Eyeglasses 3D shape reconstruction from a single face image

Yating Wang[1], Quan Wang[2], and Feng Xu*[1]

[1] BNRist and school of software, Tsinghua University
[2] SenseTime Group Limited, Beijing, China

Abstract. A complete 3D face reconstruction requires to explicitly model the eyeglasses on the face, which is less investigated in the literature. In this paper, we present an automatic system that recovers the 3D shape of eyeglasses from a single face image with an arbitrary head pose. To achieve this goal, we first trains a neural network to jointly perform glasses landmark detection and segmentation, which carry the sparse and dense glasses shape information respectively for 3D glasses pose estimation and shape recovery. To solve the ambiguity in 2D to 3D reconstruction, our system fully explores the prior knowledge including the relative motion constraint between face and glasses and the planar and symmetric shape prior feature of glasses. From the qualitative and quantitative experiments, we see that our system reconstructs promising 3D shapes of eyeglasses for various poses.

1 Introduction

Eyeglasses exist in many facial images. They can somehow be considered as extending components for human face, which influence face appearance dramatically. Reconstructing glasses explicitly is beneficial for many applications. For example, reconstructing 3D face as well as glasses on the face obviously achieves a more complete face modeling. With known glasses shape and pose, the interference caused by glasses occlusion can be eliminated in many face-related tasks such as face shape/appearance reconstruction and face authentication. Moreover, applications related to glasses can also be realized based on the glasses reconstruction, like glasses design, removal, and virtual try-on.

Reconstructing 3D glasses from a single face image is challenging, which suffers the following difficulties. First, features for reconstructing glasses are less investigated in the literature, neither the handcraft features nor the learning-based features. Second, glasses in images may vary a lot due to the large head pose changes, which increases the ambiguity in reconstructing 3D glasses from a single 2D image.

Some previous techniques have worked on this topic and tried to overcome some of the aforementioned difficulties. [13] distinguishes the face and glasses depth estimated from multi-view RGB images and reconstructs coarse 3D glasses. They propose a generic model representing the outer contour of glasses based on the glasses geometry commonality, and they optimize the contour by the

Fig. 1. Our system reconstructs eyeglasses from single input face image with an arbitrary head pose. More results can be found in the result section and supplementary materials.

symmetry shape prior feature of glasses. [28] realizes glasses segmentation from frontal face image also by the symmetry constraints, and reconstructs glasses frame by deforming the prior shape. However, these eyeglasses reconstruction techniques require either multi-view images or frontal face images. To the best of our knowledge, no previous work could reconstruct 3D glasses from a single face image with various head poses.

This paper proposes the first fully automatic system to recover glasses 3D shape from a single face image with an arbitrary head pose. Fig. 1 shows some results of our system. To guide the reconstruction of glasses, we extract two kinds of glasses features from images, i.e. the glasses landmarks and segmentation mask. We define glasses landmarks which represent the overall sparse shape of glasses as well as its pose in 3D space, which are never defined before. While the segmentation mask gives dense information describing the shape details of glasses, we observe that these two kinds of features are highly correlated and thus we propose a joint learning framework which trains one single network to perform the two tasks together.

To solve the large ambiguity in 2D to 3D estimation of glasses reconstruction, we involve various prior knowledge in our method. We leverage the well-studied face reconstruction techniques to construct motion direction constraints and contact constraints to solve the ambiguity in glasses pose estimation. Observing the planar shape of glasses (arms excluded), we frontalize the glasses so that the task of 3D shape retrieval and 3D shape deformation could be performed by 2D cues. The left-right symmetrical prior is also involved to further constrain the reconstructed 3D shape. With a technique fully exploring these priors, we successfully achieve 3D glasses reconstruction from a single face image.

2 Related works

2.1 3D face reconstruction

Faces occupy a central place in conveying human identity, expression and emotion. As a consequence, face 3D reconstruction is required in a wide range of applications. Multi-view registration [1] and shape from shading [9, 19, 8, 18] are the most common ways to achieve face reconstruction. Recently, deep learning is also applied in this task and achieves promising results [29, 6, 22, 24, 5, 16, 7]. As glasses influence human face appearance significantly, simultaneously reconstructing face and glasses will achieve better completeness, which is not fully

investigated yet. Besides, glasses cause the most common occlusions on the face, distracting face reconstruction frequently. Some methods are proposed to solve the occlusion of glasses [13] or other objects [3, 4, 23] in face reconstruction. In this paper, we explicitly reconstruct detailed glasses shape as well as the face shape, which will improve the realism and quality of face reconstruction.

2.2 Glasses reconstruction

Few works focus on glasses 3D reconstruction. [28] presents a method to reconstruct 3D glasses shape from a single frontal face image by extracting glasses frame contour and deforming existing glasses 3D template. Then the authors use the reconstructed glasses 3D model to achieve virtual glasses try-on. In [13], an approach operating on multi-view RGB images was proposed to automatically reconstruct face by ignoring the segmented depth of glasses and then use the segmented glasses depth to reconstruct glasses. But to the best of our knowledge, no previous works focus on recovering 3D glasses from a single face image of an arbitrary head pose.

