Deep 3D Portrait from a Single Image

Sicheng Xu[1*] Jiaolong Yang[2] Dong Chen[2] Fang Wen[2] Yu Deng[3] Yunde Jia[1] Xin Tong[2]
[1]Beijing Institute of Technology [2]Microsoft Research Asia [3]Tsinghua University

Abstract

In this paper, we present a learning-based approach for recovering the 3D geometry of human head from a single portrait image. Our method is learned in an unsupervised manner without any ground-truth 3D data. We represent the head geometry with a parametric 3D face model together with a depth map for other head regions including hair and ear. A two-step geometry learning scheme is proposed to learn 3D head reconstruction from in-the-wild face images, where we first learn face shape on single images using self-reconstruction and then learn hair and ear geometry using pairs of images in a stereo-matching fashion. The second step is based on the output of the first to not only improve the accuracy but also ensure the consistency of overall head geometry. We evaluate the accuracy of our method both in 3D and with pose manipulation tasks on 2D images. We alter pose based on the recovered geometry and apply a refinement network trained with adversarial learning to ameliorate the reprojected images and translate them to the real image domain. Extensive evaluations and comparison with previous methods show that our new method can produce high-fidelity 3D head geometry and head pose manipulation results.

1. Introduction

Reconstructing 3D face geometry from 2D images has been a longstanding problem in computer vision. Obtaining full head geometry will enable more applications in games and virtual reality as it provides not only a new way of 3D content creation but also image-based 3D head rotation (*i.e.*, pose manipulation). Recently, single-image 3D face reconstruction has seen remarkable progress with the enormous growth of deep convolutional neutral networks (CNN) [47, 50, 22, 20, 16]. However, most existing techniques are limited to the facial region reconstruction without addressing other head regions such as hair and ear.

Face image synthesis has also achieved rapid progress with deep learning. However, few methods can deal with head pose manipulation from a singe image, which necessitates substantial image content regeneration in the head region and beyond. Promising results have been shown for face rotation [56, 3, 26] with generative adversarial nets (GAN), but generating the whole head region with new poses is still far from being solved. One reason could be implicitly learning the complex 3D geometry of a large variety of hair styles and interpret them onto 2D pixel grid is still prohibitively challenging for GANs.

In this paper, we investigate explicit 3D geometry recovery of portrait images for head regions including face, hair and ear. We model a 3D head with two components: a face mesh by the 3D Morphable Model (3DMM) [4], and a depth map for other head parts including hair, ear and other regions not covered by the 3DMM face mesh. The 3DMM face representation facilitates easy shape manipulation given its parametric nature, and depth map provides a convenient yet powerful representation to model the complex hair geometry.

Learning single-image 3D head geometry reconstruction is a challenging task. At least two challenges need to be addressed here. First, portrait images come with ground-truth 3D geometry are too scarce for CNN training, especially for hair which can be problematic for 3D scanning. To tackle this issue, we propose an unsupervised learning pipeline for head geometry estimation. For face part, we simply follow recent 3D face reconstruction methods [47, 20, 21, 59] to learn to regress 3DMM parameters on a corpus of images via minimizing the rendering-raw input discrepancy. But for hair and ear, we propose to exploit view change and train on pairs of portrait images extracted from videos via minimizing appearance reprojection error. The second challenge is how to ensure a consistent head structure since it consists of two independent components. We propose a two-step shape learning scheme where we use the recovered face geometry as conditional input of the depth network, and the designed loss function considers the layer consistency between face and hair geometry. We show that our two-step unsupervised shape learning scheme leads to compelling 3D head reconstruction results.

Our method can be applied for portrait image head pose manipulation, the quality of which will be contingent upon the 3D reconstruction accuracy thus could be used to evalu-

*This work was done when S. Xu was an intern at MSRA.

ate our method. Specifically, we change the pose of the reconstructed head in 3D and reproject it onto 2D image plane to obtain pose manipulation results. The reprojected images require further processing. Notably, the pose changes give rise to missing regions that need to be hallucinated. To this end, we train a refinement network using both real unpaired data and synthetic paired data generated via image corruption, together with a discriminator network imposing adversarial learning. Our task here appears similar to image inpainting. However, we found the popular output formation scheme (raw image merged with network-generated missing region) in deep generative image inpainting [57, 37] leads to inferior results with obvious artifacts. We instead opt for regenerating the whole image.

Our contributions can be summarized as follows:

- We propose a novel unsupervised head geometry learning pipeline without using any ground-truth 3D data. The proposed two-step learning scheme yields consistent face-hair geometry and compelling 3D head reconstruction results.
- We propose a novel single-image head pose manipulation method which seamlessly combines learned 3D head geometry and deep image synthesis. Our method is fully CNN-based, without need for any optimization or postprocessing.
- We systematically compare against different head geometry estimation and portrait manipulation approaches in the literature using 2D/3D warping and GANs, and demonstrate the superior performance of our method.

