



Improved 3D human face reconstruction from 2D images using blended hard edges

Yueming Ding¹ · P. Y. Mok^{1,2}

Received: 5 August 2023 / Accepted: 12 April 2024 / Published online: 11 May 2024
© The Author(s) 2024

Abstract

This study reports an effective and robust edge-based scheme for the reconstruction of 3D human faces from input of single images, addressing drawbacks of existing methods in case of large face pose angles or noisy input images. Accurate 3D face reconstruction from 2D images is important, as it can enable a wide range of applications, such as face recognition, animations, games and AR/VR systems. Edge features extracted from 2D images contain wealthy and robust 3D geometric information, which were used together with landmarks for face reconstruction purpose. However, the accurate reconstruction of 3D faces from contour features is a challenging task, since traditional edge or contour detection algorithms introduce a great deal of noise, which would adversely affect the reconstruction. This paper reports on the use of a hard-blended face contour feature from a neural network and a Canny edge extractor for face reconstruction. The quantitative results indicate that our method achieves a notable improvement in face reconstruction with a Euclidean distance error of 1.64 mm and a normal vector distance error of 1.27 mm when compared to the ground truth, outperforming both traditional and other deep learning-based methods. These metrics show particularly significant advancements, especially in face shape reconstruction under large pose angles. The method also achieved higher accuracy and robustness on in-the-wild images under conditions of blurring, makeup, occlusion and poor illumination.

Keywords Computer graphics · Deep neural network · 3D face reconstruction · Blended hard edges

1 Introduction

Face reconstruction is important in computer vision and graphics and has been widely applied to animation, virtual try-on systems and face identification [1, 2]. Face reconstruction based on 3D data often relies on specialized hardware to collect and process 3D data [3, 4], which increases the cost of applications and constrains the widespread adoption. Comparatively, face reconstruction based on 2D images has more applications. An important aspect of image-based reconstruction is the use of prior knowledge of 3D face shape. The 3D morphable face model (3DMM) is a statistical face morphable model with

prior information of a face shape, and the application of 3DMM to face reconstruction can convert the problem of shape prediction into a 3D model fitting problem. Traditional face reconstruction approaches include pixel-wise reconstruction [5] and feature-based reconstruction [6]. In pixel-wise face reconstruction, the difference between an observed face image and a synthetic face image from a face model is directly minimized in a pixel-by-pixel manner, whereas the feature-based reconstruction method measures the difference between the features extracted from the observed real image and those extracted from the synthetic image based upon the estimated face model. Compared with pixel-wise face reconstruction methods, which are sensitive to occlusion and illumination, feature-based reconstruction methods are comparatively robust and computationally efficient [7].

Due to the need for robustness in real-world applications, this article mainly focuses on feature-based face reconstruction, and there exists different types of facial features in face reconstruction, such as landmark features.

✉ P. Y. Mok
tracy.mok@polyu.edu.hk

¹ The Hong Kong Polytechnic University, Hunghom, Hong Kong

² Laboratory for Artificial Intelligence in Design, Hong Kong Science Park, Hong Kong

Landmarks provide coarse facial shape information, which can be used for face tracking [8], face alignment and facial shape initialization. Nevertheless, face landmarks only present rough or approximated face shape information, they could not describe detailed geometric shape of the face, for example, whether or not the face is with high cheekbones or a prominent forehead. The locations of landmarks are often drifted or occluded when the face is posed, resulting in reconstruction errors under large pose angles [9–12]. To address this problem, some researchers [9, 10] thus proposed to discard the moved landmarks; however, this would suffer from the loss of facial boundary constraints. Alternatively, 3D landmark configurations under various poses were used, though this is only applicable for limited pose ranges [11].

Other than landmark-based reconstruction, other dense features, such as depth images [13] or optical flow [14], were used for shape reconstruction. Nevertheless, these dense features require special equipment for data collection and the methods are often computationally intensive. Compared to the dense features, edge features provide sparse yet more detailed representation of the 3D shape. Edges are abstract 2D representations of real-world objects that convey geometrical information of the 3D objects. Humans are able to perceive 3D shape information from these edges and contours, despite the lack of detailed visual cues such as texture and shadows. As a result, a number of work have been reported for 3D object reconstruction using edges and contours information [15, 16].

It is therefore attractive to use, in addition to landmarks, edge features for face reconstruction, since it provides robustness under large face pose angle situations and also provides detailed geometric face information. Traditional edge detectors, e.g. Canny detector, often generate a lot of noises, and they are sensitive to low illumination, makeup and occlusion, thereby adversely affecting the optimization of the face reconstruction.

This paper proposes to improve edge-based 3D face reconstruction using hard-blended edge features (see Fig. 1). With the development of deep learning, convolutional neural networks (CNNs) have performed well in terms of the object contours detection, enabling contour features of objects to be extracted with less noise and greater robustness to poor illumination [17, 18]. With hard-blended edge maps, which are generated from features extracted from the medial layers of a deep neural network pretrained in face alignment task [19], the new method improves the effectiveness and robustness of image-based 3D face reconstruction. This method involves four stages, as illustrated in Fig. 2. In the first stage, facial features are extracted from input images based on a deep neural network as well as landmark features. The second stage involves a method to generate hard-blended edge maps

based on features extracted from the medial layers of a deep neural network pretrained in face alignment. In the third stage, face reconstruction is initiated using landmarks features. In the fourth stage, the face model is optimized using hybrid loss function of landmarks and edges. The proposed method, as highlighted in red outline in Fig. 2, is comprehensively evaluated and compared to other face reconstruction methods, including landmark-based and edge-based, extracting from traditional edge detectors and other deep network [20]. The results show that both deep neural network-based methods represent an improvement on the face reconstruction compared with methods based on edges from traditional edge detectors and also have better performance in terms of face shape. It also shows that hard edge maps are more effective than soft edges for face reconstruction. Moreover, not only the reconstruction abilities of the different methods are compared, but also the differences in the reconstructed results for different poses are analysed, demonstrating the effectiveness of the proposed method. The key contributions of this paper are summarized as follows:

- An effective and robust edge-based scheme is designed for the reconstruction of 3D human faces from input of single images. Without relying on specialized 3D data collection equipment or matrix of multi-cameras, the input images can be face images with arbitrary face angles or in-the-wild images, the method reconstructs 3D face shapes directly from sparse facial features extracted from 2D images, including both landmarks and edges.
- It proposes the use of hard-blended edges for 3D face reconstruction, and the hard-blended edges are an organic combination of contouring edges, namely the face shape regulating edges generated from features extracted with a neural network, and those edges from a Canny edge detector. To optimize the 3D face shapes, landmarks are first extracted from the input images to initialize the face pose and shape parameters for a 3D Morphable Model (3DMM) of face, while the face shape is optimized using the hard-blended edges with an edge-to-edge loss.
- The proposed 3D face reconstruction method achieves a higher degree of accuracy and robustness on both a synthetic and an in-the-wild facial image dataset in comparison with other edge-based reconstruction methods or deep learning-based methods. In particular, the proposed method can reconstruct more accurate 3D face shapes for in-the-wild images under different conditions including blurriness, ill-illumination, makeup, and occlusion.

The rest of the paper is organized as follows. In Sect. 2, related work is reviewed. Section 3 presents the proposed

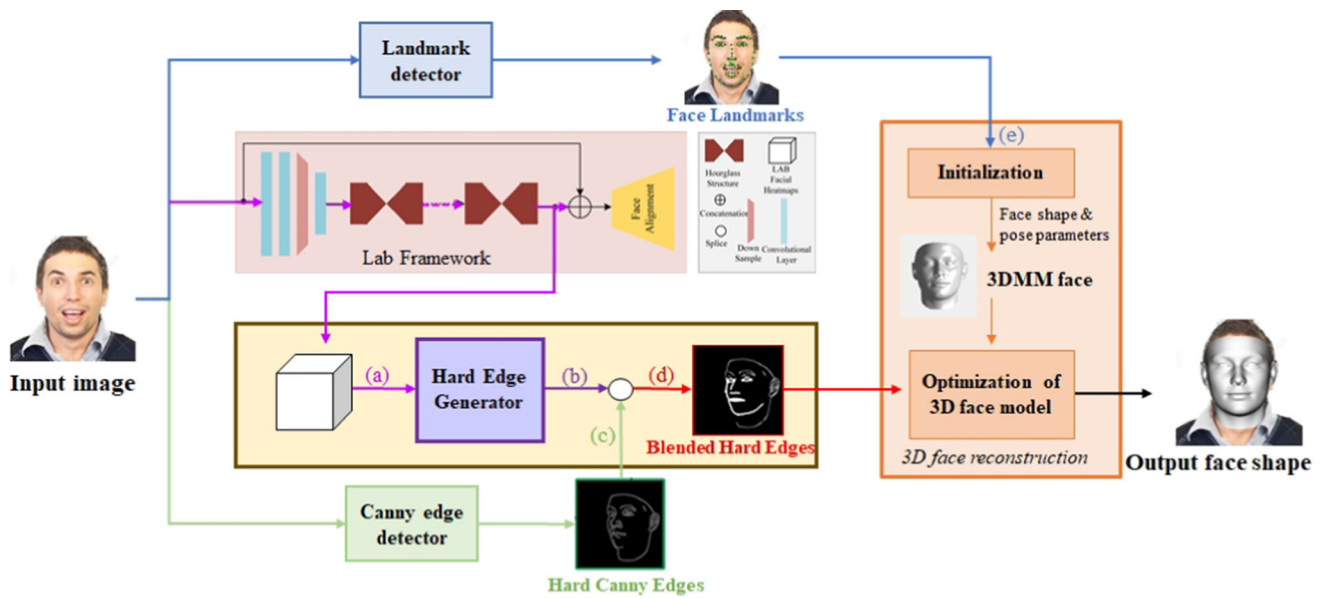


Fig. 1 3D face reconstruction from blended hard edges: an image-based reconstruction method of 3D faces that a novel feature processing method is developed to obtain rich 3D geometric

information from input face images by blending contouring features learned from deep learning network and Canny edges

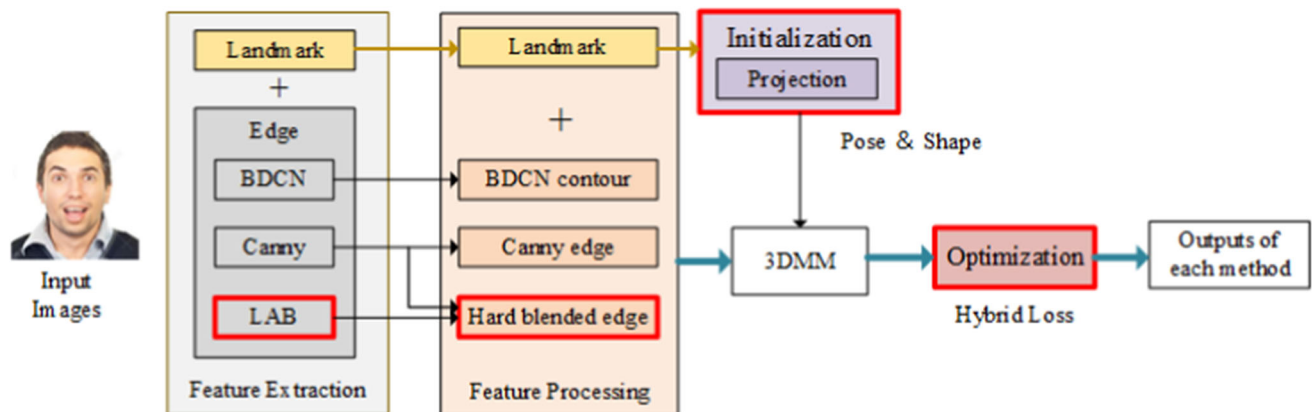


Fig. 2 Overview of the four-stage face reconstruction approach used (outlined in red bold lines) and other methods compared. With an input face image, it will be processed sequentially for feature extraction (stage 1), feature processing (stage 2); next, the processed image features will be used for 3D face reconstruction through an initialization (stage 3) and optimization (stage 4) process. Such 4-stage pipeline is a standard approach for 3D face reconstruction,

and the key novelty of current study lies in the new proposal in feature extraction and processing as well as the corresponding hybrid loss in face shape optimization, as highlighted in red outline. For comparative study, the proposed method is compared to other methods, including using only landmark features, or considering both landmarks and other features like Canny, BDNC and other deep learning-based methods in the experiment (Color figure online)

reconstruction method. Section 4 compares the proposed method with another state-of-the-art approach, and Sect. 5 presents the conclusions and suggestions for future work.