2.3 Glasses manipulation

Most works related to eyeglasses focus on glasses detection, removal and virtual try-on. As glasses cover large portions of the face, many human face applications are visibly affected by glasses. Consequently, glasses removal is of much concern in the literature. In [25], a method was proposed to automatically locate eyeglasses and fill the glasses region to synthesize a face image without glasses. [15] proposes an algorithm for glasses removal by recursive error compensation using PCA reconstruction. Notice that both these two methods operate on frontal face images. Besides, some works exploit glasses virtual try-on, by which users choose desired glasses from images or glasses database, and the chosen glasses will be blended onto the users' photos [28, 12, 14, 21, 27]. We believe that by reconstructing glasses from limited inputs, applications related to glasses could perform better. So it is interesting and also our possible future work to investigate how to utilize the 3D glasses reconstruction techniques to perform the tasks discussed in this subsection.

3 Overview

The whole pipeline is shown in Fig. 2. Our system takes a face image with glasses as input. Firstly, to guide the reconstruction, we extract image features including detecting the face and glasses landmarks and segmenting out pixels representing glasses. Then we recover the 3D face and estimate the head pose using the face landmarks. To reconstruct the 3D shape of glasses, we iterate the following three steps until convergence.

1. Using the glasses landmarks and the current 3D glasses (initialized by a default template), we estimate the glasses pose and frontalize the glasses features (i.e. the glasses landmarks and glasses mask).

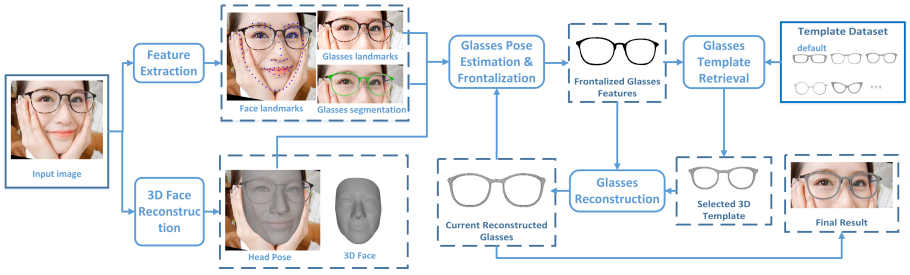

Fig. 2. Pipeline of the proposed system.

2. We select the best glasses template from a small dataset by the frontalized glasses mask.
3. After building the correspondences between the frontalized mask contour and the contour vertices of the chosen template, we deform the template to fit the shape of the input glasses.

The rest of this paper is organized as follows. Sec. 4 reviews our feature extraction including landmark detection and glasses segmentation. Sec. 5 introduces glasses pose estimation and glasses feature frontalization. Then Sec. 6 illustrates our glasses retrieval method and Sec. 7 introduces correspondences searching and glasses deformation method. Finally, Sec. 8 demonstrates the experiments to evaluate our technique.

4 Feature Extraction

For the following face and glasses reconstruction steps, we extract three types of features which are face landmarks, glasses landmarks, and glasses segmentation mask. For face landmark detection, as this is a well-investigated task, we directly use the method proposed by [26] which detects 98 face landmarks for each face in images. As there are no previous works which define and detect landmarks for glasses, we propose our technique to handle this based on our goal of glasses reconstruction. The definitions of the 21 glasses landmarks are shown in the left of Fig. 3. The glasses frame can be expressed by one outer closed curve and two inner closed curves. To reduce the semantic ambiguity of landmarks, we define the landmarks on these curves. Meanwhile, glasses segmentation in our paper is defined to segment the glasses frame (except the two arms) from images.

We use U-Net proposed in [17] to simultaneously predict the glasses landmarks and the segmentation mask. The face area, cropped by face landmarks and resized to 256*256 resolution, serves as the input of the network. The network outputs 21+1 256*256 maps. The first 21 are the heatmaps for the 21 landmarks and the last one is the segmentation feature map. As there is no available dataset for glasses segmentation or landmark prediction, we first collect 5300 face images half from the Internet and half recorded by ourselves. The internet images cover

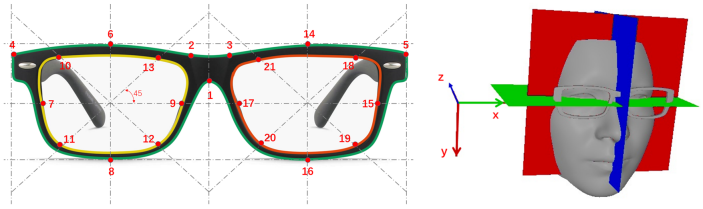

Fig. 3. The definition of glasses landmarks (left) and the face coordinate system (right).

various glasses styles while our recorded data contains large pose differences. After excluding the images with rimless or half rim or incomplete glasses which are unable to be labeled, we get 3300 images to construct our dataset. Finally, we manually label landmarks and segmentation masks for images in the dataset. We would like to release our dataset for future research. Notice that the ground truth heatmaps used to supervise landmark prediction are established by applying 2D Gaussian filtering at the labeled landmarks. And the ground truth probability maps are just the labeled binary segmentation masks.