2. Related Work

Face and hair 3D reconstruction. 3D face reconstruction has been a longstanding task. Recently, deep 3D face reconstruction [47, 50, 22, 20] has attracted considerable attention. Our method follows the unsupervised learning schemes [47, 20] that train a network without ground-truth 3D data. For hair modeling, traditional methods perform orientation-map based optimization and sometimes require manual inputs [10] or a 3D hair exemplar repository [9]. Liang *et al*. [32] and Hu *et al*. [25] leverage hairstyle database for automatic hair reconstruction. A deep 3D hair reconstruction method was proposed in [60], but the reconstructed hair strand model are not aligned with the input image thus cannot be used for our purpose.

CNN-based portrait editing and synthesis. Face image editing and synthesis have attracted considerable attention in the vision and graphics community and have seen fast growth with the deep learning technique. Most existing CNN-based methods are devoted to editing appearance attributes such as skin color [12], facial expression [12, 42, 45, 41, 18], makeup [11, 31], age [58, 12, 52], and some other local appearance attributes [41, 14, 43]. Few methods worked on head pose manipulation. Perhaps the most relevant works among them are those synthesizing novel views (*e.g*., frontal) from an input face image [28, 3, 56]. However, the goals of these methods are not portrait editing and they do not handle hair and background.

2D warping based facial animation. Some approaches have been proposed to animate a face image with 2D warping [2, 18, 53]. Averbuch-Elor *et al*. [2] proposed to animate an image by the transferring the 2D facial deformations in a driving video using anchor points. A refinement process is applied to add fine-scale details and hallucinate missing regions. A similar pipeline is proposed by Geng *et al*. [18], which uses a GAN to refine the warped images. Wiles *et al*. [53] proposes to generate 2D warping fields using neural networks. Lacking guidance from 3D geometry, there is no guarantee the face structure can be persevered by these 2D warping methods especially when head pose changes.

3D-guided view synthesis and facial animation. 3D-guided face image frontalization and profiling have been used in different domains such as face recognition [46, 62, 23] and face alignment [61]. These methods often only focus on facial region or handle hair and background naively. The most sophisticated face rotation method is perhaps due to Zhu *et al*. [61], which considers the geometry of the surrounding regions of a face. However, their heuristically-determined region and depth do not suffice for realistic synthesis, and the rotated results oftentimes exhibit obvious inconsistency with the raw portraits. Moreover, the background is warped in [61] to avoid hole filling. Several works [36, 19] have been presented to synthesize facial expression leveraging 3D models, but they do not consider hair geometry and cannot manipulate head pose.

Video and RGBD based face reenactment. Several works have been presented for face reenactment with video or RGBD inputs [48, 30, 49]. Thies *et al*. [48] transfer facial expression in a source actor video to a target actor video with the aid of 3D face reconstruction. Kim *et al*. [30] train a deep network on each given video to fit the portrait appearances therein, such that high-quality generation can be obtained. An RGBD reenactment system is presented by Thies *et al*. [49].

3. Overview and Preprocessing

The frameworks of our methods are depicted in Fig. 1. After image preprocessing (to be described below), we run two-step 3D reconstruction with two CNNs to estimate 3D head pose and shape. For head pose manipulation, we first adjust the pose of the reconstructed shape in 3D and reproject it onto image plane, and then apply a refinement CNN to obtain the final result.

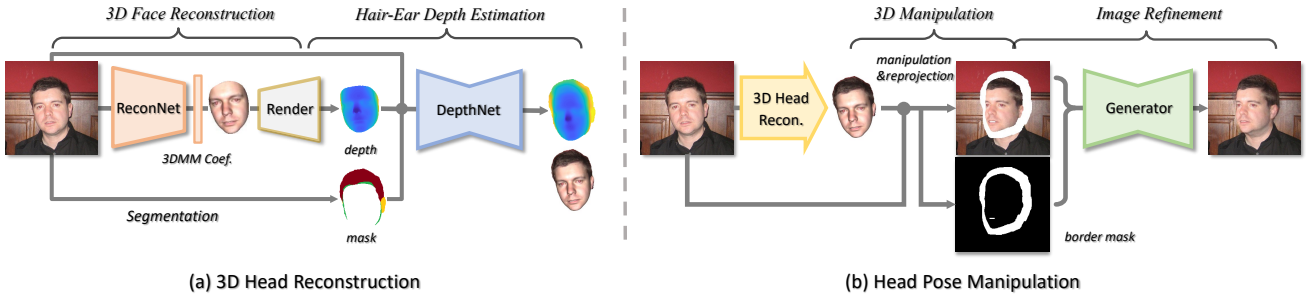


Figure 1: Overview of our single-image 3D head reconstruction and head pose manipulation methods.

Preprocessing. Given a portrait image, we perform rough alignment to centralize and rescale the detected face region (the image will be re-aligned later to accurately centralize the 3D head center after the 3D face reconstruction step). We then run a state-of-the-art face segmentation method of [33] to segment out the head region, denoted as $\mathcal{S}$, which includes face, hair and ear regions.

