Learning 3D Face Reconstruction from a Single Sketch⋆

Li Yang[a], Jing Wu[b], Jing Huo[a,*], Yu-Kun Lai[b] and Yang Gao[a]

[a]State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[b]School of Computer Science & Informatics, Cardiff University, Cardiff, Wales, UK

ARTICLE INFO

Keywords:
3D Face Reconstruction
Sketch
Characteristic Details

ABSTRACT

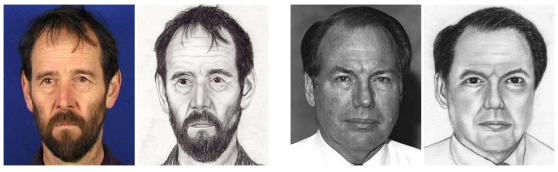
3D face reconstruction from a single image is a classic computer vision problem with many applications. However, most works achieve reconstruction from face photos, and little attention has been paid to reconstruction from other portrait forms. In this paper, we propose a learning-based approach to reconstruct a 3D face from a single face sketch. To overcome the problem of no paired sketch-3D data for supervised learning, we introduce a photo-to-sketch synthesis technique to obtain paired training data, and propose a dual-path architecture to achieve synergistic 3D reconstruction from both sketches and photos. We further propose a novel line loss function to refine the reconstruction with characteristic details depicted by lines in sketches well preserved. Our method outperforms the state-of-the-art 3D face reconstruction approaches in terms of reconstruction from face sketches. We also demonstrate the use of our method for easy editing of details on 3D face models.

1. Introduction

3D face reconstruction from a single image is a classic computer vision problem and has many applications in face recognition, animation, etc. There have been great advances from traditional methods [3, 2, 26, 21] to more recent deep learning-based methods [14, 9, 39] in this area. However, most works in this area achieve reconstruction from face photos. Reconstruction from other portrait forms, such as face sketches, has not received much attention. As depicted in Figure 1, while the photos record the facial appearance with higher fidelity from realistic shadings, the sketches are more subjective recordings with emphasized characteristic features depicted by lines, e.g. the emphasized wrinkles. Therefore, learning reconstruction from sketches is interesting and may help recover more details of faces in some cases. However, as we will later show in Section 4.2, methods for 3D reconstruction from face photos are not directly applicable to reconstruction from sketches. Based on these observations, in this work, we propose a method for 3D reconstruction from face sketches, the aim of which is to reconstruct a 3D face with a faithful overall shape and characteristic features preserved or even enhanced.

Specifically, we propose a two-stage method where the first stage uses information of photos and sketches to obtain a coarse reconstruction result and the second stage further uses depicted lines to refine the results. In the first stage, a deep learning framework is used for the proposed method, both inspired by the success of deep learning in 3D reconstruction from face photos, and considering the limitations of traditional methods on reconstruction from silhouettes or contours. However, to build a deep learning based 3D face reconstruction model, the difficulty is the lack of paired face sketches and 3D data, which is essential for learning the reconstruction in a supervised way. Overcoming this problem motivated the first idea in our work, i.e., integrating the knowledge of reconstruction from face photos. There are many data resources available for 3D reconstruction from face photos [32, 14], which provide a large amount of paired photo-3D data for deep learning-based approaches. On the other hand, in the field of sketch synthesis and recognition, sketch data are also available [35, 40]. We relate the two by proposing a dual-path architecture called DP-CoarseNet which follows the idea in [17] and decomposes the two portrait forms into their domain-specific style codes, and shared domain-invariant content codes from which the 3D shape is reconstructed. The demand for paired sketch-3D training data is thus mitigated by synergy of photo-to-sketch synthesis and 3D reconstruction from face photos using the training data available in each domain respectively. Comparing the reconstruction results of the proposed network and several baseline networks, we demonstrate the ability of the dual-path network to reconstruct more convincing overall face shapes from sketches.

A FineNet then follows to enhance the reconstructed coarse model with characteristic details. In reconstruction from pho-



(a) (b)

Figure 1: Examples of face photos and sketches from (a) the CUHK Face Sketch Database (CUFS) [35], (b) the CUHK Face Sketch FERET Database (CUFSF) [35, 40]. Sketch records more characteristic features of the face.