To train the network, we use the weighted sum of the standard MSE loss for landmark heatmaps and the cross-entropy loss for the glasses segmentation probability map:

$$loss = \lambda_{balance} * loss_{landmark} + loss_{segment} \tag{1}$$

$$loss_{landmark} = \sum_{i=1}^{n=21} (x_i - y_i)^2 \tag{2}$$

$$loss_{segment} = -[y * log\sigma(x) + (1 - y) * log(1 - \sigma(x))] \tag{3}$$

where x_i and y_i refer to the output and the ground truth heatmaps of landmark i, x and y refers to the output segmentation feature map and the ground truth segmentation mask respectively, and $\sigma(x)$ refers to the output segmentation probability map where $\sigma(.)$ refers to the sigmoid function.

In the testing, we extract pixels of the highest value in the landmark heatmaps as landmarks and pixels whose probabilities are larger than 0.5 in the segmentation probability map as the glasses segmentation mask.

5 Glasses pose estimation and frontalization

This part introduces how to estimate the glasses pose and frontalize all the aforementioned glasses features for the following glasses shape reconstruction. Besides the features, this step also requires a 3D mesh model of the glasses. The mesh model is initialized by a template and is updated according to the image information in an iterative manner as described in Sec. 3.

Our key idea here is to combine the face pose estimation with the glasses pose estimation. There are two major reasons for this. First, combining faces with

glasses can give a more complete reconstruction of face region, which is usually not considered by previous face reconstruction techniques. Second, as the face and glasses have a strong relationship in position and rotation, the well-studied face reconstruction techniques can be used to benefit glasses reconstruction, especially in determining the glasses poses.

5.1 Face reconstruction

In practice, we first solve the image-based face reconstruction problem following the method in [2]. This method takes a parametric face model, predefined 3D landmarks on the model and the 2D facial landmarks in the image space as the input. To be more specific, we assume a zero-skew perspective camera with square pixels and the principal point at the image center. Then the 3D-to-2D projection can be formulated as:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{bmatrix} * [\mathbf{R}|\mathbf{t}] * \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{4}$$

where (x, y, z) refers to a 3D vertex in the face coordinate system as shown in the right of Fig. 3, (u, v) refers to its 2D projection, and (u_0, v_0) refers to the image coordinate of the image center. $\mathbf{R} \in \mathbb{R}^{3\times3}$ and $\mathbf{t} \in \mathbb{R}^3$ define the coordinate transformation of a point from the face coordinate system to the camera coordinate system. f refers to the focal length. Notice that here $=$ means equal for two homogeneous coordinates which may have a scale difference. Then we apply the 3D-to-2D projection to the face landmarks and thus estimate the parameters θ_{face}, which includes camera parameter f, the head pose and the shape and expression parameters of the parametric face model, by minimizing the L2 distance between the projected 3D face landmarks and the detected 2D landmarks:

$$\arg\min_{\theta_{face}} \sum_{i=0}^{m-1} \left\| \mathbf{p}_i^p - \mathbf{p}_i^l \right\|_2^2 \tag{5}$$

Here m indicates the number of the face landmarks. $\mathbf{p}_i^p$ and $\mathbf{p}_i^l$ indicate the projected and the detected 2D position of the landmark i. In our experiments, we guess some values for focal length f, based on which we use ePnP algorithm [11] to calculate the closed-form solution of the head pose, and then we choose the best one (with minimum error) to be the initial value. Then we iteratively estimate all parameters in θ_{face}. Notice that as the used face model is trained with real human face data and thus already gets the scale information, we do not need to consider the face scale in the optimization. More details including some other regularization terms can be found in [2].

5.2 Glasses pose estimation

Then we solve for pose of the glasses. As we have also predefined the glasses landmarks on the glasses mesh model and detected the 2D glasses landmarks in

the image, we could solve the glasses pose using similar optimization as Eqn. 5. However, as glasses of similar shapes may vary in size, the scale of glasses needs to be solved, which is impossible for pure glasses reconstruction as the 3D-to-2D projection has an inherent scale ambiguity. Furthermore, the planarity of glasses (the arms are excluded) also aggravates the instability of pose estimation. As a consequence, we use the solved head pose to constrain the glasses pose estimation.