4. Single-Image 3D Head Reconstruction

In this work, we use the perspective camera model with an empirically-selected focal length. Head pose is determined by rotation $\mathbf{R} \in \mathrm{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$ and is parameterized by $\mathbf{p} \in \mathbb{R}^7$ with rotation represented by quaternion. We now present our method which reconstructs a 3DMM face as well as a depth map for other head regions.

4.1. Face Reconstruction and Pose Estimation

With a 3DMM, the face shape $\mathbf{F}$ and texture $\mathbf{T}$ can be represented by an affine model:

$$\begin{aligned} \mathbf{F} &= \mathbf{F}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \bar{\mathbf{F}} + \mathbf{B}_{id}\boldsymbol{\alpha} + \mathbf{B}_{exp}\boldsymbol{\beta} \\ \mathbf{T} &= \mathbf{T}(\boldsymbol{\delta}) = \bar{\mathbf{T}} + \mathbf{B}_{t}\boldsymbol{\delta} \end{aligned} \tag{1}$$

where $\bar{\mathbf{F}}$ and $\bar{\mathbf{T}}$ are the average face shape and texture; $\mathbf{B}_{id}$, $\mathbf{B}_{exp}$, and $\mathbf{B}_t$ are the PCA bases of identity, expression, and texture respectively; $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\delta}$ are the corresponding coefficient vectors. We adopt the Basel Face Model [40] for $\bar{\mathbf{F}}$, $\mathbf{B}_{id}$, $\bar{\mathbf{T}}$, and $\mathbf{B}_t$, and use the expression bases $\mathbf{B}_{exp}$ of [22] which are built from FaceWarehouse [7]. After selection of basis subsets, we have $\boldsymbol{\alpha} \in \mathbb{R}^{80}$, $\boldsymbol{\beta} \in \mathbb{R}^{64}$ and $\boldsymbol{\delta} \in \mathbb{R}^{80}$.

Since ground-truth 3D face data are scarce, we follow recent methods [47, 20, 16] to learn reconstruction in an unsupervised fashion using a large corpus of face images. Our method is adapted from [16] which uses hybrid-level supervision for training. Concretely, the unknowns to be predicted can be represented by a vector $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \mathbf{p}, \boldsymbol{\gamma}) \in \mathbb{R}^{239}$, where $\boldsymbol{\gamma} \in \mathbb{R}^9$ is the Spherical Harmonics coefficient vector for scene illumination. Let I be a training image and I' its reconstructed counterpart rendered with the network prediction, we minimize the photometric error via:

$$l_{photo} = \int_{\mathcal{F}} \|I - I'(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\delta}, \boldsymbol{\gamma}, \mathbf{p})\|_2 \tag{2}$$

where $\mathcal{F}$ denotes the rendered face region we consider here[1], and $\|\cdot\|_2$ denotes the ℓ_2 norm for residuals on r, g, b channels. We also minimize the perceptual discrepancy between the rendered and real faces via:

$$l_{per} = 1 - \frac{< f(I), f(I') >}{\|f(I)\| \cdot \|f(I')\|} \tag{3}$$

where $f(\cdot)$ denotes a face recognition network for identity feature extraction where the model from [55] is used here. Other commonly-used losses such as the 2D facial landmark loss and coefficient regularization loss are also applied, and we refer the readers to [16] for more details.

4.2. Hair&Ear Depth Estimation

Our next step is to estimate a depth map for other head region, defined as $\mathcal{H} = \mathcal{S} - (\mathcal{S}^f \bigcap \mathcal{F})$ where $\mathcal{S}^f$ denotes the face region defined by segmentation. $\mathcal{H}$ includes hair and ear as well as a small portion of segmented face region that is not covered by the projected 3DMM face. Due to lack of ground-truth depth data, we train a network using a collection of image pairs in a stereo matching setup. Note we use image pairs only for training purpose. The network always runs on a single image at test time.

Let I_1, I_2 be a training image pair of one subject (*e.g.*, two frames from a video) with different head poses $(\mathbf{R}_1, \mathbf{t}_1)$, $(\mathbf{R}_2, \mathbf{t}_2)$ recovered by our face reconstruction network. Our goal is to train a single network to predict both of their depth maps d_1 and d_2 in a siamese network scheme [13]. Before training, we first run naive triangulation on regular pixel grids of $\mathcal{H}_1$ and $\mathcal{H}_2$ to build two 2D meshes. Given depth map estimate d_1, a 3D mesh $\mathbf{H}_1$ can be constructed via inverse-projection. We can transform $\mathbf{H}_1$ to I_2's camera system via $(\mathbf{R}_2\mathbf{R}_1^{-1}, -\mathbf{R}_2\mathbf{R}_1^{-1}\mathbf{t}_1 + \mathbf{t}_2)$, and project it onto image plane to get a synthesized image I_2'. Similar process can be done for generating I_1' from I_2 and d_2. The whole process is differentiable and we use it to train our depth prediction network with the following losses.