2 Related work

This section reviews related studies on face reconstruction based on different facial features, such as landmarks, edges and other features. In addition, because the present study is

about the employment of deep learning in edge-based face reconstruction, studies involving edge detection methods based on deep learning are also covered.

2.1 Reconstruction based on facial features

2.1.1 Reconstruction based on landmark features

Although landmark features provide limited facial geometric information, it has been widely used in face

alignment, face recognition and face reconstruction [5, 21]. For example, Breuer et al. [22] employed landmarks to locate the face and classified the pose angle of the face based upon the images, and proposed a method to automatically reconstruct a textured 3D face model from a single image or a video frame. They used the support vector machine (SVM) detector to detect faces from an input image or video frame. Thereafter, the detected faces were classified into different ‘views’ (approximate poses) based on the landmarks. In the final optimisation step of face reconstruction, the 3DMM was fitted to the image using the analysis-by-synthesis method.

The other application of landmarks in face reconstruction is rough shape estimation. For example, Aldrian and Smith [23] proposed an algorithm to reconstruct a 3DDM from a face image, using limited 2D feature points to approximate the shape of the face. Because of several drawbacks, face reconstructions based on landmark feature have been mainly used in the initialization stage or for rough face reconstruction. Lv et al. [24] employed landmarks to train a model for predicting 3D face model from video; however, they can only reconstruct a coarse face shape due to only weak shape constraint being imposed by landmarks. Wood et al. [25] proposed dense landmarks for face reconstruction, but most of such dense points were estimated instead of ground true geometric values, making it difficult to evaluate its effectiveness. One disadvantage is that the landmarks are not sufficiently robust in terms of face pose variations, in which the locations of landmarks are often drifted or occluded when the face is at different pose angles. Another disadvantage is that the face landmarks only convey a very approximate face shape information.

2.1.2 Reconstruction based on edge features

Other than landmark features, many researchers have introduced contour or edge information into their face reconstruction schemes. The earliest edge-based face reconstruction was the ASM model [26]. Blanz and Vetter [5] used boundary features for face model fitting and established an active shape model. Afterwards, a number of researchers focused on edge-based face reconstruction. For example, Moghaddam et al. [27] used multi-view silhouettes to fit a 3DMM model [28]. They obtained face silhouettes from fixed multi-views and employed a bundled optimization method to fit the 3DMM model. Romdhani and Vetter [29] calculated the edge distance by a mixed energy function during the optimization stage. Fitzgibbon et al. [30] proposed a soft edge-based 2D and 3D registration method, in which a gradient magnitude threshold with non-maximum suppression was used as the edge detector, the edges were therefore smooth and soft.

Moreover, a smooth loss surface function was used in the optimization process. Romdhani and Vetter [29] applied the soft edge-based method to face reconstruction. Canny edges capture the geometric information of 3D scene in an image, facilitating both 2D image reconstruction [31] and 3D facial reconstruction [32]. Bas et al. [32] proposed their contour-based method, which was and still is the state-of-the-art method for edge-based face reconstruction. They first aligned the landmarks to initiate the pose parameters of the model; then, they used a Canny edge detector to generate hard edges from the face image and also used orthographic projection to generate the contours of the face model. A K-nearest neighbours (KNN) search was applied to compare the contours from the reconstructed face model with the edges detected from the original face image. An edge iteration optimization process was being used to minimize the distance.

Nevertheless, there still existed some challenges concerning those feature-based face reconstruction methods. Keller et al. [33] reported that the contour-based face reconstruction method represented a discontinuous and non-differentiable optimization problem, and the reconstruction can be affected by occlusions. A reconstructed face model may differ from the ground truth even though it had almost the same contours. The other disadvantage was that an edge detector can generate a lot of noise, which may adversely affect the face reconstruction process.

2.1.3 Reconstruction based on other features

Except for the landmark and contour type features, other types of features are also widely used in face reconstruction, such as other local facial features, e.g. SIFT [6]. For example, Huber et al. [6] proposed a fitting method that applied SIFT to local facial features to reconstruct a 3DMM and used cascaded-regression-based methods to derive the gradient directly from the data, rather than applying differentiation.

Shape-from-shading is another traditional method used in face reconstruction. Being based on the assumption that the illumination is invariant for each view angle, it estimates the surface orientation at each pixel directly from shading. The best orientation is then optimized by satisfying image irradiance constraints. For example, certain researchers [12, 34] employed the shadow and illumination information to estimate the shape of a human face. Amberg et al. [4] reconstructed face models from calibrated multi-view stereo images. Patel and Smith [3] used shading information to reconstruct a 3DMM and employed a surface normal error to assist in the reconstruction of the face model. They proposed a framework for predicting a per-vertex albedo map and a bump map. Khan et al. [35] proposed a coarse-to-fine face reconstruction through

coarse depth map and displacement map estimated through inverse rendering. By computing the albedo and lighting direction, the normal to the surface of the face in the image could be derived, and the difference between this and the normal to the predicted face model could be optimized in the reconstruction process.

Other approach represents 3D information through 3D discrete moments, which present 3D shapes like voxels employing discrete polynomials for representation [36–38]. Such approach ensures invariance under translation, scaling, and rotation. These methods demonstrate efficiency of computation and quality of reconstruction [39, 40]. They are also suitable to integrate with deep learning [41, 42], which were mainly applied to general 3D object classification and reconstruction. In the case of 3D face reconstruction, since face shapes exhibit complex topology and shape, the use of mesh-based representation, namely 3D Morphable Models (3DMM), remains more prevalent for industrial applications due to its effectiveness and flexibility in capturing high-quality facial geometry.

Landmark-based reconstruction or face alignment methods are usually used in small or medium pose angles, since landmarks may be occluded or drifted when the face is in large pose angles. Some researchers fitted a 3DMM to a cascaded convolutional neural network to generate dense information, such as PNCC features [43] and pixel consistency features [44] for face alignment in large pose angles. Furthermore, Shang et al. [44] used multi-view dense features and Lv et al. [45] employed face parsing for 3D face reconstruction. Nevertheless, these dense features require more computing resources and result in slower convergence.

2.2 Edge detection methods based on deep neural networks

Following the development of deep learning, many researchers have applied convolutional neural networks to detect object edges from images [46]. Shen et al. [18] used a fully convolutional neural network to detect edges from images. Multi-level features extracted from convolutional neural networks [47] have also been proposed for edge detection. Compared with traditional edge detector, those convolutional neural networks with multi-level features fusion processing can extract object geometric information with less noise. The Bi-Directional Cascade Network (BDCN) [20] employ cascaded structures to extract multi-level features and achieve great success. In the present study, BDCN is employed to extract edge features for face reconstruction and it is compared with face reconstructions based on traditional edge detectors, such as Canny edge detector.

Another kind of neural network structure, the hourglass network structure [48], can detect local and global geometric information. Because, an hourglass network structure was able to downsample and upsample the image to different scales to capture the information as well as fuse the resulting features of different scales, which made it robust to the variations of scale and able to detect both global and local geometric information.

The earliest hourglass structure was used for landmark annotation, since they were robust to scale, translation and rotation. Newell et al. [48] applied hourglass networks to the task of human pose landmark location, while Yang et al. [49] also used it for the task of face landmark location. Bulat and Tzimiropoulos [50] combined a state-of-the-art hourglass network structure with a residual block in 2D and 3D landmark location tasks.

In previous studies, deep neural networks have been widely used in face landmark annotations [51] but not for face contour extraction purposes. Inspired by the research involving human pose annotation, Wu et al. [19] introduced hourglass network structures into boundary-based face alignment. They stacked several hourglass architectures to extract face boundary heatmaps with Gaussian distribution, which they then used to assist the face alignment from in-the-wild images. Since the present study is focused on face alignment, different from object contour detection, the medial-level features of Wu et al. [19]’s network, which contain more valuable facial information, are used here to extract facial edge features for the face model reconstruction.

3 Method

3.1 Overview of the 4-stage method

This study reconstructs face model from a 2D image based on a proposed hard-blended edge feature in a four-stage framework, as illustrated in Fig. 2. The key concept of the present study was based on the observation that the edges present more detailed shape information and are also more robust under large pose angles compared with landmarks. Nevertheless, face reconstruction from edge features is a complex task, as reviewed in Sect. 2.1.2. The traditional edge detector is sensitive to noise, occlusion, and varied illumination conditions, resulting in a great deal of noise generated, which adversely affects the quality of the reconstructed faces. The present study introduces a deep neural network into the face edges extraction, which detects edges with more detailed shape information and improves robustness for large pose angles, occlusions and poor-illumination. Nevertheless, existing state-of-the-art edge detection networks, e.g. BDCN [20], were not trained

on face images and thus do not perform well in detecting facial edge features in case of poor-illumination or occlusion. Hence, in the present study, a neural network, pretrained for face alignment, was introduced to generate facial edges for application in face reconstruction.

The overall method involves four stages (see Fig. 2):

- (1) The *extraction of facial features* from input images based on deep neural networks, which include landmarks, traditional edge features and shape regulating edge features from deep neural networks;
- (2) The development of a *feature processing* method to generate *hard-blended edge* maps based on medial features of LAB network, a deep neural network pretrained for face alignment;
- (3) The *initialization* of pose and shape parameters for a face model using landmarks; and
- (4) The face model is *optimized* with *hybrid loss function* containing both landmarks and edges.

In the experimental verification, the proposed method will be compared to other methods for face reconstructions using other landmarks and features.

3.2 Facial feature extraction

Both landmark and edge features are extracted and used in the proposed method. A landmark detector [52] is used to extract landmarks from face images, as previously in [32]. Landmark features are used in the initial stage of the method. As reviewed in Sect. 2.1.1, face models can be reconstructed purely based on landmarks, and this will be compared with the proposed method in later Sect. 4.