We first manually pose the template glasses on the template face. Then we could represent the glasses in the face coordinate system and the global motion of the glasses could be expressed as:

$$[\mathbf{R}|\mathbf{t}] = \left[\mathbf{R}^f|\mathbf{t}^f\right] * [\mathbf{R}^g|\mathbf{t}^g] * s^g \tag{6}$$

where $[\mathbf{R}^g|\mathbf{t}^g]$ is the relative motion between the glasses and the face and s^g is the scale factor of the glasses. In most cases, the initial pose of the glasses on the face is almost correct and thus $\mathbf{R}^g$ should be close to $\mathbf{I}$ and $\mathbf{t}^g$ should be $\mathbf{0}$. Thus given the new projection formulation of the points on the glasses, we have the new energy to be minimized:

$$\arg\min_{\mathbf{R}^g,\mathbf{t}^g,s^g} \sum_{j=0}^{n-1} \left\|\mathbf{p}_j^p - \mathbf{p}_j^l\right\|_2^2 + \lambda \left\|\mathbf{R}^g - \mathbf{I}\right\|_2^2 + \lambda \left\|\mathbf{t}^g - \mathbf{0}\right\|_2^2 \tag{7}$$

where n is the number of glasses landmarks, λ controls the weights of different terms. The parameters like f, $\mathbf{R}^f$, and $\mathbf{t}^f$ have already been estimated in the face reconstruction step.

However, glasses may not always be in the pose as shown in the right of Fig. 3. Sometimes, glasses could be on the forehead or on the nose tip as shown in Fig. 1. To handle these cases, we does not directly constrain $\mathbf{R}^g$ and $\mathbf{t}^g$, but transfer them into 7 motion parameters $\theta_{glasses} = \{r_x^g, r_y^g, r_z^g, t_x^g, t_y^g, t_z^g, s^g\}$ after adding s^g and constrain $\theta_{glasses}^{sub} = \{r_y^g, r_z^g, t_x^g\}$ to be $\mathbf{0}$. This constraint is based on the observation that even for the uncommon cases in Fig. 1, the glasses will still not have the rotation around the y and z-axis or the translation on the x-axis.

However, even with the constraint on $\theta_{glasses}^{sub}$, the scale ambiguity in the 3D-to-2D projection still exists. To further solve this ambiguity, we propose a physical constraint that the two nose pads should be constrained on the face. So the final optimization for glasses pose estimation is:

$$\arg\min_{\theta_{glasses}} \sum_{j=0}^{n-1} \left\|\mathbf{p}_j^p - \mathbf{p}_j^l\right\|_2^2 + \lambda \left\|\theta_{glasses}^{sub} - \mathbf{0}\right\|_2^2 + \gamma \sum_{k=0}^{1} \left\|\mathbf{P}_k^g - \mathbf{P}_k^f\right\|_2^2 \tag{8}$$

where $\mathbf{P}_k^g$ denotes a manually defined 3D point representing one nose pad and $\mathbf{P}_k^f$ is its contacting point on the face. In practice, Eqn. 8 is optimized in an iterative manner and for each iteration, $\mathbf{P}_k^f$ is the closest point of $\mathbf{P}_k^g$ on the face. λ and γ are chosen to be very large to make the constraints firmly satisfied. Notice that

as our template glasses models may not have nose pads, we manually label two virtual points as the contact points on nose pads, which have fixed orientations to the geometry center of the glasses.

5.3 Frontalization

After obtaining θ_{face}, $\theta_{glasses}$ and the face and glasses mesh models, we get the 3D reconstruction of the face and glasses, which will be the final outputs when they are obtained by the last iteration. For the earlier iterations, we need to frontalize the glasses features (only the segmentation mask) by the following steps. For a pixel (u, v) on the glasses, we calculate its corresponding (x, y, z) in the camera coordinate system by θ_{face} and $\theta_{glasses}$. Notice that in the early iterations, the shape of the reconstructed glasses is not accurate that a pixel may not be able to back-projected onto the 3D glasses model. Since the glasses frame is almost on a plane, we fit a plane to get the 3D positions of the glasses pixels. The projected 3D points form our proxy glasses M_{pry}. Then by setting a proper θ^{frt}_{face} and $\theta^{frt}_{glasses}$, we can make the proxy glasses face the camera center along the camera's z-axis (denoted as M^{frt}_{pry}) and get the frontalized 2D glasses m^{frt}_{pry} by image projection.

6 Glasses template retrieval

In this section, we will use the frontalized glasses mask m^{frt}_{pry} (Fig. 4(a)) to find the best glasses mesh model in our glasses dataset. Actually, our dataset only contains 9 glasses mesh models with large shape differences. We do not require too many glasses models because we also have a shape deformation step that can deform a glasses mesh to the specific glasses shape in the input image. In practice, we find that these 9 models are almost enough to handle most daily glasses. If there is a new pair of glasses with a very unique shape, we just need to ask an artist to make one model with a similar shape and add it to our dataset.

The reason that we frontalize the glasses features is that most glasses models are plane-like. Notice that we do not consider glasses arms in this work. In this situation, the frontalized 2D shape contains the major information of the glasses and some 3D tasks could be simplified to 2D. Here, the 3D shape retrieval is performed in the 2D space.