As in stereo matching, we first enforce color constancy

[1] For brevity, in our loss functions we drop the notation of pixel variable in the area integral. We also drop the normalization factors (*e.g.*, $\frac{1}{N_{\mathcal{F}}}$ in Eq. 2 where $N_{\mathcal{F}}$ is the number of pixels in region $\mathcal{F}$).

constraint by minimizing the brightness error

$$l_{color} = \int_{\mathcal{H}'_2} \|I'_2(d_1) - I_2\|_1 + \int_{\mathcal{H}'_1} \|I'_1(d_2) - I_1\|_1 \quad (4)$$

where $\mathcal{H}'_2 = \mathcal{H}'_2(\mathcal{H}_1, d_1)$ is the warped region from $\mathcal{H}_1$ computed by head poses and d_1 in the transformation process described above; similarly for $\mathcal{H}'_1 = \mathcal{H}'_1(\mathcal{H}_2, d_2)$. We also apply a gradient discrepancy loss which is robust to illumination change thus widely adopted in stereo and optical flow estimation [6, 5, 54]:

$$l_{grad} = \int_{\mathcal{H}'_2} \|\nabla I'_2(d_1) - \nabla I_2\|_1 + \int_{\mathcal{H}'_1} \|\nabla I'_1(d_2) - \nabla I_1\|_1 \quad (5)$$

where ∇ denotes the gradient operator. To impose a spatial smoothness prior, we add a second-order smoothness loss

$$l_{smooth} = \int_{\mathcal{H}_1} |\Delta d_1| + \int_{\mathcal{H}_2} |\Delta d_2| \quad (6)$$

where Δ denotes the Laplace operator.

Face depth as condition and output. Instead of directly estimating hair and ear depth from the input image I, we project the reconstructed face shape $\mathbf{F}$ onto image plane to get a face depth map d^f. We make d^f an extra conditional input concatenated with I. Note d^f provides beneficial information (*e.g.*, head pose, camera distance) for hair and ear depth estimation. In addition, it allows the known face depth around the contour to be easily propagated to the adjacent regions with unknown depth.

More importantly, we train the network to also predict the depth of the facial region using d^f as target:

$$l_{face} = \int_{\mathcal{F}_1 - \mathcal{S}_1^h \cap \mathcal{F}_1} |d_1 - d_1^f| + \int_{\mathcal{F}_2 - \mathcal{S}_2^h \cap \mathcal{F}_2} |d_2 - d_2^f| \quad (7)$$

where $\mathcal{S}^h$ denotes the hair region defined by segmentation. Note learning face depth via l_{face} should not introduce much extra burden for the network since d^f is provided as input. But crucially, we can now easily enforce the consistency between the reconstructed 3D face and the estimated 3D geometry in other regions, as in this case we calculate the smoothness loss across whole head regions $\mathcal{S}_1, \mathcal{S}_2$:

$$l_{smooth} = \int_{\mathcal{S}_1} |\Delta d_1| + \int_{\mathcal{S}_2} |\Delta d_2| \quad (8)$$

Figure 2 (2nd and 3rd columns) compares the results with and without face depth. We also show quantitative comparisons in Table 1 (2nd and 3rd columns). As can be observed, using face depth significantly improves head geometry consistency and reconstruction accuracy.

Layer-order loss. Hair can often occlude a part of facial region, leading to two depth layers. To ensure correct relative position between the hair and occluded face region (*i.e.*, the former should be in front of the latter), we introduced a layer-order loss defined as:

$$l_{layer} = \int_{\mathcal{S}_1^h \cap \mathcal{F}_1} \max(0, d_1 - d_1^f) + \int_{\mathcal{S}_2^h \cap \mathcal{F}_2} \max(0, d_2 - d_2^f) \quad (9)$$

which penalizes incorrect layer order. As shown in Fig. 2, the reconstructed shapes are more accurate with l_{layer}.

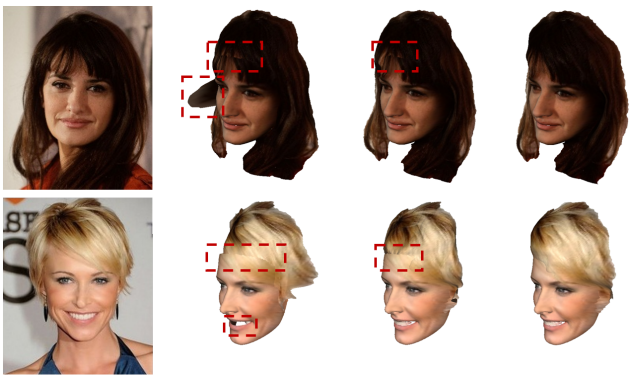


Figure 2: 3D head reconstruction result of our method with different settings.