Edge features are the most important in the face shape reconstruction exercise. To compare the capability of deep learning edge-based face reconstruction methods with methods based on traditional edge detector, three kinds of edges features are investigated. The first one is generated by the traditional Canny edge detector. The second method involved contours from a state-of-the-art contour detector based on deep neural network, BDCN [20]. The third one is the proposal of this study, namely blended hard edge

features, which are generated from medial level features extracted from a deep neural network pretrained for face alignment [19].

3.3 Feature processing

3.3.1 Hard-blended edge generation

Most edge detection deep neural networks focus on extracting the contour of the object, in which the semantic information of eyes, nose and mouth would not be detected. In the present study, a deep neural network pretrained for face alignment [19], which can detect a great deal of geometric information in its low-layers or medial-layers, was selected.

Extracting feature maps from a LAB deep neural network

The present 3D face reconstruction method utilizes the LAB framework [19] for contour extraction, with convolutional layers being able to extract the geometric information of an image. The initial convolutional layers can capture low-level local geometric information, while the later convolutional layers can capture high-level global geometric information but with relatively low resolution. Considering the characteristics of a convolutional neural network, an hourglass neural network structure [48], with top-down and bottom-up designs, can capture well both local and global information from an image. This structure is used in human pose landmark extractions as well as face alignment, such as the LAB neural network [19]. The LAB network stacks several hourglass network structures to extract face contour heatmap features and then concatenates features extracted after those stacked hourglass network structures for face alignment, as shown in the LAB framework of Fig. 1.

Generating hard-blended edges In the present study, the heatmaps generated from the stacked hourglass network structures of the LAB framework [19] are extracted, representing a Gaussian face contour heatmap (soft edges). The soft contours are transformed into hard edges in the present study (see Fig. 3).

Algorithm 1 Generation of hard LAB edge maps

Input: X = soft facial heatmap features of size $m \times n$
 $X_{(i,j)}$ = pixel value of X at coordinate (i,j)
 T = threshold

Output: Hard blended edge map $S_{Hard_LAB_edge}$

Symbol: m, n : number of row and column
 S : set of pixels with coordinates and pixel-values
 ∇ : discrete gradients
 $N(\cdot)$: the number of elements in a set
 $Sign(\cdot)$: obtain the sign of a number
 $append(\cdot)$: add elements to a set

```

1: Row direction:
2:  $S_{row} = \emptyset$ 
3: for  $i \in (1, m)$  do
4:    $S_t = \emptyset$ 
5:    $S_{t_{conv}} = \emptyset$ 
6:   for  $j \in (1, n)$  do
7:     if  $X_{i,j} > T$  then
8:        $\nabla_j = \frac{\partial X(i,j)}{\partial j}$ 
9:        $S_t$   $append(\nabla_j)$ ;
10:       $S_{t_{conv}}$   $append(X(i,j))$ ;
11:     else if  $N(S_t) > 1$ ; then
12:        $K = N(S_t)$ 
13:        $t=0$ ;  $k_{max} = 0$ ;
14:       for  $k \in (1, K)$  do
15:         if  $Sign(\nabla_{j-K+k}) \neq Sign(\nabla_{j-K+k+1})$ 
16:           then
17:             if  $t < S_{t_{conv}}(i, j - K + k)$  then
18:                $t = \max(t, S_{t_{conv}}(i, j - K + k))$ ;
19:                $k_{max} = k$ 
20:              $S_{row} \cdot append((i, j - K + k_{max}))$ ;
21:              $S_t = \emptyset$ 
22:              $S_{t_{conv}} = \emptyset$ 
23: Column direction:
24: Repeat steps 3-21 to calculate  $\nabla_j$  and  $S_{col}$ 
25:  $S_{Hard\_LAB\_edge} = S_{row} \cup S_{col}$ 

```

As illustrated in Fig. 3, hard-blended edge generation method combines hard LAB facial edges and hard Canny edges for feature processing. The inputs of the hard LAB edges generation are features extracted from stacked hourglass structure from the LAB network. These input features with initial size of $13 \times 64 \times 64$ are first upsampled by linear interpolation to the size of $13 \times 384 \times 384$, matching the size of input image. The input facial heatmaps have the following characteristics and/or limitations. First, the features are blur soft pixels with Gaussian distribution. Second, the heatmap features are non-convex at some regions, such as the eye region and the corners of the mouth. Hence, it is hard to localize pixels on the heatmaps to form a hard clear edge map.

An algorithm is developed to generate hard LAB edges from the soft heatmaps by segmenting the soft edges into convex subsets. Convex set and non-convex set are concepts in geometry and mathematical analysis that describe types of set based on their structural properties. A convex set is defined as a set of points in which the line segment

connecting any two points within the set lies entirely within the set itself. In other words, a set is convex if every point between two points in the set is also in the set. Intuitively, a convex set is one that does not have any ‘holes’ or ‘dents’ in it. A non-convex set is the negation of a convex set. The proposed algorithm generates hard LAB edges as follows. First, the non-convex feature heatmap is segmented into two subsets along row and column directions. Then, the gradient of the pixel values in the row or column directions is calculated as a vector of size 1×384 . By the gradient signs, these convex regions are separated from the non-convex subsets along column or row directions, in which the coordinates of each region are recorded as a convex subset. Hence, since each subset of the column or row directions is convex, the coordinates with the largest pixel values can be selected to generate a coordinate set. *The Gaussian contour heatmaps are anisotropic in the row and column directions; for example, the contours of the chin show larger variance in the horizontal direction than in the vertical direction.* The coordinate sets from convex subsets are then combined in both the row and column directions to generate a binary hard LAB edge map.

Since the LAB network only considers the face region, without information about the shapes of the ears or neck, the generated LAB hard edges are blended with the edges from the Canny edge detector for generating a more complete hard facial edge map. A bounding box, which enclosing the facial region closely, is detected in hard LAB edge maps.

3.3.2 Landmarks

Landmarks extracted by [52] are used in the initialization and the comparison part of the study. As mentioned in related work, there are two kinds of landmarks, namely 2D landmarks and 3D landmarks.

Among 2D and 3D landmarks, some are *static landmarks* [2], shown as blue points in Fig. 4, which represent the common parts of the two kinds of landmarks and annotate the eyes, nose, mouth and eyebrow regions. These landmarks can be mapped onto particular vertices lying on the same regions on the 3D face model. The difference between two kinds of landmarks is the annotations on the jawline and chin regions, shown as green points and red points, respectively, in Fig. 4. The 3D landmarks [50] shown as the green points in Fig. 4 mark the jawbone region, mapping onto the particular vertices of the 3D face model. However, these 3D landmarks may not depict the ground true profile or contours of the face when the face is posed. Hence, 3D landmarks will not be used to represent face shape information in the present study. In contrast to the green points (3D landmarks) along jawbone, the red points in Fig. 4 represent 2D landmarks, which will shift

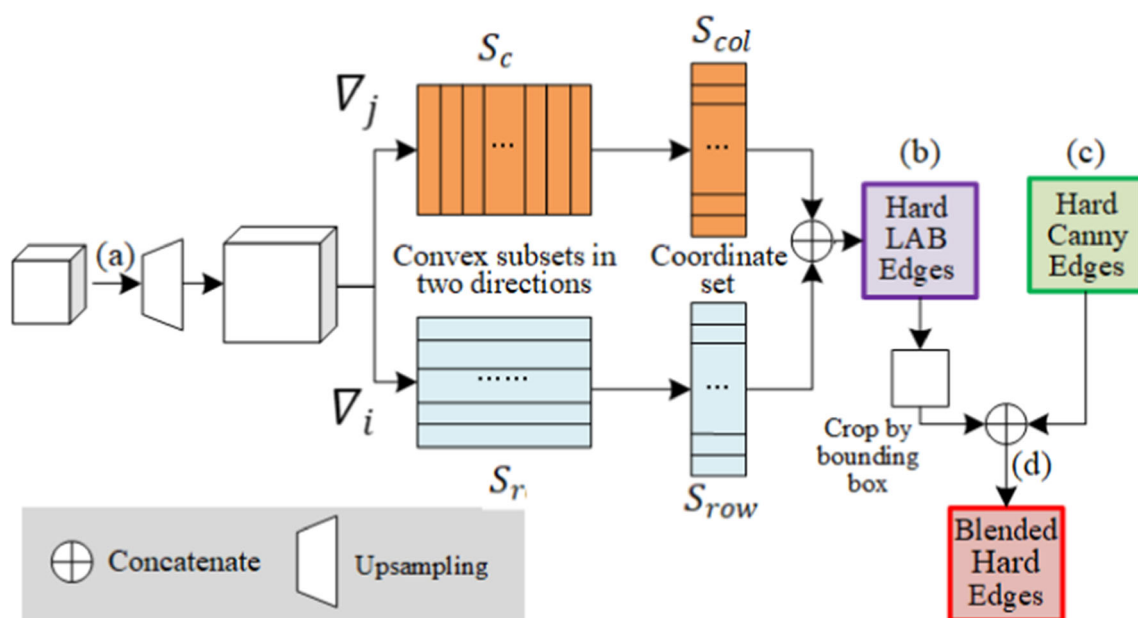


Fig. 3 Feature processing: hard-blended edge generation: the features extracted from the LAB framework (labelled as (a) in this figure as well as in Fig. 1) are unsampled to the size of the input image. Such features are soft heatmap features and are noisy; then, the gradients ∇ along column direction S_c and row direction S_r are calculated to separate convex subsets along these gradients using a threshold. From

the convex subsets, the peak pixels are selected as S_{col} and S_{row} , which are combined to generate the hard LAB edges (labelled as (b)). Finally, the hard LAB edges are combined with hard canny edges (c) defined within the bounding box of face region to obtain the blended hard edges (labelled as (d)) for subsequent face reconstruction use

with a change in the head pose/orientation. These landmarks are not mapped onto specific vertices of the 3D face model, and they are defined as *dynamic landmarks*. Nevertheless, the dynamic landmarks can depict well the actual contour of the face; hence, they are employed to represent the face shape information in the present study.

Static landmark points are employed to initialize the pose and shape parameters. The dynamic landmarks, namely 2D landmarks located on the jawline region, are employed to represent the face shape. The dynamic landmarks cannot be defined on the 3D face template model [2, 53], because their positions can move along the changes of the head pose (see the last row in Fig. 4). In 3D face reconstruction, the vertices on face template model paired with dynamic landmarks are estimated for face shape reconstruction. For estimating those vertices, the vertices on the occluding contours of the 3D face model are projected onto the screen, and the KNN algorithm is used to detect the nearest projected vertex for each dynamic landmark point.

3.3.3 Edges

For comparative study of edge-based face reconstruction, the traditional Canny edge detection method and an end-to-end deep neural network methods BDCN [20] are

employed, respectively, to extract edges from images, which are further used to reconstruct face models.

There are differences in the edges extracted from the above two methods, respectively. Edges extracted from a deep neural network are, as shown in red lines Fig. 5, a contour related to face shape (also called shape regulating edges), which separates the face from the background and is invariant to illumination. Nevertheless, such contour-based edges cannot convey valuable details, such as the location and shape of the eyes and nose. In contrast to this, edges extracted from the Canny edge detector, shown as the blue lines in Fig. 5, are the textured edges, which have a great amount of details but also some noises caused by makeup, occlusions and illumination.