To be specific, the frontalized glasses mask m^{frt}_{pry} is first normalized to be an N*N square image (Fig. 4(b)). At the same time, the 9 glasses in the dataset are also transformed and projected to be frontalized (Fig. 4:Left) and then normalized similarly. After extracting the segmentation masks of the 9 glasses in the normalized images, we calculate IOU between the query mask and each of the 9 masks in dataset to represent the similarity. Notice that different parts of the glasses have different difficulties to deform to another shape. Based on this observation, different parts should have different weights in calculating similarity. So before IOU calculation, we manually assign higher weights on parts that are hard to deform for every candidate model, like glasses bridge or hinges. In

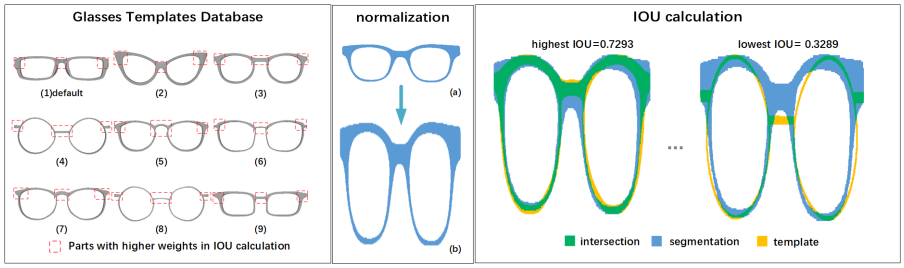

Fig. 4. An illustration for shape retrieval. From left to right: masks of all glasses in the dataset; a frontalized input glasses mask before and after the normalization; IOU calculation. In IOU calculation, we show the highest and lowest IOU between the input and the templates (3) and (4) in the dataset, respectively. We got the highest IOU between the input and (3), so (3) is the "best" template for the input. It is clear that (3) is much closer to the input than (4) and is much easier to be deformed to the shape of the input.

this manner, these regions will contribute more to the similarity. Notice that the weight map for each template will be normalized to make sure the retrieval is fair for all the templates. Finally, the glasses with the highest IOU are chosen from the dataset for the next step which will deform this glasses to fit the input image.

The retrieval is performed in all the iterations of the glasses reconstruction method as shown on the right side of Fig. 2. Except for the early iterations, we add the current deformed glasses model as the 10th model for retrieval. Not surprisingly, in most cases, the 10th model will be chosen as the retrieval result as it becomes more and more similar to the real shape of the input. However, for some input, if the glasses shape is too different from the initial template, the pose estimation may be wrong in the beginning, leading to a wrong retrieval result in the first iteration. In the following iterations, as the pose becomes better, the wrong retrieval could be corrected in this retrieval step and the system could recover from the errors. This is the reason we perform retrieval in every iteration.

7 Glasses reconstruction

In this section, we will deform the retrieved glasses mesh model to fit the glasses in the input image. Remember that the glasses features and the glasses mesh have both been frontalized by the method in Sec. 5. We again use the frontalized 2D information to guide the deformation. We first extract contours from the segmentation mask m_{pry}^{frt} using OpenCV. And then build dense correspondences between the contour vertices on the model and pixels of the contours. Next, a Laplacian-based deformation is performed on the model guided by the correspondences. With the last two steps iterating, a better and better glasses shape is obtained.

7.1 Correspondence search

In the previous section, we use the normalized images to calculate similarity. Here, we again use the normalized images to find correspondences between the predefined contour vertexes on the retrieved mesh model and the contours of m_{pry}^{frt}. In the normalized images, we find that the closet points are good approximation of the correspondences as the large scale shape differences between glasses are dramatically compensated in the normalized images. With the iterations in the glasses deformation, the glasses shape becomes better and better, and thus the closest points also become better and better in finding the real correspondences. So we use this ICP method on the normalized images to find correspondences. Notice that the errors caused by inner and outer contour mismatching are very likely to happen. To eliminate this, for each point, we calculate the direction from this point to the average position of its 2D neighbors in a small circle, and then use a threshold to filter out the candidates whose direction is quite different, as the inner and outer contour points have very different directions.

7.2 Glasses deformation

This part elaborates on how to deform the retrieved glasses mesh driven by the correspondences. We follow the Laplacian deformation approach presented in [20], using the correspondences searched in the previous step as constraints and enforcing the deformed model to keep horizontally symmetrical. In Sec. 5, the glasses mesh is placed to be symmetrical about plane $x = 0$. We define V and $V^{'}$ denoting the sets of vertices of the original and deformed 3D mesh, and $\mathbf{v} = (x, y, z)$ and $\mathbf{v}^{'} = (x^{'}, y^{'}, z^{'})$ denoting a particular vertex, respectively. We denote the set of correspondences by C, and $\mathbf{v}_j$ in C has a corresponding 3D point $\mathbf{p}_j = (u_j, v_j, w_j)$ in M_{pry}^{frt}. Notice that we have fitted a plane to the glasses model to perform the frontalization, which may bring some errors in the value of w_j. So we keep z-coordinates of $\mathbf{v}_j$ unchanged through the deformation, and $\mathbf{p}_j$ in C is replaced by $\mathbf{c}_j = (u_j, v_j, z_j)$.