Network structure. We apply a simple encoder-decoder structure using a ResNet-18 [24] as backbone. We discard its global average pooling and the last fc layers, and append several transposed convolutional layers to upsample the feature maps to the full resolution. Skip connections are added at 64×64, 32×32 and 16×16 resolutions. The input image size is 256×256. More details of the network structure can be found in the *suppl. material*.

5. Single-Image Head Pose Manipulation

Given the 3D head model reconstructed from the input portrait image, we modify its pose and synthesize new portrait images, described as follows.

5.1. 3D Pose Manipulation and Projection

To change the head pose, one simply needs to apply a rigid transformation in 3D for the 3DMM face $\mathbf{F}$ and hair-ear mesh $\mathbf{H}$ given the target pose $\bar{\mathbf{p}}$ or displacement $\delta\mathbf{p}$. After the pose is changed, we reproject the 3D model onto 2D image plane to get coarse synthesis results. Two examples are shown in Fig. 3.

5.2. Image Refinement with Adversarial Learning

The reprojected images suffer from several issues. Notably, due to pose and expression change, some holes may appear, where the missing background and/or head region should be hallucinated akin to an image inpainting process. Besides, the reprojection procedure may also induce certain artifacts due to imperfect rendering.

Input Mixing output l_{color} $l_{color}+l_{adv}$ Ours

Figure 3: Results of the image refinement network trained with different settings.

To address these issues, we apply a deep network G to process these images. For stronger supervision, we leverage both paired (*i.e.*, images with ground truth label) and unpaired data (*i.e.*, our coarse results) to train such a network. To obtain paired data, we take some real images with various head poses, and synthetically masked out some regions along the head segmentation boundaries. Let J be an unpaired coarse result, and $(R, \hat{R})$ be the paired data where R denotes the corrupted image and $\hat{R}$ its corresponding real image, we apply the ℓ_1 color loss via

$$l_{color}(G) = \mathbb{E}_J\Big[\int_{\mathcal{B}} \|G(J) - J\|_1\Big] + \mathbb{E}_R\Big[\int \|G(R) - \hat{R}\|_1\Big] \quad (10)$$

where $\mathcal{B}$ denotes the background and the warped head regions of J.

We apply adversarial learning to improve the realism of the generated images. We introduce a discriminator D to distinguish the outputs of G from real images, and train G to fool D. The LS-GAN [35] framework is used, and our adversarial loss functions for G and D can be written as

$$l_{adv}(G) = \frac{1}{2}\mathbb{E}_J\big[(D(G(J))-1)^2\big] + \frac{1}{2}\mathbb{E}_R\big[(D(G(R))-1)^2\big], \quad (11)$$

$$l_{adv}(D) = \frac{1}{2}\mathbb{E}_J\big[(D(G(J))-0)^2\big] + \frac{1}{2}\mathbb{E}_R\big[(D(G(R))-0)^2\big] + \mathbb{E}_R\big[(D(\hat{R})-1)^2\big], \quad (12)$$

respectively. As shown in Fig. 3, with the aid of the adversarial loss, our model generates much sharper results. However, some unwanted artifacts are introduced, possibly due to unstable GAN training.

To remove these artifacts, we further apply a deep feature loss, also known as perceptual loss [29], for paired data via

$$l_{feat}(G) = \sum_i \frac{1}{N_i}\|\phi_i(G(R)) - \phi_i(\hat{R})\|_1 \quad (13)$$

where ϕ_i is the i-th activation layer of the VGG-19 network [44] pretrained on ImageNet [15]. We use the first layers in all blocks. Figure 3 shows that our final results appear quite realistic. They are sharp and artifact-free.

Difference with image inpainting. In our task, the reprojected head portrait, though visually quite realistic to human observer, may contain some unique characteristics that can be detected by the discriminator. We tried generating refined results to be fed into D by mixing G's partial output and the raw input – a popular formulation in deep image inpainting [57] – via $J' = M \odot G(J) + (1-M) \odot J$ where M is the missing region mask. However, the results are consistently worse than our full image output strategy (see Fig. 3 for a comparison).

Table 1: Average 3D reconstruction error evaluated on RGBD images from the Biwi dataset [17].

Error (mm)	Zhu [61]	Ours$_{w/o\ df}$	Ours
Face	5.05	4.31	**3.88**
Non-face	8.56	7.39	**6.78**

Network structure. Our network structures for G and D are adapted from [51]. The input and output image size is 256×256. More details can be found in the *suppl. material*.

6. Experiments

Implementation Details. Our method is implemented with Tensorflow [1].[2] The face reconstruction network is trained with 180K in-the-wild face images from multiple sources such as CelebA [34], 300W-LP [61] and LFW [27]. To train the head depth network, we collected 11K image pairs from 316 videos of 316 subjects that contain human head movements[3]. The relative rotations are mostly within 5 to 15 degrees. The training took 15 hours on 1 NVIDIA M40 GPU card. To train the image refinement network, we collected 37K paired data and 30K unpaired data, and the training took 40 hours on 4 M40 GPU cards. Due to space limitation, more implementation details and results are shown in the *suppl. material*.