⋆This work is supported by Science and Technology Innovation 2030-"New Generation Artificial Intelligence" Major Project (No.: 2018AAA0100905), the National Natural Science Foundation of China (Granted No.: 61806092), the Natural Science Foundation of Jiangsu Province (Granted No.: BK20180326) and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

*Corresponding author
yangl@smail.nju.edu.cn (L. Yang); wuj11@cardiff.ac.uk (J. Wu); huojing@nju.edu.cn (J. Huo); laiy4@cardiff.ac.uk (Y. Lai); gaoy@nju.edu.cn (Y. Gao)

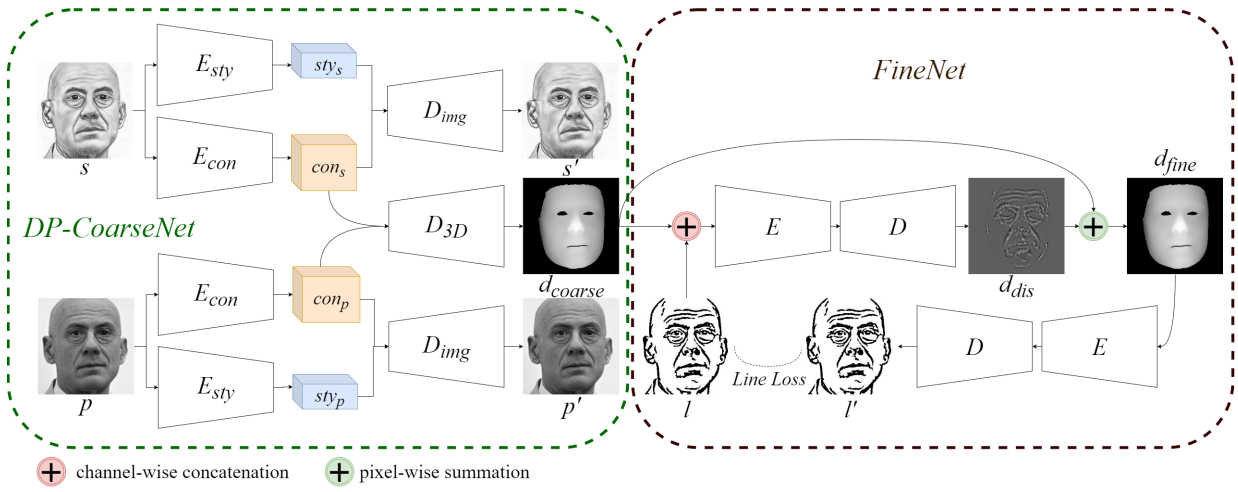


Figure 2: The overall architecture of our proposed network. The DP-CoarseNet decomposes each portrait form into their domain-specific style codes and shared content codes, and constructs the coarse 3D shape from the latter. The FineNet further refines the coarse shape by matching the lines extracted from the fine 3D shape with the characteristic lines extracted from the input sketch.

tos, the fine geometric details are captured by making use of shading variations [21]. However, shading is either missing or inaccurate in sketches, making this refine-from-shading approach inappropriate for reconstruction from sketches. Instead, sketched faces exhibit their characteristic features through lines, as shown in Figure 1. This motivated the second idea in our work, i.e., to refine the coarse model such that the reconstructed fine model reflects the characteristic lines depicted in the sketch. We extract these characteristic lines from sketches using a method based on XDoG [36], and develop a CNN-based module to simulate the same line extraction operation on reconstructed face models. We then integrate this module into the FineNet, and devise a novel line loss function. Minimization of the line loss guides the refinement of the coarse model to preserve the characteristic features in sketches. As this module is built on line drawings, we demonstrate that it is also effective for editing details by adding or removing lines on the extracted line drawing of a portrait.

The overall architecture of the proposed network is shown in Figure 2. We evaluate our method on both real sketches and synthetic sketches. The reconstruction results are compared with those achieved from the face photos [43, 10, 9, 39] and from the hand-drawn line sketches [15]. We also demonstrate the use of our method for easy editing of details on reconstructed face models.

In summary, the contributions of our work are as follows:

- We develop a deep neural network for 3D face reconstruction from sketches, integrating the knowledge of reconstruction from face photos.
- We design a novel line loss function and demonstrate its ability to preserve the characteristic details in face sketches during the reconstruction process.
- We demonstrate that our method can be used for easy editing of 3D face models with characteristic details added or removed.