Then the deformation can be formulated as the optimization of the following object function:

$$E(V^{'}) = \lambda_C E_C(V^{'}) + \lambda_L E_L(V, V^{'}) + \lambda_S E_S(V^{'}) \tag{9}$$

where E_C calculates the euclidean distance between the corresponding points pair:

$$E_C = \sum_{j \in C} \left\| \mathbf{v}_j^{'} - \mathbf{c}_j \right\|_2^2. \tag{10}$$

E_L strives to preserve the Laplacian coordinates, resulting in detail-preserving and smooth deformation:

$$E_L = \left\| L(V) - L(V^{'}) \right\|_2^2, \tag{11}$$

Fig. 5. Feature extraction of our joint training network. **Top**:landmark prediction. **Bottom**:segmentation.

where $L(.)$ is the transformation from Cartesian coordinates to Laplacian coordinates. E_S enforces that the deformed model to be horizontally symmetrical:

$$E_S = \sum_{i=1}^{N} |x_i^{'} + x_{k(i)}^{'}|^2 + |y_i^{'} - y_{k(i)}^{'}|^2 + |z_i^{'} - z_{k(i)}^{'}|^2. \tag{12}$$

Here, before the deformation, for each vertex v_i of each glasses model in dataset, we find a vertex $v_{k(i)}$ that is the nearest vertex to the point $(-x_i, y_i, z_i)$ as the symmetric vertex of v_i. Notice that this step only needs to be performed once. Finally, $\mathbf{x}^{'}$,$\mathbf{y}^{'}$,$\mathbf{z}^{'}$ can be solved though linear optimization.

8 Experimental results

In this section, we first introduce implementation details of our method. Then we evaluate two key components, the joint trained feature extraction network and the glasses pose estimation aided by head pose. Next, we evaluate whole system by analyzing the final results. Finally, we discuss limitations of our work.

8.1 Implementation details

The network for joint landmark detection and segmentation is implemented in PyTorch. We set the sigma of Gaussian filtering used in heatmaps generation to be 5. The network is trained for 25 epochs on a GTX2080 using Adam[10] as the optimizer. In our experiments, the landmark prediction task reaches convergence faster than segmentation. Firstly, we set $\lambda_{balance}$ to be 200 and the learning rate to be 0.001. When the landmark loss no longer declines (12 epochs), we set $\lambda_{balance}$ to be 50 and the learning rate to be 0.0005 to achieve better segmentation. We randomly split our dataset to 3000 images for training and 300 for testing. For the rest of our technique, we set $\lambda = 400$, $\gamma = 1500$, $\lambda_C = 25$, $\lambda_L = 1$, and $\lambda_S = 100$, which are tuned for the best performance.

8.2 Feature extraction network

Firstly, we demonstrate effectiveness of the joint training network. We compare it with training the two tasks separately with the same network structure. We train

a U-Net to predict landmark heatmaps using Adam with learning rate= 0.001, which converges in 16 epochs. Meanwhile, we train another U-Net to predict segmentation, which converges in 20 epochs. In Table. 1, we show the quantitative results of the three networks on 300 testing images of 256*256 resolution. We use dice coefficient $2 \times (|X \bigcap Y|)/(|X| + |Y|)$ to measure the performance of the segmentation and the average L2 distance to measure landmark prediction. Here X and Y denote the output and the ground truth masks. It can be seen our joint training method achieves better feature extraction results. Another benefit is that only one network is needed for the two tasks. Fig. 5 shows some qualitative results. We can see that our method handles various head and glasses poses, glasses styles, illuminations, and occlusions. Notice that the dice coefficient is not high in our task. This is because the frame of glasses is very thin, thus a litter mismatch may cause a very low dice coefficient. But in this situation, the final result may not look bad, which is the common case in our experiments.

Table 1. Quantitative evaluation of joint training for glasses landmark detection and segmentation.

-	landmark L2 error	segmentation dice coeff
joint training	0.8378	0.8183
landmark only	0.8701	-
segmentation only	-	0.7801

8.3 Glasses pose estimation

Here, we evaluate our glasses pose estimation method. In Fig. 6, we show the glasses pose estimation results of different solutions: our proposed method, ours without physical constraint and ours without physical and $\theta^{sub}_{glasses}$ constraints. We render template glasses and face from two perspectives for better visualization. As Fig. 6 shows, without physical constraint, the solved glasses may float on face (d, top) or interact with face (d, bottom). Without both two kinds of constraints, the glasses may have wrong z-direction rotations (e, top) or x-direction translations(e, bottom). The wrong estimation is due to the inherent ambiguity in 2D to 3D estimation and will lead to the failure of the whole system.

8.4 Final results

Fig. 7 and Fig. 1 show our 3D reconstruction results for different glasses styles with different glasses poses. With the correctly extracted glasses features and reconstructed faces, our method achieves promising results in both glasses shape and pose estimation. More results can be found in our supplementary material.