6.1. Results

The results from our method will be presented as follows. Note all the results here are from our test set where the images were not used for training.

3D head reconstruction. Figure 4 shows some typical samples of our single-image 3D head reconstruction results. As can be observed, our reconstruction networks can produce quality face and hair geometry given a single portrait image, despite we did not use any ground-truth 3D data for training. Various hair styles can be well handled as shown in the figure. Although the face region and the hair-ear part have different, disjoint representations (3DMM *vs.* depth map) and are reconstructed in two steps, they appear highly consistent with each other and the resultant head models are visually pleasing.

For quantitative evaluation and ablation study, we use the RGBD images from Biwi Kinect Head Pose Database [17]

[2] Code and trained model will be released publicly.

[3] We assume hair undergoes rigid motion within small time windows.

Figure 4: Typical single-image head reconstruction results. Our method can deal with a large variety of face shapes and hair styles, generating high-quality 3D head models. Note our method is trained without any ground-truth 3D data.

which contains 20 subjects with various hair styles. We compute the head reconstruction errors of our method using the depth images as ground-truth geometry. The error is computed as the average point distances in 3D between the outputs and ground-truth shapes after 3D alignment. The results are presented in Table 1, which shows the decent 3D reconstruction accuracy of our method. It also shows that the accuracy decreases if face depth is not used as the input for the depth estimation network, demonstrating the efficacy of our algorithm design.

Pose Manipulation. Figure 5 presents some pose manipulation results from our method. It can be seen that our method can generate realistic images with new head poses. Not only the face identity is well preserved, but also the hair shapes are highly consistent across different poses. The background is not disrupted by pose change.

6.2. Comparison with Prior Art

Comparison with Zhu *et al.* [61]. Zhu *et al.* [61] proposed a CNN-based 3D face reconstruction and alignment

Figure 5: Typical pose manipulation results. The left column shows the input images to our method, and the other columns show our synthesized images with altered head poses.

Figure 6: Comparison with the methods of Zhu *et al*. [61], Chai *et al*. [8], and Wiles *et al*. [53].

approach for single images. It also provides a warping-based portrait rotation method, originally developed for training data generation, based on 3D face geometry. To obtain reasonable warping for hair and ear regions, it defines a surrounding region of the face and heuristically determines its depth based on face depth. Figure 6 compares the face rotation results of [61] and ours. It can be seen that [61]'s results may suffer from obvious distortions. In contrast, our method can generate new views that are not only more realistic but also more consistent with the input images.

Table 2: Average perceptual similarity (deep feature cosine similarity) between the input and rotated face images.

	Chai *et al*. [8]	Ours
Cosine distance	0.829	**0.856**

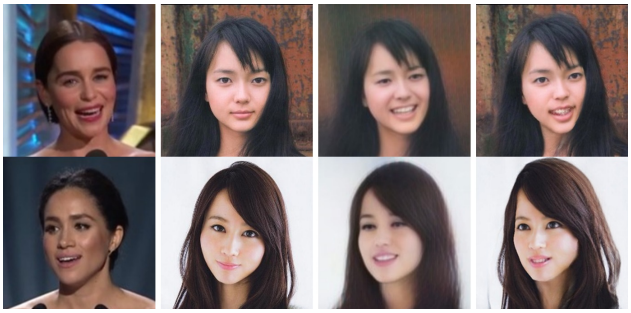
Also note that [61] simply warps the whole image including the background region. Background change is undesired for portrait manipulation but is commonly seen in previous 2D/3D warping based methods [8, 62, 61, 2]. In contrast, our method can well preserve the background.

Table 1 compares the 3D reconstruction error of [61] and our method using images from the Biwi dataset [17]. It shows that our method outperforms [61] by an appreciable margin: our errors are 23.17% and 20.79% lower in face and non-face regions, respectively.

Comparison with Chai *et al*. [8]. We then compare with Chai *et al*. [8], which is a traditional optimization-based method developed for hair modeling. It also estimates face

Input [2] Ours [49]

Figure 7: Comparison with Averbuch-Elor *et al*. [2]. The input image and result of [2] are from [49].



Source Target FSGAN [38] Ours

Figure 8: Comparison with the FSGAN method of Nirkin *et al*. [38]. Images are from [38].

shape by facial landmark fitting. We run the program released by [8], which requires a few user-provided strokes before reconstruction and provides 3D view of the input image after running reconstruction. As shown in Fig. 6, the method of [8] also leads to some distortions, while our results are more consistent with the input faces. The differences in background regions are also prominent.