In the edge-based face reconstruction, the 3D face morphable model will be optimized through minimizing the edge distances from 2D face images and generated from 3D face morphable model. To contour edges from 3D face models, contours are generated by computing the visibility and position of triangular faces on the mesh model. More specifically, if the mean vertex normal of adjacent faces of the mesh model have an opposite direction (i.e. opposite signs, one positive and one negative, of the z-coordinates for two adjacent triangles), the vertices lying on the adjacent edges will be deemed to belong to occluding contours. These invisible boundaries will be filtered through Z-buffer. A Z-buffer, also known as a

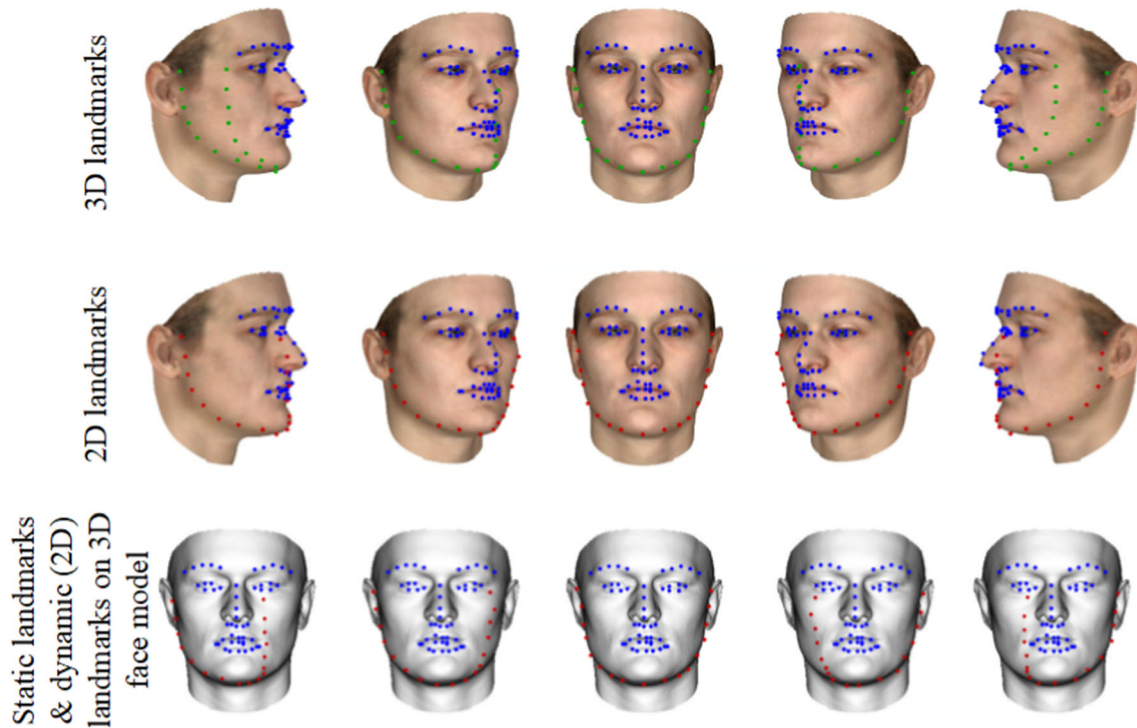


Fig. 4 Illustrations of various landmarks: green points represent 3D facial landmarks, blue points represent static landmarks, and red points represent dynamic landmarks

depth buffer, is commonly used in computer graphics to render three-dimensional scenes. It manages the depth information of each pixel in a scene relative to the viewer's perspective. In the proposed method, the Z-buffer value determines the visibility of contours projected from 3D face shape. As each pixel lying on contours is processed, the Z-buffer is updated to store the depth information of the current pixel. This depth information is then compared with the existing depth value stored in the Z-buffer. If the new pixel is closer to the viewer, its depth value replaces the existing one in the Z-buffer, indicating that this pixel is more visible. Conversely, if the new pixel is farther away, it is discarded as it is occluded by the closer pixel. By doing so, the 2D coordinates of edges lying on the contours of the 3D face are obtained through orthographic projection.

3.4 Initialization

The pose and shape parameters of the 3D morphable face model (3DMM) require initialization before further optimization, because the face model reconstruction based on boundaries is a non-convex problem [33]. In this initialization stage, the landmarks from 2D image are used as a reference for pose prediction. The landmarks are detected by means of a facial landmark extractor [52]. As introduced in previous section, because the static parts of the

landmarks can have an invariant mapping with vertices on the 3D face model, these landmarks can be used to initialize the pose and shape of the 3DMM.

3.4.1 3D morphable model (3DMM)

3DMM is a widely used parametric face model, in which facial shapes are generated based on a linear combination of shape and appearance eigenvectors. In the present study, a publicly available 3D face model, the Basel face model (BFM) [28], is used,

$$V = V_{mean} + P_s \cdot \alpha \quad (1)$$

$$T = T_{mean} + P_t \cdot \beta \quad (2)$$

where V_{mean} and T_{mean} represent the mean face shape and the texture map, respectively; P_s is a matrix containing the principal component vectors for the shape, and α is the parameter vector for the shape. Similarly, P_t represents the principal component vectors for the appearance, and β is the parameter vector for the texture.

3.4.2 Orthographic projection

An orthographic projection, $P(\cdot)$, is used as the camera model to predict the pose parameters. Each vertex in the 3D face mesh is projected onto a 2D image using a rotation matrix R , a scaling factor s and a translation vector t :



Fig. 5 Texture edges and occluding contours being illustrated on input face images: in-the-wild and synthetic face images and the detected edges being overplotted on top of the input face images with red lines representing occluding contours, while blue lines representing texture edges

$$P(\cdot) = sRV + t \quad (3)$$

where V represents the coordinates of vertices in the 3D face model. Orthographic projection $P(\cdot)$ is also known as a weak perspective projection.

A rotation matrix R can be generated by singular value decomposition:

$$R = uv^T \quad (4)$$

$$H = (V'_{lmk} - \bar{V}'_{lmk})(V_{lmk} - \bar{V}_{lmk})^T \quad (5)$$

$$[u, s, v] = SVD(H) \quad (6)$$

In Eq. (5), V'_{lmk} and V_{lmk} represent the coordinates of a landmark in the 2D image and the predefined corresponding points in the 3D face model, respectively. \bar{V}'_{lmk} and \bar{V}_{lmk} are the centre point coordinates of V'_{lmk} and V_{lmk} , respectively. $SVD(\cdot)$ in Eq. (6) is the singular value decomposition operation, and $[u, s, v]$ represent the output of the singular value decomposition process.

3.5 Optimization

After the initialization of the face pose and shape model, an iterative optimization method is employed to predict the 3DMM shape parameters. In the first step, the pose and shape parameters are initialized via landmark alignment. The 2D coordinates of the vertices lying on occluding contours are generated from the 3D face model. The KNN algorithm is used to find the corresponding points lying on the detected edges from the 2D image for each vertex lying on the occluding contours. Next, the pose and shape parameters are updated by minimizing the distances between the points lying on the edges extracted from the 2D image and on the occluding contours extracted from the predicted 3D model. In the final optimisation, the pose parameters are fixed, and the shape parameters are optimized to give an error less than a given threshold. A hybrid loss function is employed at this stage to combine the weighted landmark function with the edge loss function

and the penalties for the shape parameters. A trust-region-reflective algorithm is applied for optimization to obtain the final shape parameters.

Landmark loss function The landmark loss is used in both the initialization and formal training stages. The square of distance between the landmarks predicted from the 2D image and the corresponding projected pre-defined vertices on the 3D face model is calculated by,

$$E^{lmk} = \frac{1}{N} \sum_{i=1}^N \|v_i^{lmk} - P(v_i^{lmk})\|^2 \quad (7)$$

where $P(\cdot)$ represents the orthographic projection operation given in Eq. (3); v_i^{lmk} denotes N landmark points predicted from the input image; and v_i^{lmk} denotes the predefined corresponding points on the 3D mesh.

Contour loss function The contour loss calculates the distance between the points lying on the facial edges detected from a 2D image and the points on the projected occluding contours from a 3D face model.

$$E^{edge} = \frac{1}{N_C} \sum_{j \in C} \|v_j^{edge} - P(v_j^{cont})\|^2 \quad (8)$$

In Eq. (8), v_j^{cont} represents vertex j on the occluding contour of the predicted mesh model, and $j \in C$, where C is the set of vertices on the contour and N_C is the total number of vertices on the contour. v_j^{edge} represents point coordinates of the input image that correspond to the vertex j on the occluding contour.

Penalty loss function The shape parameters of the 3DMM model are expected to follow a normal distribution [7]. To prevent the face shape from diverging, a regularization term is applied to normalize the face shape model by

$$E^p = \sum_{k=1}^N \frac{\alpha_k^2}{\sigma_k^2} \quad (9)$$



Fig. 6 Edge features and 3D face reconstructions for images from the synthetic dataset. From the left to right: inputs, landmarks, BDCN edges, hard Canny edges, LAB heatmaps, and hard-blended edges

where σ_k^2 represents variance associated with the k th principal component, α_k is the shape parameter vector in Eq. (1), which weight N principal components about the shape of the face model.

Hybrid loss function A hybrid loss function is used in the final optimization stage to combine the weighted landmark loss function, Eq. (7), with the contour loss function, Eq. (8), and the penalty function, Eq. (9), as follows.

$$E^{hybrid} = \omega_1 E^{lmk} + \omega_2 E^{edge} + \omega_p E^p \quad (10)$$

where ω_1 represents the weight for the landmark loss function E^{lmk} , ω_2 represents the weight for the edge loss function E^{edge} , and ω_p represents the weight for the penalty function E^p . Only landmark loss function, Eq. (7), will be used in the initialization stage, while the hybrid loss of Eq. (10) is used in the training stage.

4 Experiments

In the experimental evaluation, all the methods were tested on synthetic as well as in-the-wild image datasets. In the synthetic image dataset, the accuracy of the proposed method for face reconstruction was first compared to previous studies, evaluating the reconstruction ability under different face pose angles. As for the in-the-wild image dataset, the face reconstruction ability was evaluated under real world conditions of makeup, blurring, occlusion and poor illumination.