2. Related Work

3D Face Modeling from Photos. 3D face reconstruction from a single face photo is a topic that has been studied extensively, from traditional methods [3, 21] to more recent deep learning-based methods [9, 39]. 3D morphable model (3DMM) [3] is a popular approach for this purpose. Methods based on 3DMM [4, 27, 28, 1] are guided by the top-down process in the perception of facial shapes, i.e., making use of facial priors. 3DMM constructs a parametric model from a set of example faces, which ensures the validity of the reconstructed faces, but the low-dimensional parametric representation often fails to capture fine geometric details. To address this limitation, Bas et al. [1] refine the 3DMM reconstruction by fitting the model to edges with hard correspondence. Their work shares some similarity with ours by taking into account facial features and characteristics for 3D reconstruction, although their method reconstructs 3D faces from photos rather than sketches. Compared to 3DMM-based methods, shape from shading (SFS)-based methods [21] follow the bottom-up process, recovering 3D shapes from the shading variations presented at the pixel level. Although pure SFS methods have inherent ambiguities, they are good at capturing fine geometric details such as wrinkles. More recent methods [28, 19] combine the two approaches by first reconstructing a coarse 3D face model based on 3DMM and then refining the model based on SFS. In the last years, deep learning, especially CNN-based methods, have become more and more popular in 3D face reconstruction [43, 10, 9, 39, 14]. These methods can reconstruct coarse 3DMM model parameters [43, 9], UV position maps [10], as well as coarse models plus fine details [14, 39]. Specifically, the training data of [10] is generated by 3DMM, and [39] directly generates depth maps based on SFS. Compared to the traditional 3DMM or SFS approaches, deep learning-based methods have the advantage of reconstruction time efficiency which

makes them suitable for real-time reconstruction from videos [14]. Our work is inspired by the success of deep learning for reconstruction from photos, and follows the coarse-to-fine multiple stage reconstruction as in [14, 39].

3D Face Modeling from Other Portrait Forms. In contrast to the large number of works in 3D face modeling from photos, modeling from other forms of portraits has not received as much attention. As will be shown in Section 4.2, direct application of the above methods to face sketches leads to unsatisfactory results. This is because different portrait forms reveal 3D shapes from different cues. A few works have investigated 3D face modeling from silhouettes [24], contours [20], and line sketches [37, 15]. As silhouettes only encode the boundary information, multiple images are necessary to reconstruct shapes [24]. Occluding contours have been shown useful to estimate poses, but insufficient to reconstruct shapes from single images [20]. On the other hand, suggestive contours [8], which were developed in computer graphics for non-photo realistic rendering, have been demonstrated to be effective in conveying shape information. In [37], a functional mapping from the space of suggestive contours to the space of 3D shapes has been learned, which has been used to recover 3D shapes from sketches drawn by artists. A more recent work [15] has used deep learning to map hand-drawn line sketches to 3D face models, where the extended 3D face database [5] enables the modeling of exaggerated caricature faces. However, because of the sparsity, contours are still either insufficient to reconstruct shapes from a single image [37, 8] or unable to reconstruct shapes with sufficient details. More advanced methods are needed to address these limitations. In our work, we extract characteristic lines from both the sketches and the reconstructed 3D models, and propose a novel line loss function to guide the refinement of coarse models to preserve characteristic features in sketches.

Photo-to-Sketch Synthesis. Sketch synthesis is a classic problem in computer vision. Traditional methods can be divided into three main categories according to model construction technology [34]: subspace learning-based methods [30, 31, 23], sparse representation-based methods [33, 11, 6], and Bayesian inference-based methods [41, 12, 38, 35]. In recent years, generation methods have been used in photo-to-sketch synthesis. Generative Adversarial Networks (GANs) have achieved impressive results in image generation by learning the underlying distribution of images [13]. Isola et al. proposed a unified framework for image-to-image translation based on conditional GANs [18]. Zhu et al. proposed CycleGAN to learn image-to-image translation in an unsupervised way [42]. In our work, CycleGAN is used for photo-to-sketch synthesis to mitigate the demand for paired sketch-3D training data.

3. Approach

As illustrated in Figure 2, our coarse-to-fine framework consists of two sub-networks: CoarseNet and FineNet. The CoarseNet takes a single sketch as input and outputs a face model with a faithful overall shape. The output face model and a line drawing extracted from the input sketch are then fed into the FineNet, where a line loss is used to enhance characteristic features for finer detail reconstruction. In our work, face models are in the form of depth maps. We will detail individual components in the following subsections.