As we do not have the ground truth 3D glasses models, quantitative evaluation is performed by using the standard dice coefficient metric between the ground truth segmentation masks and the projected masks of the reconstructed

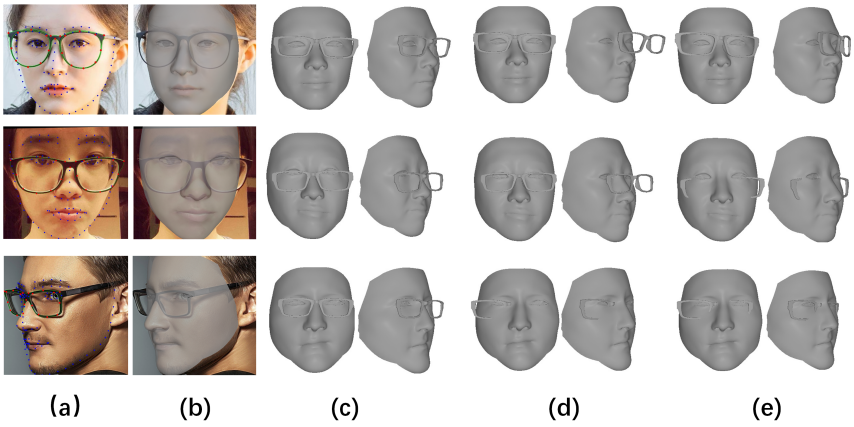

Fig. 6. Comparison of glasses pose estimation among different constraining methods. (a): face/glasses landmarks and glasses segmentation. (b): face reconstruction. (c): glasses pose of our proposed method. (d): glasses pose without physical constraints. (e): glasses pose without physical and $\theta^{sub}_{glasses}$ constraints. (c-e) are rendered under two perspectives while reserving the relative motion between the glasses and the face.

glasses. Here we first compare our method with the only single image based glasses reconstruction method [28]. The comparison is performed on their test images as they have published their results. From Tab. 2, we can see our method and [28] are comparable and we believe the difference is majorly due to the used templates. Please notice that this comparison is performed on frontal face images as [28] does not handle other head poses by design.

Table 2. Dice coefficient between the ground truth segmentation mask and the projection of the reconstructed glasses.

	[28] on [28]'s test set	ours on [28]'s test set	ours on our test set
dice coeff.	0.7306	0.6885	0.5552

The focus of our method is to handle extreme head poses so we also report the quantitative results on our 300 test images in Tab. 2, where extreme head poses and various glasses styles are included. Though the dice coefficient is worse as the test set is much more challenging, visual reconstruction results are still very promising, which are shown in Fig. 7, Fig. 1 and the supplementary material.

8.5 Limitations

Our method relies on glasses features extraction, so it cannot handle cases where the two tasks fail, like rimless or half rim glasses, and will fail if glasses is heavily occluded or outside the image. Besides, our results do not contain glasses arms

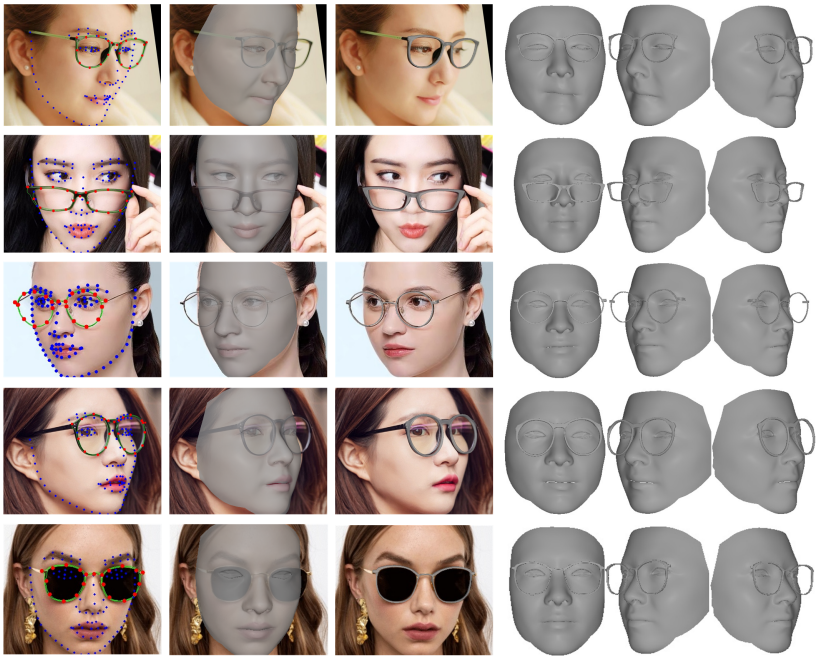

Fig. 7. Results of our method. From left to right, we show glasses/face landmarks and glasses segmentation, face reconstruction, glasses reconstruction, full reconstruction under different perspectives.

as our glasses models do not contain them. More complete reconstruction will be achieved if glasses arms are considered.

9 Conclusions

We propose a system that reconstructs eyeglasses 3D shape from a single portrait image. The system is capable to handle extreme head and glasses poses. The trained neural network jointly predicts glasses landmark and segmentation from images. The iterative reconstruction method leverages face reconstruction and utilizes shape commonality of glasses to achieve glasses reconstruction of extreme poses. Our method promotes the technique of full face reconstruction with glasses 3D shape and pose estimation.