For quantitative comparison, we consider a face recognition setup and compute the perceptual similarity between the original images and the warped ones. For fair comparison, we use the 10 images shown in [8] (Figure 13) and the corresponding results therein. For our method, we rotate the raw faces to same poses as theirs. We use VGG-Face [39] for deep face feature extraction, and Table 2 shows the higher perceptual similarity of our results.

Comparison with Averbuch-Elor *et al*. [2]. In Fig. 7, we qualitatively compare with a 2D-warping based face reenactment method of Averbuch-Elor *et al*. [2], which drives a face image with a reference face video for animation. As can be seen, pose change is problematic for 2D warping and the result exhibits obvious distortions. Ours contains much less distortion and appears more realistic. For reference, we also present in Fig. 7 the result of Thies *et al*. [49], which works on RGBD images and builds a target actor 3D model offline for high-quality reenactment.

Comparison with X2Face [53]. We also compare with another 2D warping-based face manipulation method, X2Face [53], where the warping fields are generated by neural networks. As shown in Fig. 6, the results of [53] suffer from obvious distortions and cannot handle missing regions, whereas ours appear much more natural.

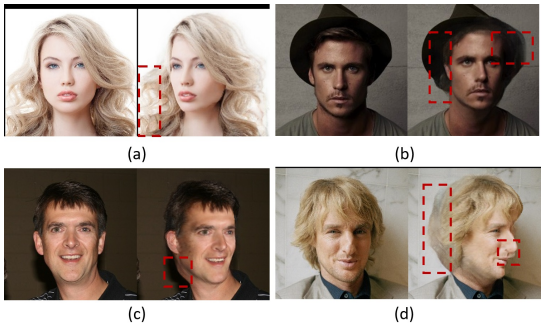


Figure 9: Failure cases due to (a) wrong segmentation, (b) obstruction, (c) inaccurate rotation center estimate, and (d) extreme pose.

Comparison with FSGAN [38]. Finally, we compare with a recent GAN-based face swapping and reenactment method, FSGAN [38]. As shown in Fig. 8, the results of [38] tend to be over-smoothed. We believe there's a still great hurdle for GANs to directly generate fine details that are geometrically correct, given the complex geometry of hairs. However, the expression of [38]'s results appears more vivid than ours especially for the first example. One of our future works would be integrating fine-grained expression manipulation into our pipeline.

Failure cases. Our method may fail under several situations, as illustrated in Fig. 9. For example, erroneous segmentation and obstructions may lead to apparent artifacts. Our head center is estimated in the face reconstruction step, and artifacts may appear for a few cases with inaccurate head center estimates. Our current method can not handle extreme poses, which we leave as our further work.

Running time. Tested on an NVIDIA M40 GPU, our face reconstruction, depth estimation and image refinement networks take $13ms$, $9.5ms$, and $11ms$ respectively to run on one single image. Segmentation takes $15ms$. Interactive portrait manipulation can be easily achieved by our method.

7. Conclusion

We presented a novel approach for single-image 3D portrait modeling, a challenging task that is not properly addressed by existing methods. A new CNN-based 3D head reconstruction pipeline is proposed to learn 3D head geometry effectively without any ground-truth 3D data. Extensive experiments and comparisons collectively demonstrated the efficacy of our proposed method for both 3D head reconstruction and single-image pose manipulation.

Acknowledgment. This work was partially supported by the National Natural Science Foundation of China under Grants No. 61773062.

References

[1] Martí Abadi et al. TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org/, 2015. 5

[2] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, 36(6):196, 2017. 2, 7, 8

[3] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6722, 2018. 1, 2

[4] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, pages 187–194, 1999. 1

[5] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *British Conference on Computer Vision (BMVC)*, volume 11, pages 1–11, 2011. 4

[6] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert. High accuracy optical flow estimation based on a theory for warping. In *European Conference on Computer Vision (ECCV)*, pages 25–36, 2004. 4

[7] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. FaceWarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014. 3

[8] Menglei Chai, Linjie Luo, Kalyan Sunkavalli, Nathan Carr, Sunil Hadap, and Kun Zhou. High-quality hair modeling from a single portrait photo. *ACM Transactions on Graphics (TOG)*, 34(6):204, 2015. 7, 8

[9] Menglei Chai, Tianjia Shao, Hongzhi Wu, Yanlin Weng, and Kun Zhou. Autohair: Fully automatic hair modeling from a single image. *ACM Transactions on Graphics (TOG)*, 35(4), 2016. 2

[10] Menglei Chai, Lvdi Wang, Yanlin Weng, Xiaogang Jin, and Kun Zhou. Dynamic hair manipulation in images and videos. *ACM Transactions on Graphics (TOG)*, 32(4):75, 2013. 2

[11] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. PairedCycleGAN: Asymmetric style transfer for applying and removing makeup. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 40–48, 2018. 2

[12] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018. 2

[13] Sumit Chopra, Raia Hadsell, Yann LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 539–546, 2005. 3