4.1 Datasets and implementation detail

Synthetic image dataset The synthetic database used in the experiments was built up from the Basel Face Model (BFM) [28]. The BFM includes a shape model and an appearance model, the shape model consisting of 53,490

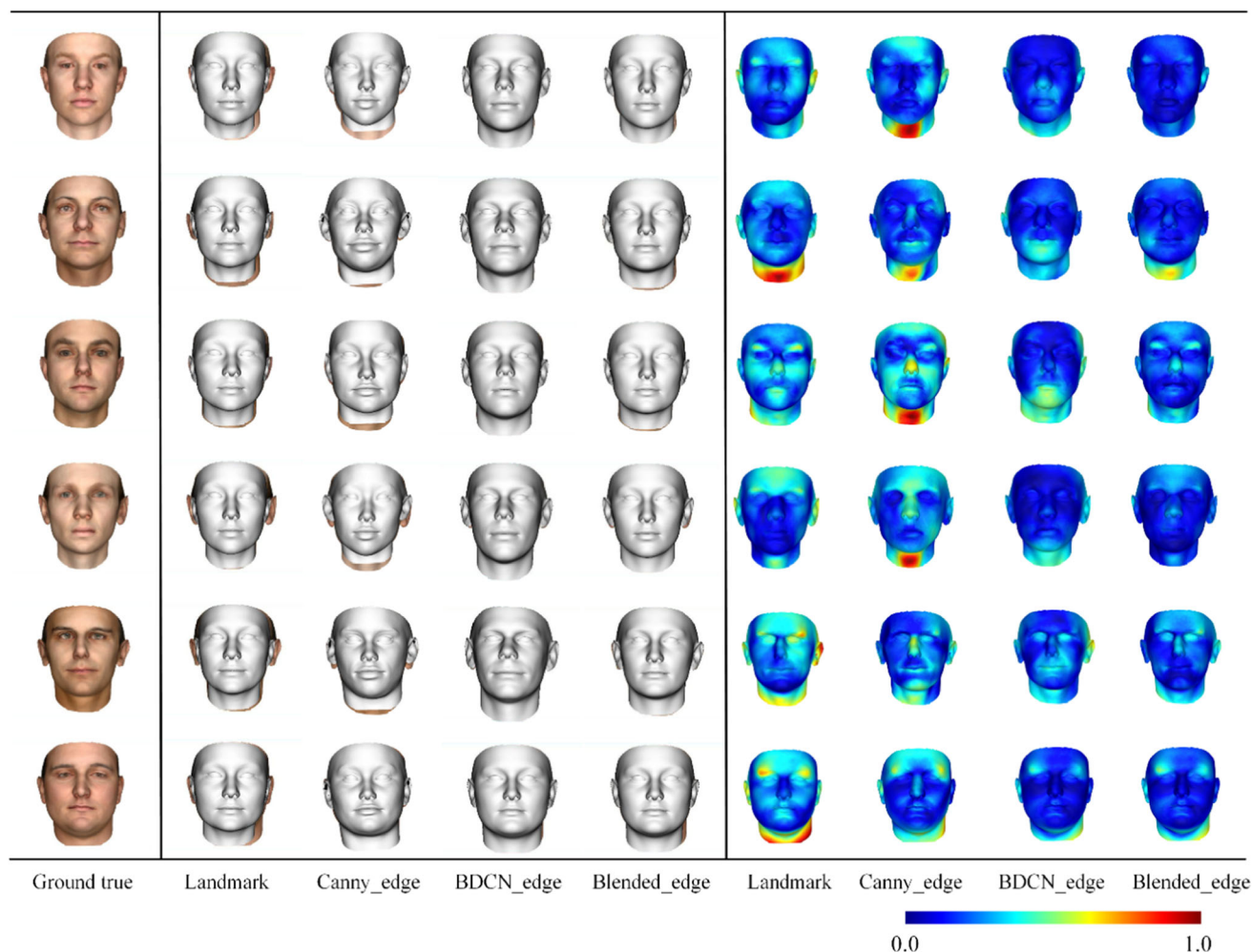


Fig. 7 Reconstructions using different methods and thermodynamic error images: column 1: input image; column 2 to 5: reconstructions from left to right using landmarks, hard Canny edges, BDCN edges

vertices, 159,955 edges and 106,466 faces. Each vertex is represented by a 3D coordinate $v_i = (x_i, y_i, z_i) \in R^3$. In the appearance model, albedo maps can be controlled to generate the RGB face texture map. By controlling the shape and texture parameters, different face identities can be generated. Thus, 100 random shape parameters and texture parameters were generated, which can be considered as 100 different identities.

The 100 different face model will be used to generate the synthetic image dataset. The pose parameters were separated at angular intervals of 10° . The pitch angles of the faces ranged from -80° to 80° , with a total of 17 pose values ($0^\circ, \pm 10^\circ, \pm 20^\circ, \pm 30^\circ, \pm 40^\circ, \pm 50^\circ, \pm 60^\circ, \pm 70^\circ, \pm 80^\circ$). Poses outside this range were not considered. For each of the 100 synthetic face models, the face model was rotated according to the pose parameters and then rendered with random background to generate synthetic 2D images with different poses by means of

and blended edge; column 6 to 9, relevant thermodynamic errors from left to right for results based on landmarks, hard Canny edges, BDCN edges and blended edges

orthographic projection, resulting in a synthetic dataset of 1700 face images paired with ground truth 3D face models. This synthetic dataset was used in the experiment for comprehensive evaluation across different methods. When comparing with other studies of similar nature in the literature, such as [34, 58], the dataset of such sample size is regarded as sufficient for evaluation purpose.

In-the-wild image dataset The facial in-the-wild WFLW dataset [19] was employed in the experiment. This dataset includes 10,000 face images gathered from the internet, with no constraints on the resolution, lighting conditions or camera parameters. Each image was labelled based on various attributes, such as occlusion and blurring, and these images are annotated with 98 ground truth facial landmarks. We sampled a subset of the data, with four attributes of occlusion, illumination, makeup, and blurring that are likely to be affected by edge detection and we used 100 faces for each attribute for evaluation purpose. This dataset

Table 1 Comparison of different reconstruction methods using average Euclidean vertex distance (mm) and ablation study on the effectiveness of different process along the proposed pipeline (marked as (a)–(e)) based on the synthetic dataset

Method	Ablation Study	Rotation angle									Mean
		0°	± 10°	± 20°	± 30°	± 40°	± 50°	± 60°	± 70°	± 80°	
Landmark [50]	(e)	2.47	2.57	2.40	2.31	2.23	2.49	2.59	2.85	3.20	2.57
Soft edges [29]		2.59	2.61	2.58	2.54	2.52	2.50	2.65	2.57	2.63	2.57
Soft LAB edges [19]	(a)	2.20	2.12	2.18	2.25	2.55	2.60	2.71	2.88	3.00	2.50
PRN [58]		2.33	–	–	–	2.13	–	–	–	2.52	2.32
DECA [59]		2.14	–	2.19	–	2.21	–	2.24	–	2.26	2.21
Deep3DFaceRecon [57]		1.81	–	–	–	1.82	–	–	–	2.22	1.95
Proposed (Hard LAB edges)	(b)	2.15	1.97	2.04	1.99	1.62	1.56	1.63	1.69	2.05	1.86
Hard Canny edges [32]	(c)	1.80	2.03	1.95	1.79	1.71	1.53	1.65	1.93	2.25	1.85
BDCN edges [20]		1.69	1.69	1.61	1.52	1.45	1.38	1.50	1.84	2.20	1.65
Proposed blended hard edges	(d)	1.70	1.72	1.58	1.48	1.42	1.42	1.54	1.75	2.10	1.64

Table 2 Comparison of different reconstruction methods using average normal vertex distance (mm) based on the synthetic dataset

Method	Rotation angle									Mean
	0°	± 10°	± 20°	± 30°	± 40°	± 50°	± 60°	± 70°	± 80°	
Landmark [50]	1.65	1.69	1.56	1.50	1.41	1.45	1.43	1.53	1.71	1.55
Hard Canny edges [32]	1.33	1.32	1.28	1.28	1.26	1.28	1.31	1.39	1.43	1.32
BDCN edges [20]	1.31	1.30	1.27	1.26	1.26	1.25	1.29	1.38	1.48	1.31
Proposed blended hard edges	1.29	1.29	1.24	1.22	1.22	1.23	1.26	1.31	1.38	1.27

was used to evaluate the performance and robustness of the proposed method in different image conditions.

Implementation Detail The LAB network [19] pretrained using WFLW dataset for face alignment was used to extract facial edge heatmap features from the intermedia layers. The extracted features were processed into hard-blended edges using the proposed algorithm. We then used the proposed method for 3D face shape reconstruction. Our algorithm was implemented in the Ubuntu 20 environment with NVIDIA GeForce RTX 2070, using Caffe and MATLAB.

4.2 Evaluation metrics

4.2.1 Procrustes analysis

For a comprehensive evaluation of the proposed method, Procrustes analysis (PA) was first used to align and normalize two meshes before calculating the point-to-point Euclidean distance. The PA is widely used in performance evaluation [54] because the 3D shapes reconstructed from various methods may likely exhibit diverse scales,

orientations, and centroid positions, the mesh models must be aligned and normalized for comparison purpose. The PA method, proposed by [55], is a statistical shape analysis to transform a source set of points to a target set, minimizing the Procrustes distance through scaling and translation as follows:

$$d(\hat{V}, \hat{V}^{gt}) = |\hat{V} - \hat{V}^{gt}| \quad (11a)$$

$$\hat{V} = \frac{1}{\sigma(V)}(V - \bar{V}) \quad (11b)$$

$$\hat{V}^{gt} = \frac{1}{\sigma(V^{gt})}(V^{gt} - \bar{V}^{gt}) \quad (11c)$$

\hat{V}^{gt} and \hat{V} represent the *normalized* ground truth model and the normalized predicted model, and PA is done by minimizing the Procrustes distance, $d(\cdot)$, between the two models of Eq. (11a). $|\cdot|$ represents the mean vertex Euclidean distance. The respective model normalization is calculated by Eqs. (11b) and (11c), respectively, where V and V^{gt} represent vertices of the predicted face model and the ground truth face model, while \bar{V} and \bar{V}^{gt} represent the centre points of the predicted and ground truth models,