Acknowledgements. This work was supported by the National Key R&D Program of China 2018YFA0704000, the NSFC (No.61822111, 61727808, 61671268) and Beijing Natural Science Foundation (JQ19015, L182052). Feng Xu is the corresponding author.

References

1. Beeler, T., Bickel, B., Beardsley, P., Sumner, B., Gross, M.: High-quality single-shot capture of facial geometry. ACM Transactions on Graphics (ToG) pp. 1–9 (2010)
2. Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K.: Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics (TVCG) **20**(3), 413–425 (2013)
3. De Smet, M., Fransens, R., Van Gool, L.: A generalized em approach for 3d model based face recognition under occlusions. In: Proceedings of the IEEE Computer Conference on Computer Vision and Pattern Recognition (CVPR). vol. 2, pp. 1423–1430 (2006)
4. Egger, B., Schneider, A., Blumer, C., Forster, A., Schönborn, S., Vetter, T.: Occlusion-aware 3d morphable face models. In: Proceedings of the British Machine Vision Conference (BMVC). vol. 2, p. 4 (2016)
5. Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X.: Joint 3d face reconstruction and dense alignment with position map regression network. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 534–551 (2018)
6. 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 (CVPR). pp. 8377–8386 (2018)
7. Huynh, L., Chen, W., Saito, S., Xing, J., Nagano, K., Jones, A., Debevec, P., Li, H.: Mesoscopic facial geometry inference using deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8407–8416 (2018)
8. Jiang, L., Zhang, J., Deng, B., Li, H., Liu, L.: 3d face reconstruction with geometry details from a single image. IEEE Transactions on Image Processing (TIP) **27**(10), 4756–4770 (2018)
9. Kemelmacher-Shlizerman, I., Basri, R.: 3d face reconstruction from a single image using a single reference face shape. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **33**(2), 394–405 (2010)
10. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Lepetit, V., Moreno-Noguer, F., Fua, P.: Epnp: An accurateo(n) solution to the pnp problem. International Journal of Computer Vision (IJCV) **81**(2), 155–166
12. Li, J., Yang, J.: Eyeglasses try-on based on improved poisson equations. In: Proceedings of the International Conference on Multimedia Technology (ICMT). pp. 3058–3061 (2011)
13. Maninchedda, F., Oswald, M.R., Pollefeys, M.: Fast 3d reconstruction of faces with glasses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6599–6608 (2017)
14. Niswar, A., Khan, I.R., Farbiz, F.: Virtual try-on of eyeglasses using 3d model of the head. In: Proceedings of the International Conference on Virtual Reality Continuum and its Applications in Industry (VRCAI). pp. 435–438 (2011)
15. Park, J.S., Oh, Y.H., Ahn, S.C., Lee, S.W.: Glasses removal from facial image using recursive error compensation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **27**(5), 805–811 (2005)
16. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1259–1268 (2017)

17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2015)
18. Roth, J., Tong, Y., Liu, X.: Unconstrained 3d face reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2606–2615 (2015)
19. Smith, W.A., Hancock, E.R.: Recovering facial shape using a statistical model of surface normal direction. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **28**(12), 1914–1930 (2006)
20. Sorkine, O.: Differential representations for mesh processing. Computer Graphics Forum (4), 789–807 (2006)
21. Tang, D., Zhang, J., Tang, K., Xu, L., Fang, L.: Making 3d eyeglasses try-on practical. In: Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW). pp. 1–6 (2014)
22. Tewari, A., Zollhöfer, M., Garrido, P., Bernard, F., Kim, H., Pérez, P., Theobalt, C.: 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 (CVPR). pp. 2549–2559 (2018)
23. Tran, A.T., Hassner, T., Masi, I., Paz, E., Nirkin, Y., Medioni, G.G.: Extreme 3d face reconstruction: Seeing through occlusions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3935–3944 (2018)
24. Tran, L., Liu, X.: Nonlinear 3d face morphable model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7346–7355 (2018)
25. Wu, C., Liu, C., Shum, H.Y., Xy, Y.Q., Zhang, Z.: Automatic eyeglasses removal from face images. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **26**(3), 322–336 (2004)
26. Wu, W., Qian, C., Yang, S., Wang, Q., Cai, Y., Zhou, Q.: Look at boundary: A boundary-aware face alignment algorithm. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
27. Yuan, M., Khan, I., Farbiz, F., Niswar, A., Huang, Z.: A mixed reality system for virtual glasses try-on. Proceedings of the International Conference on Virtual Reality Continuum and Its Applications in Industry (VRCAI) (12 2011)
28. Yuan, X., Tang, D., Liu, Y., Ling, Q., Fang, L.: Magic glasses: from 2d to 3d. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) **27**(4), 843–854 (2016)
29. Zeng, X., Peng, X., Qiao, Y.: Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 2315–2324 (2019)