[14] Tali Dekel, Chuang Gan, Dilip Krishnan, Ce Liu, and William T Freeman. Sparse, smart contours to represent and edit images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3511–3520, 2018. 2

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009. 5

[16] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 1, 3

[17] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *International Journal of Computer Vision (IJCV)*, 2013. 5, 7

[18] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics (TOG)*, 37(6), 2018. 2

[19] Zhenglin Geng, Chen Cao, and Sergey Tulyakov. 3d guided fine-grained face manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9821–9830, 2019. 2

[20] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4362–4371, 2018. 1, 2, 3

[21] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 270–279, 2017. 1

[22] Yudong Guo, Juyong Zhang Zhang, Jianfei Cai, Boyi Jiang, and Jianmin Zheng. CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 1, 2, 3

[23] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4295–4304, 2015. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[25] Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (TOG)*, 36(6):195, 2017. 2

[26] Yibo Hu, Xiang Wu, Bing Yu, Ran He, and Zhenan Sun. Pose-guided photorealistic face rotation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8398–8406, 2018. 1

[27] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 5

[28] Rui Huang, Shu Zhang, Tianyu Li, and Ran He. Beyond face rotation: Global and local perception gan for photorealistic and identity preserving frontal view synthesis. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2439–2448, 2017. 2

[29] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 5

[30] Hyeongwoo Kim, Pablo Carrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018. 2

[31] Tingting Li, Ruihe Qian, Chao Dong, Si Liu, Qiong Yan, Wenwu Zhu, and Liang Lin. Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In *ACM International Conference on Multimedia*, pages 645–653, 2018. 2

[32] Shu Liang, Xiufeng Huang, Xianyu Meng, Kunyao Chen, Linda G Shapiro, and Ira Kemelmacher-Shlizerman. Video to fully automatic 3d hair model. In *SIGGRAPH Asia*, page 206, 2018. 2

[33] Jinpeng Lin, Hao Yang, Dong Chen, Ming Zeng, Fang Wen, and Lu Yuan. Face parsing with roi tanh-warping. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 5

[35] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2794–2802, 2017. 5

[36] Koki Nagano, Jaewoo Seo, Jun Xing, Lingyu Wei, Zimo Li, Shunsuke Saito, Aviral Agarwal, Jens Fursund, and Hao Li. paGAN: real-time avatars using dynamic textures. In *SIGGRAPH Asia*, page 258, 2018. 2

[37] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 2

[38] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7184–7193, 2019. 8

[39] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, et al. Deep face recognition. In *British Conference on Computer Vision (BMVC)*, page 6, 2015. 8

[40] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, 2009. 3

[41] Tiziano Portenier, Qiyang Hu, Attila Szabo, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. Faceshop: Deep sketch-based face image editing. *ACM Transactions on Graphics (TOG)*, 37(4):99, 2018. 2

[42] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. GANimation: Anatomically-aware facial animation from a single image. In *European Conference on Computer Vision (ECCV)*, pages 818–833, 2018. 2

[43] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5541–5550, 2017. 2

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 5

[45] Lingxiao Song, Zhihe Lu, Ran He, Zhenan Sun, and Tieniu Tan. Geometry guided adversarial facial expression synthesis. In *ACM International Conference on Multimedia*, pages 627–635, 2018. 2

[46] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1701–1708, 2014. 2

[47] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. MoFa: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 1274–1283, 2017. 1, 2, 3

[48] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2387–2395, 2016. 2

[49] Justus Thies, Michael Zollhöfer, Christian Theobalt, Marc Stamminger, and Matthias Nießner. Headon: real-time reenactment of human portrait videos. *ACM Transactions on Graphics (TOG)*, 37(4):164, 2018. 2, 8

[50] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1502, 2017. 1, 2

[51] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018. 5

[52] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face aging with identity-preserved conditional generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7939–7947, 2018. 2

[53] Olivia Wiles, A Sophia Koepke, and Andrew Zisserman. X2face: A network for controlling face generation using im-

ages, audio, and pose codes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 670–686, 2018. 2, 7, 8

[54] Jiaolong Yang and Hongdong Li. Dense, accurate optical flow estimation with piecewise parametric model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1019–1027, 2015. 4

[55] Jiaolong Yang, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li, and Gang Hua. Neural aggregation network for video face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4362–4371, 2017. 3

[56] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *International Conference on Computer Vision (ICCV)*, pages 3990–3999, 2017. 1, 2

[57] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5505–5514, 2018. 2, 5

[58] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5810–5818, 2017. 2

[59] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, 2017. 1

[60] Yi Zhou, Liwen Hu, Jun Xing, Weikai Chen, Han-Wei Kung, Xin Tong, and Hao Li. Hairnet: Single-view hair reconstruction using convolutional neural networks. In *European Conference on Computer Vision (ECCV)*, pages 235–251, 2018. 2

[61] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 146–155, 2016. 2, 5, 6, 7

[62] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 787–796, 2015. 2, 7