3.1. CoarseNet

To mitigate the demand for paired sketch-3D training data, we introduce two components into our method: photo-to-sketch synthesis and the synergy of 3D reconstruction from face photos. The former will synthesize sketches for input photos. With paired photo-3D data available, it will then build the bridge between the synthesized sketches and the 3D data to give the synthesized sketch-3D pairs. Specifically, we train the photo-to-sketch synthesis model using CycleGAN [42], and use the photos from FineData [14] and the sketches from CUFSF [35, 40] for training. Sketches are then synthesized for photos in FineData, which expands the photo-3D pairs to sketch-photo-3D tuples. We then use the synthesized sketch-3D pairs for training a baseline network, namely single-path CoarseNet (SP-CoarseNet), for reconstruction of 3D face models from sketches. Based on SP-CoarseNet, we further propose dual path CoarseNet (DP-CoarseNet) which integrates 3D reconstruction from face photos as an additional path in order to exploit the knowledge in reconstruction from a different modality.

3.1.1. SP-CoarseNet

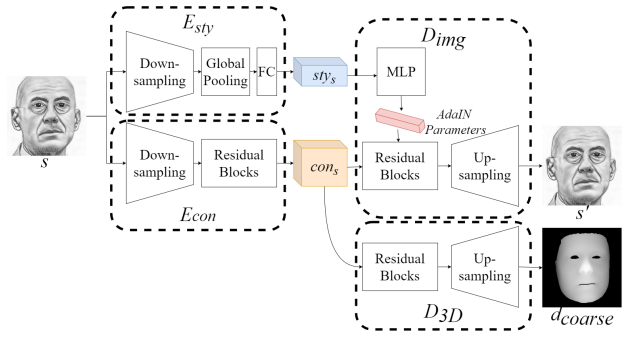
We assume that face shapes are independent of image styles (photos, sketches, paintings, etc.). Following the idea in [17], we thus decompose a sketch into the content space and the style space, where the content space contains face shape information and can further recover 3D face models.

As shown in Figure 3 (a), our SP-CoarseNet consists of a style encoder E_{sty}, a content encoder E_{con}, and two corresponding decoders D_{img} and D_{3D}. E_{sty} and E_{con} are used to generate style codes and content codes respectively, and D_{img} and D_{3D} are used to reconstruct input image and 3D face model accordingly. The architectures of the encoders and decoders are following MUNIT [17] with 3 downsampling and upsampling layers, and 128 filters in MLP. We also removed the Adaptive Instance Normalization (AdaIN) layer [16] in D_{3D}, as AdaIN is used for style transfer irrelevant to 3D reconstruction here. D_{3D} thus consists of residual blocks with consecutive upsampling layers.

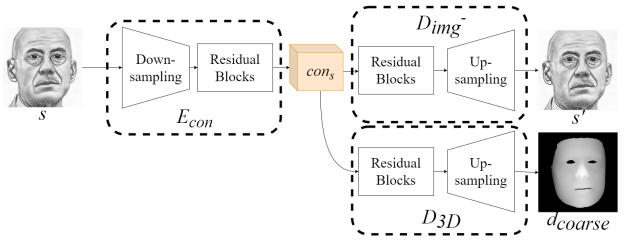
As a comparison, we also simplify SP-CoarseNet to SP-CoarseNet$^-$ as shown in Figure 3 (b), where the decomposition into style and content space is removed. A single code encodes both the image style and the face shape, and is used to reconstruct the image and the 3D face model. The reconstruction results of both networks will be compared in Section 4.3.

Two loss functions are used to train SP-CoarseNet and SP-CoarseNet$^-$. The first is a pixel-wise ℓ_2 loss measuring the similarity between the input sketch s and the reconstructed sketch s',

$$L_{rec\_s} = \frac{1}{\mathcal{P}} \|s - s'\|_2 \tag{1}$$



(a) SP-CoarseNet



(b) SP-CoarseNet$^-$

Figure 3: The architecture of our SP-CoarseNet and SP-CoarseNet$^-$.

where $\mathcal{P}$ is the number of pixels in the image. The second is a pixel-wise ℓ_2 loss between the ground-truth face depth map d_{gt} and the recovered face depth map d_{coarse_s},

$$L_{3d\_s} = \frac{1}{\mathcal{P}} \| d_{gt} - d_{coarse\_s} \|_2 \tag{2}$$

The final loss is a weighted sum of L_{rec_s} and L_{3d_s}:

$$L = \lambda_{rec} \cdot L_{rec\_s} + \lambda_{3d} \cdot L_{3d\_s} \tag{3}$$

where λ_{rec} and λ_{3d} are two parameters to balance the influence of the two loss terms.

