN, combining CNNs with the 3DMM model and separately predicting expression and identity parameters at different network layers. Richardson [2] et al. proposed a progressively refined 3D face reconstruction algorithm consisting of two network modules: CoarseNet and FineNet. CoarseNet first recovers a coarse facial shape using the 3DMM method, which FineNet then continuously refines. Zhu et al. [1] proposed a 3D face reconstruction method based on cascaded CNNs, fitting the 3DMM model to input 2D facial images through cascaded networks for dense face reconstruction and alignment, demonstrating excellent reconstruction and alignment performance.

VRNet [5] treats the face as 200 cross-sections from the ear plane to the nose tip plane, using a CNN to directly regress each section for 3D face reconstruction. However, this method introduces issues with face alignment and limited reconstruction resolution scalability. PRNet [6] proposes a 2D representation based on UV position maps that can record 3D shape information of single facial images in UV space, reconstructing 3D facial shapes by training a CNN.

Although deep learning-based 3D face reconstruction methods have achieved rapid development, the lack of 3D face data and corresponding annotations severely limits model performance. To address this, researchers have proposed unsupervised 3D face reconstruction methods to alleviate the shortage of 3D data labels. Unsupervised methods render the generated 3D face model to 2D and optimize the reconstruction network by continuously fitting it to the input image. Pan [7] et al. utilized GAN networks to mine 3D geometric cues from 2D images, achieving unsupervised 3D shape recovery from single 2D facial images. Shang [8] et al. generated multi-view facial images from a single image and used consistency constraints across different views of the same face to enhance the expressive power of 3D face models. Thewlis [9] et al. used equivariance to learn dense landmarks and employed the learned landmark information to recover 3D geometry corresponding to 2D objects. Li [10] et al. exploited consistency between 3D meshes and depth images in an unsupervised manner to fit and optimize networks, enabling 3D face reconstruction under poor lighting conditions. DEA [11] further decomposes image albedo and shading and constrains an autoencoder with a small bottleneck embedding to predict 3D facial shape. Zhang [12] et al. proposed a novel learning aggregation and personalization framework that addresses the poor performance of unsupervised 3D face reconstruction under large poses and higher resolutions. Building upon these works, current popular methods have achieved significant breakthroughs. Wu [13] proposed an unsupervised method that disentangles 2D facial images into viewpoint, illumination, albedo, and depth factors through symmetry and combines them to obtain reconstructed 3D face models.

Chapter 2: Key Technologies

The theoretical and code foundation of this report originates from the paper: *Learning an Animatable Detailed 3D Face Model from In-The-Wild Images*. The report presents a two-stage face reconstruction approach. The first stage reconstructs a coarse facial shape, while the second stage recovers facial details.

[Figure 1: see original paper]

As shown in Figure 1, during the coarse reconstruction stage of the first phase, a facial image is input into an encoder (ResNet50 in the original paper) to obtain FLAME face model parameters including shape coefficients, expression coefficients, texture coefficients, pose, illumination, camera parameters, and reflectance. These parameters are then used with a differentiable renderer to

generate a 2D image. Finally, a loss is computed between the rendered 2D image and the original input image to update network parameters. By feeding the shape coefficients, expression coefficients, and pose coefficients from coarse reconstruction into the FLAME decoder, we obtain the 3D face vertices—the 3D face itself.

After coarse reconstruction training is completed, its network parameters are fixed, and detail reconstruction begins. The input for detail reconstruction is also a 2D image. An encoder (ResNet50 in the original paper) encodes the input image into a 128-dimensional detail code, which is then combined with expression parameters and jaw pose parameters from coarse reconstruction and fed into a decoder to generate a displacement map. Applying this displacement map to the coarse reconstruction result yields the final output.

The 3D face model used is FLAME, a statistical 3D face model with 5,023 vertices that can generate different 3D faces using shape coefficients, expression coefficients, and pose parameters.

2.1 Loss Functions

The total loss function is expressed as:

The facial landmark loss is expressed as:

This calculates the distance between 68 keypoints on the input facial image and the corresponding 68 keypoints projected onto the image plane from the generated FLAME face model.

The eye closure loss is expressed as:

Similar to the facial landmark loss, the eye closure loss computes the relative offset between eye keypoints.

The photometric loss:

where I is the original input image, I_r is the rendered image, and V_I is the mask corresponding to the input image, representing the visible facial region.

The identity consistency loss:

where $f(I)$ is the feature extracted from the original input facial image using a pre-trained face recognition network, and $f(I_r)$ is the feature extracted from the rendered image.

The shape consistency loss:

For different photos of the same person, the shape coefficients should remain constant. With shape coefficients fixed and other coefficients unchanged, all the above loss values are computed.

2.2 Loss Functions for Detail Reconstruction

The total loss function is expressed as:

The detail photometric loss function:

Similar to the photometric loss in coarse reconstruction, here I_r is the image after applying the displacement map to the rendered image.