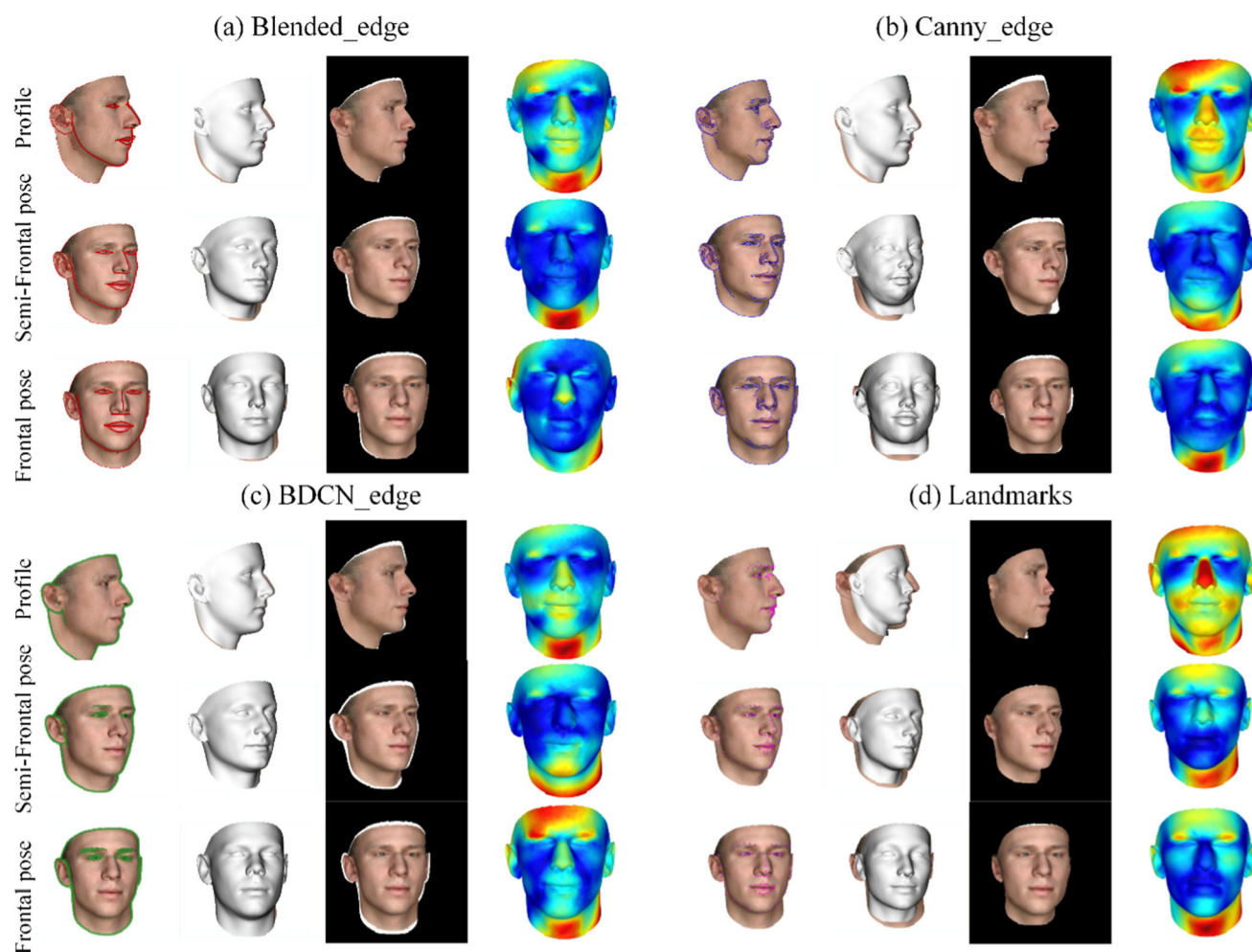


Fig. 8 Detail comparison of face reconstructions from different pose angles by different methods: **a** the proposed hard-blended edge feature-based method, **b** method based Canny edge features [32], **c** method based on BDCN edge features [20], and **d** method purely based on landmark features. In each method, the first column is input image overlaid with detected features from different pose angles for profile

angle at the top row, semi-frontal angle in the middle and frontal pose angle at the bottom, the second column present reconstructed face shape, the third column present the contour difference between the reconstructed face and ground-truth face shape, the last column present thermodynamic diagram visualizing vertex-to-vertex distance between reconstructed face and ground-truth face

respectively. $\sigma(\cdot)$ represents the calculation of standard variation.

4.2.2 Evaluation metrics for the synthetic dataset

Euclidean distance metric After PA, the performance of different methods can be evaluated by calculating the mean Euclidean distance between paired vertices of the ground truth and the predicted mesh models as follows.

$$E^{Eud} = \sqrt{\sum_{i=1}^N (v_i - v'_i)^2} \quad (12)$$

where $v_i \in \hat{V}$ represents a normalized vertex v_i of the predicted model while $v'_i \in \hat{V}^{gt}$ represents the corresponding vertex of ground truth model normalized by PA. This

metric was proposed by Piotraschke and Blanz [56] and is commonly used in 3D reconstruction evaluation.

Considering that the Euclidean distance-based metric of Eq. (12) is sensitive to the accuracy of face alignment and scale, *normal distance metric*, also proposed by Piotraschke and Blanz [56], measures the shape difference by further eliminating scale interference.

$$E^{normal} = \frac{1}{N} \sum_{i=1}^N \arccos \frac{n_i \cdot n'_i}{\|n_i\| \cdot \|n'_i\|} \quad (13)$$

where n_i represents the normal vector of vertex $v_i \in \hat{V}$, and $\|\cdot\|$ represents the modulus of the vector. As shown in Eq. (13), this metric measures the cosine distance between the normal vectors of two vertices without the influence of the scale factor or rotation. The normal distance metric

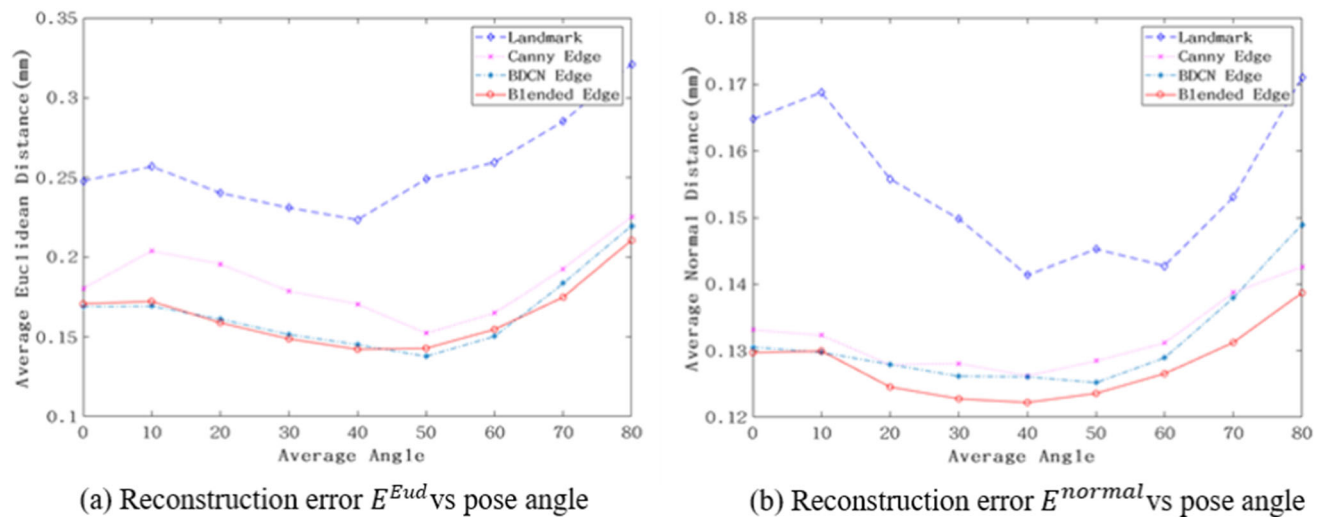


Fig. 9 Comparison of the reconstruction errors, in terms of **a** average Euclidean distance and **b** average normal distance, at different pose angles using different methods: method purely based on landmark

features [50], method based on Canny edge features [32], method based on BDCN edge features [20], and the proposed method using hard-blended edge features

therefore can measure shape-related differences. For evaluation on the synthetic dataset, this normal distance metric is used to evaluate the differences in shape between the synthesized models (ground truths) and the predicted models.

4.2.3 Evaluation metrics for the in-the-wild image dataset

There is no ground truth 3D face model for the in-the-wild dataset, but ground truth landmarks. Thus, some work [32] employed mean Euclidean vertex of landmark as substitute. In the evaluation process, the distances between the ground truth landmarks and the contour generated from the estimated face model are calculated using the KNN-search algorithm. The evaluation metric is given as:

$$E^{edge} = \frac{1}{N} \sum_{i \in N} \left| \hat{x}_i - \text{Knnsearch} \left(sRv_j^{cont} + t \right) \right| \quad (14)$$

where $\hat{x}_i, i \in N$ are the coordinates of the N ground truth landmarks, $v_j^{cont}, j \in C$ are the vertices on the occluding contour C , and $\text{KNNsearch}(\bullet)$ represents the KNN search operation.

4.3 Evaluation on the synthetic images

4.3.1 Feature analysis

The facial features, detected by the application of different methods on the synthetic image dataset, are shown in Fig. 6. The Canny edge detector generates a lot of noise in the synthetic images. Compared with the Canny edges, the LAB edges contain significantly less noise. The proposed

hard-blended edges not only contain less noise but also provide more geometric information.

4.3.2 Reconstruction analysis

The Lambertian renderings and thermodynamic images of four experiments are shown in Fig. 7, which visualize and compare the reconstruction ability of different methods qualitatively through six samples selected from 100 synthetic face models. The Lambertian renderings are shown from column 2 to 5. The thermodynamic images from column 6 to 9 represent the per-vertex Euclidean distance between the predicted face model and the ground truth. The colour changing from dark blue to red indicates the distance increasing from small to large. For the landmark-based method, column 6 represents the thermodynamic diagrams of the results generated from the landmark-based methods, showing larger errors than those of the hard-blended edge and BDCN methods. The Lambertian renderings in column 2 show that the ears and neck are not well fitted due to the lack of points in the relevant regions. The thermodynamic images in column 8 show that the reconstruction errors of the BDCN method are relatively small. The BDCN method can fit the contours of the head well (as shown in column 4), although the reconstruction of the eyes and nose lack details compared to the hard-blended edge method. The reconstructions from the hard-blended edge method contain less noise and more detail than the other methods, as shown in column 5. The thermodynamic images in column 9 show that the errors of reconstructions from the hard-blended edges are small.