3.1.2. DP-CoarseNet

SP-CoarseNet makes use of synthesized sketch-3D pairs to learn the reconstruction from sketches. Meanwhile, the photo-3D pairs are also available. We assume that 3D shape reconstructed from sketches and from photos is the same. We thus propose a dual-path architecture called DP-CoarseNet to achieve synergistic 3D reconstruction from both the sketch and the photo. As shown in Figure 2, each path is dedicated to reconstructing from one portrait form. We decompose the two portrait forms into their domain-specific style codes and the shared domain-invariant content codes, and the 3D shape is reconstructed from the latter. The idea is to make use of the complementary information in photos to assist the 3D reconstruction from sketches. The architecture of each component in DP-CoarseNet (E_{sty}, E_{con}, D_{img}, and D_{3D}) is the same as those in SP-CoarseNet.

For training DP-CoarseNet, in addition to the L_{rec_s} and L_{3d_s} losses in SP-CoarseNet, L_{rec_p} and L_{3d_p} are also used to enforce the similarity between the reconstructed photo/3D and the input/ground-truth along the reconstruction path from photo.

$$L_{rec\_p} = \frac{1}{\mathcal{P}} \| p - p' \|_2 \tag{4}$$

$$L_{3d\_p} = \frac{1}{\mathcal{P}} \| d_{gt} - d_{coarse\_p} \|_2 \tag{5}$$

where p is the input photo, p' is the reconstructed photo, d_{coarse_p} is the reconstructed face depth map from p, and d_{gt} is the ground-truth face depth map.

Moreover, the dual path also enables the reconstruction of images from the other portrait forms. Thus two further loss functions measuring the similarity between the input portrait and the portrait reconstructed from the other form are also used for training the DP-CoarseNet.

$$L_{cyt\_p} = \frac{1}{\mathcal{P}} \| p - s_{from\_p} \|_2 \tag{6}$$

$$L_{cyt\_s} = \frac{1}{\mathcal{P}} \| s - p_{from\_s} \|_2 \tag{7}$$

where s_{from_p} represents the sketch reconstructed from a photo, which is reconstructed using the content code of sketch (con_s) and style code of photo (sty_p), p_{from_s} is the photo reconstructed from a sketch, which is reconstructed from the content code of photo (con_p) and style code of sketch (sty_s). The final loss is a weighted sum of these loss terms:

$$\begin{aligned} L = \lambda_{rec} \cdot (L_{rec\_p} + L_{rec\_s}) + \lambda_{cyt} \cdot (L_{cyt\_p} + \\ L_{cyt\_s}) + \lambda_{3d} \cdot (L_{3d\_p} + L_{3d\_s}) \end{aligned} \tag{8}$$

λ_{rec}, λ_{cyt} and λ_{3d} are weighting parameters of the three loss terms.

3.2. Coarse-to-Fine Enhancement by Lines

We propose FineNet to refine the coarse face model obtained by CoarseNet. The details are enhanced through the characteristic lines extracted.

3.2.1. Line Drawing Extraction

As shown in Figure 2, two types of line drawings are introduced into FineNet. One is extracted from the input sketch, and the other is extracted from the reconstructed fine face model. We refine the coarse model by comparing these two types of line drawings with the aim to preserve in the refined fine model all the characteristic lines presented in the input sketch.

As shown in Figure 2, FineNet has two sets of encoder-decoders. One is a face model refinement module for refining coarse models, and the other is a line extraction module for extracting the characteristic lines from fine models. Specifically, we extract characteristic lines l from the input sketch using the XDoG edge detection method [36]. In order to extract the same type of lines from reconstructed face models, we train a line extraction module using the extracted XDoG lines from the synthesized sketches paired with their corresponding face models. The coarse model together with the extracted line drawing l are input into the face model refinement module, which outputs a displacement map to refine the coarse model. The refined model is input into the

pre-trained line extraction module to obtain its corresponding line drawing l', which will be compared with the input line drawing l to guide the restoration of facial details. The architecture of the encoders and the decoders are the same as E_{con} and D_{3D} in Figure 3 respectively.

3.2.2. Line Loss Function

We expect the refined face model can preserve all the characteristic details depicted in the sketch. Reflected in comparison of the two line drawings, a principle is to let l' preserve as much as possible the characteristic lines in l. We thus define the line loss function,

$$L_{line} = \|(1 - l) \cdot l'\|_1 \tag{9}$$

In our line drawings, characteristic lines are depicted by black pixels with intensity value of 0, and the background are white pixels with intensity value of 1. Specifically, if l' contains all characteristic lines in l then the value of L_{line} is 0. Otherwise, $L_{line} > 0$. Minimization of L_{line} allows us to refine the coarse model so that the reconstructed fine model reflects the characteristic lines depicted in the sketch.