The Implicit Diversified Markov Random Field (ID-MRF) loss:

This is computed on layers 3_2 and 4_2 of VGG19.

The soft symmetry loss:

To increase robustness to self-occlusion, a soft symmetry loss is added to regularize invisible facial parts. V_{uv} represents the facial skin mask in UV space, $flip$ denotes horizontal flipping operation, and D is the normal map obtained from the displacement map.

Chapter 3: Experimental Results

Based on the open-source code of the original paper, we retrained and evaluated the model using a collected face dataset.

Dataset Composition: VggFace2, CelebA, and AFLW, partitioned into training and validation sets at an 8:2 ratio. All other parameters follow the original paper settings, with batch size set to 4.

Reconstruction results on the NoW dataset test set are shown in Figure 2. The first row shows original input images; the second row shows predicted 68 2D facial landmarks; the third row shows predicted 68 3D facial landmarks, where green points are visible and red points are invisible; the fourth row shows reconstructed coarse facial shapes; the fifth row shows shapes with displacement maps applied; the sixth row shows shapes with texture maps applied. The results demonstrate good reconstruction quality, even for large poses and occluded cases.

[Figure 2: see original paper]

Quantitative evaluation was performed on the NoW dataset validation set (available at https://now.is.tue.mpg.de/). The evaluation metric computes the distance between predicted facial vertices and ground-truth vertices, calculating the mean, median, and standard deviation of all distances. Using the official code and validation set, the reconstruction errors are: median: 1.18mm, mean: 1.46mm, standard deviation: 1.25mm. The NoW dataset test set reconstruction error leaderboard is shown in Figure 3.

[Figure 3: see original paper]

Chapter 4: Conclusion

The method proposed in this paper learns an animatable detailed model from in-the-wild facial image datasets to reconstruct expressive and animatable face models with rich details from single images. The method is trained on in-the-wild datasets without 2D-to-3D supervision. The proposed detail reconstruction can disentangle person-specific details from expression-specific details, enabling the generation of animatable facial details applicable to animation, shape transformation, and wrinkle transfer. Due to its accuracy, reliability, and speed, the method can be applied to face reenactment and digital human creation.

The proposed face reconstruction method demonstrates good robustness to occlusion and large expressions but may fail under extreme poses. Additionally, the method can recover person-specific facial details from single images, with these details varying according to expressions. The method is fully open-source, providing significant guidance for 3D face reconstruction research.

References

[1] Zhu X, Yang F, Huang D, et al. Beyond 3dmm space: Towards fine-grained 3D face reconstruction[C]//Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part VII 16. Springer International Publishing, 2020: 343-358.

[2] Richardson E, Sela M, Or-El R, et al. Learning detailed face reconstruction from a single image[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1259-1268.

[3] Dou P, Shah S K, Kakadiaris I A. End-to-end 3D face reconstruction with deep neural networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5908-5917.

[4] Zhu X, Liu X, Lei Z, et al. Face alignment in full pose range: A 3D total solution[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 41(1): 78-92.

[5] Jackson A S, Bulat A, Argyriou V, et al. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression[C]//Proceedings of IEEE international conference on computer vision. 2017: 1031-1039.

[6] Feng Y, Wu F, Shao X, et al. Joint 3D face reconstruction and dense alignment with position map regression network[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 534-551.

[7] Pan X, Dai B, Liu Z, et al. Do 2D GANs know 3D shape? Unsupervised 3D shape reconstruction from 2D image GANs[J]. arXiv preprint arXiv:2011.00844, 2020.

[8] Shang J, Shen T, Li S, et al. Self-supervised monocular 3D face reconstruction with occlusion-aware multi-view geometry consistency[C]//Computer

Vision-ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XV. Cham: Springer International Publishing, 2020: 53-70.

[9] Thewlis J, Bilen H, Vedaldi A. Modelling and unsupervised learning of symmetric deformable object categories[J]. Advances in Neural Information Processing Systems, 2018, 31.

[10] Li P, Pei Y, Zhong Y, et al. An unsupervised approach for 3D face reconstruction from a single depth image[C]//Advances in Computer Graphics: 37th Computer Graphics International Conference, CGI 2020, Geneva, Switzerland, October 20-23, 2020, Proceedings 37. Springer International Publishing, 2020: 206-219.

[11] Shu Z, Sahasrabudhe M, Guler R A, et al. Deforming autoencoders: Unsupervised disentangling of shape and appearance[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 650-665.

[12] Zhang Z, Ge Y, Chen R, et al. Learning to aggregate and personalize 3D face from in-the-wild photo collection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14214-14224.

[13] Wu S, Rupprecht C, Vedaldi A. Unsupervised learning of probably symmetric deformable 3D objects from images in the wild[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 1-10.

Note: Figure translations are in progress. See original paper for figures.