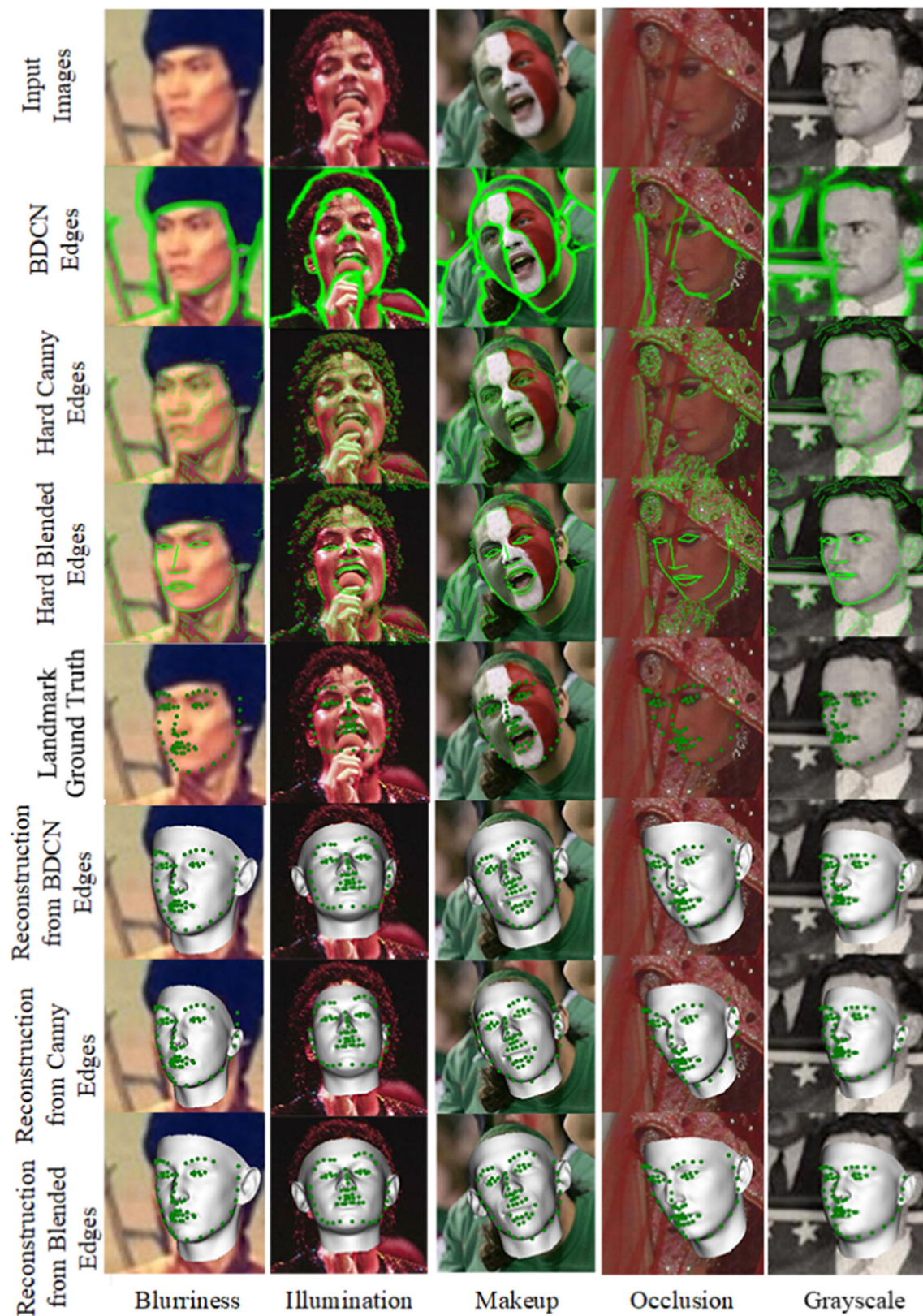


Fig. 10 Qualitative results for the in-the-wild dataset: from left to right: blurriness, illumination, makeup, occlusion and grayscale. From top to bottom: input images, reconstructions based on hard Canny edges and hard-blended edges

Table 3 Comparison of different reconstruction methods using normalized mean error (NME) based on the in-the-wild dataset

Method	Blurring	Illumination	Makeup	Occlusion	Greyscale	Mean
Hard Canny edges [32]	22.29	22.57	27.37	27.31	20.27	23.96
BDCN edges [20]	21.74	21.30	22.38	27.27	19.82	22.50
Proposed (hard LAB edges)	21.89	21.96	21.96	27.02	19.61	22.49
Proposed (hard-blended edges)	21.72	21.26	21.16	26.74	19.55	22.08

4.3.3 Quantitative analysis

Table 1 presents a comparison based on the per-vertex Euclidean distance metrics (Eq. (11a)) among different methods, including a landmark-based method [50] and edge-based methods such as those reconstructed from soft edges [29], Canny edges [32], and BDCN edges [20], which is a deep learning based method for edge detection. Moreover, it also compares with three recent deep-learning based methods [57–59] for face reconstruction directly from images. Romdhani and Vetter [29] proposed a soft edge cost function and determined the influence of an edge based on distance. This method leads to a wider radius of convergence and robustness, but the precision is limited due to local minima problem. The results show that hard edge features are more effective than soft edges in 3DMM reconstruction, which is similar to the findings of [32]. Recently, new deep learning-based methods were proposed for 3D face reconstruction [57] and facial texture reconstruction [58] using the Basel Face Model. Feng et al. [58] trained a neural network to predict the pose and shape parameters of BFM. Feng et al. [59] proposed a network to predict shape and expression based on TF-Flame model. Deng et al. [57] incorporated differentiable rendering into a deep neural network to reconstruct face shape and texture. However, these methods do not perform well under large pose conditions because they employed 3D landmarks for alignment and face shape reconstruction due to the inherent limitations of 3D landmarks strategy, as illustrated in Fig. 4. Table 1 shows that the proposed method outperforms all other methods, in terms of Euclidean vertex distances (Eq. (12)), using the synthetic dataset. The best results are highlighted in bold in the table and hereafter in other tables.

Table 2 compares different methods by means of the per-vertex normal distance metric (Eq. (13)) and shows that the pure landmark-based method gives a larger distance than the edge-based methods. On the normal distance metric, the present method using hard-blended edges gave a better accuracy than the two methods involving hard Canny edges [32] and BDCN edges [20], respectively, indicating that the hard-blended edges method produce a better performance on the basis of the reconstructed shapes.

4.3.4 Evaluation of results generated with different pose angles

As previously reported [33], although the contour of a face model may be closely fitted to the edge map, there may still be a significant difference between the ground truth and predicted models. Figure 8 compares the results of face reconstructions using the landmarks [50], hard Canny edges [32] and BDCN-based [20] methods and the proposed blended hard edge-based method with different pose angles.

For frontal and semi-frontal poses reconstructions, the cheek contour area exhibits high accuracy, whereas the forehead and nose areas show relatively large errors. For profile view reconstructions, the cheek areas exhibit larger errors than the other poses. Furthermore, the landmark-based reconstruction method exhibits larger errors than the edge-based methods.

The reconstructions using different methods based on images with different pose angles are compared in Fig. 9, and it is shown that the best per-vertex accuracy occurred on a range of semi-frontal poses between 30° and 50°, both for the Euclidean and normal distance metrics. Among the different methods, the landmark-based method is obvious inferior to edge-related methods, and the hard-blended edge method (red line) exhibits the highest accuracy.

The per-vertex shape accuracy for the face models generated from the proposed model at different pose angles was explored. In general, the mean error distribution indicated larger reconstruction errors for greater pose angles for all the methods. With the frontal pose, the contours fit the model well, although the errors are not the smallest. A semi-frontal pose gives a smaller Euclidean error distribution for most of the methods. These effects have not been fully considered or explored in previous studies. For example, He et al. [56] combined reconstructions from different view angles to achieve greater accuracy. In another study [60], a recurrent neural network was used to carry out reconstruction in different poses and then combine these reconstructions. However, the reconstructions in the different poses were all treated equally.

4.4 Ablation study

An ablation study was conducted and the results are also shown in Table 1, in which the impacts of different processes along the proposed pipeline on the reconstruction results are compared. The first row of Table 1 represents face reconstruction without using any edge features but solely based on landmark features (labelled as (e) in Fig. 1 and Table 1). The row marked with (a) in Table 1 depicts results obtained by initializing with landmark features (i.e. (e)) and subsequently incorporating unprocessed LAB edge features (with reference to (a) in Figs. 1 and 3). From the results, it is evident that the inclusion of unprocessed soft lab features does not significantly enhance the reconstruction results. In the row marked with (b) in the table, the lab features are converted into hard edges (feature labelled as (b) in Figs. 1 and 3) being used together with landmark features, resulting in a notable improvement in the reconstruction performance. The last row of the table, marked as row (d), depicts outcomes from the proposed hard-blended edge by combining the processed hard lab edges (b) with Canny edge (c) features, demonstrating further enhancement on the reconstruction results.

4.5 Evaluation on the in-the-wild images

In this section, the hard-blended edge method was applied to the in-the-wild dataset in comparison with hard Canny edge method [32] and BDCN edge method [20]. The qualitative and quantitative analyses are discussed in the following sections.

4.5.1 Qualitative analysis

The qualitative results of the face reconstructions generated by the proposed hard-blended edge method using in-the-wild database are illustrated in Fig. 10. Images annotated as containing makeup, blurring, occlusion, with low illumination, and in greyscale were specifically selected.

Under blurring and illumination conditions, the Canny edge detector is less effective and can hardly detect the face shapes. In contrast, the BDCN edges can effectively extract the face contour in the images, but details of eyes and mouth are missing. Comparatively, the proposed method can detect both the global shape and local details of face in these images. As for the makeup and occlusion conditions, a great deal of noise is produced by the Canny edge detectors as well as BDCN edge detection neural network, the sixth and seventh rows showing that the reconstructed face models can be aligned to the wrong edges under such conditions. The reconstructed 3D face images using the proposed method are shown in the eighth

row and illustrate that they better fit to the input images. The ground-truth landmarks of the input images (the fifth row) are also overlaid on the constructed faces output from the corresponding methods (in the sixth to eighth rows) for illustration purpose. In sum, the proposed hard-blended edge method can obtain cleaner and less noisy edges and is more robust to in-the-wild images, comparing to traditional Canny edge-based method and deep learning BDCN edge-based method. When inputs are black-and-white images, our result show that it will not deteriorate the performance of all the compared methods, this is because extraction of landmarks and Canny edges involves transforming the images into binary space, while black-and-white images are included in the training dataset of the deep learning-based methods, including the face alignment LAB framework [19] of our proposed method. To conclude, the main factors that affect the processing of landmarks and edges are the blur and illumination levels of the images and the impact of occlusions.

4.5.2 Quantitative analysis

The results of the quantitative evaluation of the in-the-wild image dataset [19] are shown in Table 3. The normalized mean error (NME) was used as a metric Eq. (14) for the distance between the ground truth landmarks and their corresponding vertices from the predicted 3D model. Analyses were also carried out on images from the WFLW dataset with attributes of blurring, low illumination, makeup, occlusion, and greyscale (Table 3). The results show that the blended edge method performed more robustly under conditions of blurring, poor illumination and occlusion than the other three methods.

5 Conclusion

In this study, an edge-feature-based face reconstruction with hard-blended edge features generated from deep neural network is proposed. Edge-based 3D face reconstruction is an approach to reconstruct 3D face shapes using edge features in addition to facial landmarks, taking advantage of more geometric information of the target 3D shapes that can be provided from edge features of the input images. Nevertheless, existing methods for feature extraction have a known drawback that the extracted edges are often noisy, in particular when input images are blurry, or have problems such as illumination, makeup, and occlusion. For improving the quality of edge features for face reconstruction, a new method is developed to extract hard-blended edge features. The use of hard-blended edges can therefore improve contour-based face reconstruction, achieving a large degree of accuracy on a synthetic dataset.

Furthermore, those edge features are robust when applying to in-the-wild images, even under blurring, poor illumination, makeup, and occlusion conditions.

The quantitative results showed that reconstructions based upon semi-frontal poses achieved the greatest accuracy in general, and that the accuracy of local regions was largely affected by pose angles. Reconstructions based upon the frontal pose were more accurate for the cheek contour region, whereas those of the forehead, neck and mouth regions showed less accuracy and greater divergence. When the pose was large, the global accuracy of the face was lower. This effect is believed to be related to the capability of the parametric face model as well as the diversity of training data. The proposed method has been applied in customizing 3D avatars for fashion presentation using augmented reality technology in a mobile app.

The current method has a few limitations. First of all, comparing with other edge or landmark-based reconstruction methods, the current approach uses deep neural network for edge feature extraction, which are then used to optimize 3DMM parameters. It is not end-to-end process and is computationally expensive; the current face reconstruction adopts a server-based implementation. In the future, lightweight implementation for real-time application directly on terminal devices like smartphones will be explored. On the other hand, the current study mainly focused on face shape reconstruction from images of large pose angles, heavy makeup, poor illumination, blurriness, and occlusion, without considering facial expressions. This aspect remains a subject for future research endeavours. Lastly, the current method reconstructs face shapes from image input, without considering video data. Video-based reconstruction will be a key agenda for future research efforts.