In addition to the line loss, a pixel-wise ℓ_2 loss is also used to encourage small displacements,

$$L_{dis} = \frac{1}{\mathcal{P}} \left\| d_{dis} \right\|_2 \tag{10}$$

where d_{dis} is the displacement map from the refinement module, which is used to refine the coarse model by adding pixel-wise geometric details. The final loss is a weighted sum of L_{line} and L_{dis}:

$$L = \lambda_{line} \cdot L_{line} + \lambda_{dis} \cdot L_{dis} \tag{11}$$

where λ_{line} and λ_{dis} are two parameters to balance the influence of the two loss terms.

4. Experiments

In this section, we first give an overview of the datasets, the evaluation metrics, and the implementation details. We then demonstrate 1) the necessity of designing a dedicated sketch 3D reconstruction network, and 2) the effectiveness of the dual-path design and the line loss, through both qualitative and quantitative evaluations. We compare the results of our method with the previous state-of-the-art 3D face reconstruction methods. We also show the use of our method for editing the details on the reconstructed face models, and the versatility of the method for line drawing sketches.

4.1. Experimental Details

We first give details of the datasets, the implementation, and the metrics that are used in the following experiments.

4.1.1. Dataset Preprocessing

The CUHK Face Sketch FERET Database (CUFSF) [35, 40], the FineData [14], and the ZJU-VIPA Line Drawing Face Database [22] are used in the experiments. FineData has about 3131×30 photo-depth pairs augmented from the 3131 images in 300-W [25], and constructed by transferring the details from other images to the original images. Each face image is augmented 30 times with different details. We randomly split the dataset into a training set of 2501×30 pairs and a test set of 630×30 pairs. CUFSF has 1194 photo-sketch pairs. The sketches in CUFSF are drawn by an artist when viewing the corresponding photos. We randomly split the dataset into a training set of 900 pairs and a test set of 294 pairs. To align the faces in CUFSF and FineData, we crop the faces in CUFSF according to the coordinates of the eyes. When training the CycleGAN model for photo-to-sketch synthesis, to address the head pose variance in FineData, we augment CUFSF by rotating the sketch images by some random small angles and double the number of sketches. In addition, we randomly selected 2000 photos from the training set of FineData, of which only one of the 30 augmented images was chosen so that the numbers of training images from the two datasets are similar. The ZJU-VIPA Line Drawing Face Database includes 188 line drawing type sketches which are drawn based on the face photos for the CUHK Face Sketch Database (CUFS) [35]. These line drawings are used in our experiments to evaluate the versatility of our method for different types of sketches. All the images used in this work are resized to 256×256.

4.1.2. Evaluation Metrics

One of our goals is to reconstruct a face model with faithful overall shape and characteristic features preserved or even enhanced. To achieve this aim, it is important to both achieve accuracy in the reconstruction, and preserve the diversity between different faces. An observation is that most methods have the tendency to reconstruct similar face shapes for different people. Thus, in addition to visual qualitative evaluation, we conducted two quantitative analysis to evaluate both the accuracy and the diversity of the reconstructed face shapes. The two evaluation metrics used are: 1) the depth error Err, and 2) the variance Var calculated from the determinant of the covariance matrix.

The depth error measures the difference between the reconstructed face model and the ground-truth face model,

$$\mathrm{Err} = \frac{1}{\mathrm{K}} \sum_{\mathrm{i}=1}^{\mathrm{K}} \frac{1}{\mathcal{P}} \|\mathrm{d}_{\mathrm{gt_i}} - \mathrm{d}_{\mathrm{out_i}}\|_2 \tag{12}$$

where $\mathcal{P}$ is the number of pixels, K is the number of test samples, d_{gt_i} is the i-th ground-truth face depth map, d_{out_i} is the i-th reconstructed face depth map.

The variance, calculated from the determinant of the covariance matrix, measures the diversity of a set of face models. The larger the variance, the better the diversity of the set of face models. Specifically, the set of face depth maps is arranged into a data-matrix, where each column of the data-matrix is a vector representing a face depth map. Based on the data-matrix, we obtain the eigenvalues of its covariance matrix using the snap-shot method [29], then the determinant is the product of all the eigenvalues. In order to limit the range of the measured variance, Var is defined as the n-th

Figure 4: Qualitative comparison results of PhotoNet on CUFSF. The first column shows the input photos. The second to sixth columns show the results of 3DDFA [43], PRNet [10], Deng et al. [9], DF^{2}Net [39] of the photos. The seventh column shows the input sketches. The eighth to twelfth columns show the results of 3DDFA [43] (retrained using sketch-3D paired data), PRNet [10], Deng et al. [9], DF^{2}Net [39] of the sketches. The last column shows the results of our FineNet.