Acknowledgements The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Grant Number 152112/19E). This work was also partially supported by the Laboratory for Artificial Intelligence in Design (Project Code: RP1-1), InnoHK Research Cluster, Hong Kong Special Administrative Region.

Funding Open access funding provided by The Hong Kong Polytechnic University.

Data availability The authors declare that the data supporting the findings of this study are available within the paper. The code will be released once the paper is published.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing,

adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Fan X, Cheng S, Huyan K, Hou M, Liu R, Luo Z (2020) Dual neural networks coupling data regression with explicit priors for monocular 3D face reconstruction. *IEEE Trans Multimed* 23:1252–1263
2. Li T, Bolkart T, Black MJ, Li H, Romero J (2017) Learning a model of facial shape and expression from 4D scans. *ACM Trans Graph* 36(6):194:1–194:17
3. Patel A, Smith WA (2012) Driving 3D morphable models using shading cues. *Pattern Recognit* 45(5):1993–2004
4. Amberg B, Blake A, Fitzgibbon A, Romdhani S, Vetter T (2007) Reconstructing high quality face-surfaces using model based stereo. In: 2007 IEEE 11th international conference on computer vision. IEEE
5. Blanz V, Vetter T (2023) A morphable model for the synthesis of 3D faces. In *Seminal Graphics Papers: Pushing the Boundaries*, Volume 2 (pp. 157–164)
6. Huber P, Feng Z-H, Christmas W, Kittler J, Rätsch M (2015) Fitting 3d morphable face models using local features. In: 2015 IEEE international conference on image processing (ICIP). IEEE
7. Egger B, Smith WA, Tewari A, Wuhler S, Zollhoefer M, Beeler T, Bernard F, Bolkart T, Kortylewski A, Romdhani S (2020) 3d morphable face models—past, present, and future. *ACM Trans Graph (TOG)* 39(5):1–38
8. Hassner T, Harel S, Paz E, Enbar R (2015) Effective face frontalization in unconstrained images. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4295–4304)
9. Lee YJ, Lee SJ, Park KR, Jo J, Kim J (2012) Single view-based 3D face reconstruction robust to self-occlusion. *EURASIP J Adv Signal Process* 2012:1–20
10. Qu C, Monari E, Schuchert T, Beyerer J (2014) Fast, robust and automatic 3D face model reconstruction from videos. In: 2014 11th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE
11. Asthana A, Zafeiriou S, Cheng S, Pantic M (2013) Robust discriminative response map fitting with constrained local models. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3444–3451)
12. Zhu X, Lei Z, Yan J, Yi D, Li SZ (2015) High-fidelity pose and expression normalization for face recognition in the wild. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 787–796)
13. Thies J, Zollhöfer M, Nießner M, Valgaerts L, Stamminger M, Theobalt C (2015) Real-time expression transfer for facial reenactment. *ACM Trans Graph* 34(6):183:1–183:14
14. Cao C, Chai M, Woodford O, Luo L (2018) Stabilized real-time face tracking via a learned dynamic rigidity prior. *ACM Trans Graph (TOG)* 37(6):1–11

15. Wang J, Lin J, Yu Q, Liu R, Chen Y, Yu SX (2022) 3d shape reconstruction from free-hand sketches. In: European conference on computer vision. Springer
16. Delanoy J, Aubry M, Isola P, Efros AA, Bousseau A (2018) 3d sketching using multi-view deep volumetric prediction. In: Proceedings of the ACM on computer graphics and interactive techniques, vol 1, no 1, pp 1–22
17. Yu Z, Feng C, Liu MY, Ramalingam S (2017) CASENet: deep category-aware semantic edge detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5964–5973)
18. Shen W, Wang B, Jiang Y, Wang Y, Yuille A (2017) Multi-stage multi-recursive-input fully convolutional networks for neuronal boundary detection. In: Proceedings of the IEEE international conference on computer vision (pp. 2391–2400)
19. Wu W, Qian C, Yang S, Wang Q, Cai Y, Zhou Q (2018) Look at boundary: a boundary-aware face alignment algorithm. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2129–2138)
20. He J, Zhang S, Yang M, Shan Y, Huang T (2019) Bi-directional cascade network for perceptual edge detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3828–3837)
21. Schönborn S, Forster A, Egger B, Vetter T (2013) A Monte Carlo strategy to integrate detection and model-based face analysis. In: Pattern recognition: 35th German conference, GCPR 2013, Saarbrücken, Germany, September 3–6, 2013. Proceedings 35. Springer
22. Breuer P, Kim K-I, Kienle W, Scholkopf B, Blanz V (2008) Automatic 3D face reconstruction from single images or video. In: 2008 8th IEEE international conference on automatic face & gesture recognition. IEEE
23. Aldrian O, Smith WA (2012) Inverse rendering of faces with a 3D morphable model. *IEEE Trans Pattern Anal Mach Intell* 35(5):1080–1093
24. Lv Z (2020) Robust3D: a robust 3D face reconstruction application. *Neural Comput Appl* 32(13):8893–8900
25. Wood E, Baltrušaitis T, Hewitt C, Johnson M, Shen J, Milosavljević N, Wilde D, Garbin S, Sharp T, Stojiljković I (2022) 3d face reconstruction with dense landmarks. In: European conference on computer vision. Springer
26. Cootes TF, Taylor CJ, Cooper DH, Graham J (1995) Active shape models-their training and application. *Comput Vis Image Underst* 61(1):38–59
27. Moghaddam B, Lee J, Pfister H, Machiraju R (2003) Model-based 3D face capture with shape-from-silhouettes. In: 2003 IEEE international SOI conference. Proceedings (Cat. No. 03CH37443). IEEE
28. Paysan P, Knothe R, Amberg B, Romdhani S, Vetter T (2009) A 3D face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. IEEE
29. Romdhani S, Vetter T (2005) Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). IEEE
30. Fitzgibbon AW (2003) Robust registration of 2D and 3D point sets. *Image Vis Comput* 21(13–14):1145–1153
31. Langari B, Vaseghi S, Prochazka A, Vaziri B, Aria FT (2016) Edge-guided image gap interpolation using multi-scale transformation. *IEEE Trans Image Process* 25(9):4394–4405
32. Bas A, Smith WA, Bolkart T, Wuhler S (2016) Fitting a 3D morphable model to edges: a comparison between hard and soft correspondences. In: Asian conference on computer vision. Springer
33. Keller M, Knothe R, Vetter T (2007) 3D reconstruction of human faces from occluding contours. In: International conference on computer vision/computer graphics collaboration techniques and applications. Springer
34. Suwajanakorn S, Kemelmacher-Shlizerman I, Seitz SM (2014) Total moving face reconstruction. In: European conference on computer vision. Springer
35. Khan A, Hayat S, Ahmad M, Cao J, Tahir MF, Ullah A, Javed MS (2021) Learning-detailed 3D face reconstruction based on convolutional neural networks from a single image. *Neural Comput Appl* 33:5951–5964
36. Karmouni H, Yamni M, El Ogri O, Daoui A, Sayyouri M, Qjidaa H, Tahiri A, Maaroufi M, Alami B (2021) Fast computation of 3D discrete invariant moments based on 3D cuboid for 3D image classification. *Circuits Syst Signal Process* 40:3782–3812
37. El Ogri O, Daoui A, Yamni M, Karmouni H, Sayyouri M, Qjidaa H (2020) New set of fractional-order generalized Laguerre moment invariants for pattern recognition. *Multimed Tools Appl* 79:23261–23294
38. Daoui A, Karmouni H, Sayyouri M, Qjidaa H (2022) Efficient methods for signal processing using Charlier moments and artificial bee Colony algorithm. *Circuits Syst Signal Process* 41(1):166–195
39. Yamni M, Daoui A, Karmouni H, Sayyouri M, Qjidaa H (2019) Influence of Krawtchouk and Charlier moment's parameters on image reconstruction and classification. *Procedia Comput Sci* 148:418–427
40. Tahiri MA, Bencherqui A, Karmouni H, Jamil MO, Sayyouri M, Qjidaa H (2022) Optimal 3D object reconstruction and classification by separable moments via the Firefly algorithm. In: 2022 International conference on intelligent systems and computer vision (ISCV). IEEE
41. Karmouni H, Yamni M, El Ogri O, Daoui A, Sayyouri M, Qjidaa H (2020) Fast computation of 3D Meixner's invariant moments using 3D image cuboid representation for 3D image classification. *Multimed Tools Appl* 79:29121–29144
42. El Ogri O, Karmouni H, Sayyouri M, Qjidaa H (2021) 3D image recognition using new set of fractional-order Legendre moments and deep neural networks. *Signal Process Image Commun* 98:116410
43. Zhu X, Lei Z, Liu X, Shi H, Li SZ (2016) Face alignment across large poses: a 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 146–155)
44. Shang J, Shen T, Li S, Zhou L, Zhen M, Fang T, Quan L (2020) Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In: European conference on computer vision. Springer
45. Lv J, Shao X, Xing J, Cheng C, Zhou X (2017) A deep regression architecture with two-stage re-initialization for high performance facial landmark detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3317–3326)
46. Sharma S, Kumar V (2022) 3D face reconstruction in deep learning era: a survey. *Arch Comput Methods Eng* 29(5):3475–3507
47. Shen W, Wang X, Wang Y, Bai X, Zhang Z (2015) Deepcontour: a deep convolutional feature learned by positive-sharing loss for contour detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3982–3991)
48. Newell A, Yang K, Deng J (2016) Stacked hourglass networks for human pose estimation. In: European conference on computer vision. Springer
49. Yang J, Liu Q, Zhang K (2017) Stacked hourglass network for robust facial landmark localisation. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 79–87)

50. Bulat A, Tzimiropoulos G (2017) How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In: Proceedings of the IEEE international conference on computer vision (pp. 1021–1030)
51. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process Lett* 23(10):1499–1503
52. Zhu X, Ramanan D (2012) Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE
53. Deng J, Trigeorgis G, Zhou Y, Zafeiriou S (2019) Joint multi-view face alignment in the wild. *IEEE Trans Image Process* 28(7):3636–3648
54. Zhang J, Luximon Y, Wan J, Li P (2023) Capture my head: a convenient and accessible approach combining 3D shape reconstruction and size measurement from 2D images for headwear design. *Comput Aided Des* 159:103487
55. Gower JC (1975) Generalized procrustes analysis. *Psychometrika* 40:33–51
56. Pietraschke M, Blanz V (2016) Automated 3d face reconstruction from multiple images using quality measures. In: Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3418–3427)
57. Deng Y, Yang J, Xu S, Chen D, Jia Y, Tong X (2019) Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (pp. 1–11)
58. Feng Y, Wu F, Shao X, Wang Y, Zhou X (2018) 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)
59. Feng Y, Feng H, Black MJ, Bolkart T (2021) Learning an animatable detailed 3D face model from in-the-wild images. *ACM Trans Graph (TOG)* 40(4):1–13
60. Wu F, Bao L, Chen Y, Ling Y, Song Y, Li S, Ngan KN, Liu W (2019) MVF-Net: multi-view 3d face morphable model regression. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 959–968)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.