Method	Var(Photo)	Var(Sketch)
3DDFA [43]	154.8393	76.1249
PRNet [10]	313.6973	243.6036
Deng et al. [9]	185.9193	115.5553
DF^{2}Net [39]	330.2838	366.0855

Table 1
Quantitative comparison results of PhotoNet on CUFSF. The result is calculated from 2D images rendered by 3D face models. The larger the value of Var, the better the diversity of the face models is maintained. The diversity of reconstructed face models from sketches is lower than that from photo. For DF^{2}Net, larger Var value for sketches is due to unreasonable distortions (in particular see the side views).

root of the determinant value:

$$\mathrm{Var} = (\prod_{\mathrm{i}=1}^{\mathrm{n}} \lambda_{\mathrm{i}})^{\frac{1}{\mathrm{n}}} \tag{13}$$

where n is the number of eigenvalues , λ_i is the i-th eigenvalue.

4.1.3. Implementation Details

Our framework is implemented through PyTorch. We train our network in a two-stage scheme. In the first stage, we train SP-CoarseNet and SP-CoarseNet$^-$ with a batch size of 8, and train DP-CoarseNet with a batch size of 2. In the second stage, we fix CoarseNet and train FineNet with a batch size of 4. We empirically set $\lambda_{3d} = 0.001$, $\lambda_{rec} = 1$, $\lambda_{cvt} = 1$, $\lambda_{line} = 0.85$, and $\lambda_{dis} = 0.15$. We use ADAM optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ to train our network, and the learning rate is set to 0.0001. ReLU is used as activation of our network. We fix the dimension of the input image and output face model (depth map) to 256×256, then the dimension of the content codes and style codes are $512 \times 32 \times 32$ and $8 \times 1 \times 1$ respectively. We perform training on an NVIDIA GeForce RTX 2080 Ti GPU, and the training of SP-CoarseNet and SP-CoarseNet$^-$ takes about 30 hours, while DP-coarseNet takes about 80 hours, and FineNet takes about 8 hours.

4.2. Experiments on PhotoNets

As described in Section 2, there are many existing methods to reconstruct 3D face models from photos, such as 3DMM-based methods [43, 10, 9] and SFS-based methods [39], which are referred to as PhotoNets in this paper. To demonstrate the need to design a dedicated sketch 3D reconstruction network instead of using an existing PhotoNet, we input photos and sketches into several state-of-the-art PhotoNets respectively. Among these PhotoNets, only the training code of 3DDFA [43] is available. We thus retrain the 3DDFA network using the sketch-3D paired data for a fair comparison.

We conduct experiments on CUFSF [35, 40]. Figure 4 shows a qualitative comparison. 3DMM-based methods fail to capture fine geometric details and ignore the characteristic details no matter for photos or sketches [43, 10, 9]. SFS-based methods can generate details, but the unrealistic shading in sketches distorts the generated overall shape [39]. As the sketches in CUFSF do not have ground-truth face models for comparison, we quantitatively compare the diversity of the reconstructed coarse face models, measured by the Var value. Table 1 shows a quantitative comparison. As the face models reconstructed by different PhotoNets are different in representation and facial area coverage, comparison between different methods is difficult. Therefore, Table 1 is to provide the comparison between the reconstructions from photos and from sketches using the *same* PhotoNet (row-wise), rather than the comparison between different PhotoNets. From Table 1, we can see that for 3DMM-based methods, the diversity of reconstructed face models from sketches is lower than that from photos [43, 10, 9], while for SFS-based methods, the reconstructed face models of sketches are more diverse, but this is due to unreasonable distortions in face mod-

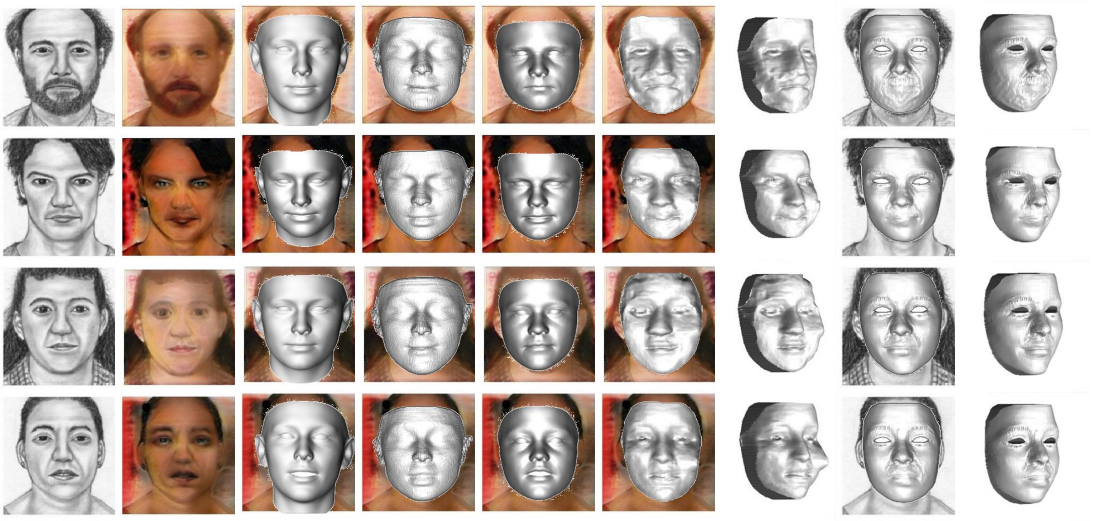


Figure 5: Qualitative comparison of 3DDFA [43], PRNet [10], Deng et al. [9], DF^{2}Net [39] and our FineNet on CUFSF by translating the input sketch into a realistic photo. The photos in this figure are translated by CycleGAN [42].

els [39]. We also display in the last column of Figure 4 the results reconstructed using the proposed FineNet. It can be observed that FineNet can both reconstruct the overall face shape and preserve the characteristic details.

To further explore the photo-based methods and ensure they are not disadvantaged, we conduct additional experiments by translating the input sketches into realistic photos before applying the PhotoNets. CycleGAN [42] and DeepFaceDrawing [7] are used for translating sketches into photos, which are representative methods for general-purpose image translation and sketch-to-photo translation, respective. The reconstruction results from the photos translated using the two methods are shown in Figure 5 and Figure 6 respectively. It can be observed that this strategy largely depends on the quality of the sketch-to-photo translation, especially for DF^{2}Net [39] which is shading-based. The blurry translation using CycleGAN (Figure 5) results in poor overall shape reconstruction using DF^{2}Net [39]. On the other hand, DeepFaceDrawing [7] can generate high-quality face photos but lacks fidelity (Figure 6), resulting in the reconstructed 3D face models not resembling the original sketches. It is also observed that no matter using CycleGAN or DeepFaceDrawing, the 3DMM-based methods [43, 10, 9] fail to capture the fine geometric details of the face. The reconstructions using the proposed FineNet are displayed in the last two columns of Figures 5 and 6 for comparison.

The above two experiments both demonstrate that the existing PhotoNets are not suitable for 3D face reconstruction from face sketches, and a new reconstruction method needs to be designed.

4.3. Experiments on CoarseNet

We first compare the reconstructed coarse models using DP-CoarseNet, SP-CoarseNet and SP-CoarseNet$^-$. We also try our best to reproduce the D-Net, which is the coarse model of DF^{2}Net [39], carefully following the description of the paper. The reproduced method is used for generation of coarse face models in comparison with ours. We also carry out ablation studies on the loss function of the DP-CoarseNet (Eq.(8)) to verify its effectiveness. Our experiments are conducted on both FineData and CUFSF.

Method	Err
D-Net [39]	20.8882
SP-CoarseNet$^-$	16.2249
SP-CoarseNet	16.1479
DP-CoarseNet	**15.6372**

Table 2
Quantitative comparison results of CoarseNet on FineData.

4.3.1. Comparison of CoarseNet Architectures

FineData: As there are synthesized sketch-3D pairs, the reconstructed coarse face models from sketches can be directly compared with the ground-truth. Table 2 shows the reconstruction errors using the four network architectures, from which the DP-CoarseNet achieves the lowest depth error. Figure 7 shows the qualitative comparison of the results from the four network architectures. Although the difference in the overall shapes are subtle, the enlarged views show that DP-CoarseNet can recover clearer and more faithful face shapes.

CUFSF: As the sketches in CUFSF do not have ground-truth face models for comparison, we quantitatively compare the diversity of the reconstructed coarse face models, measured by the Var value. As illustrated in Table 3, DP-CoarseNet achieves the highest variance among the recovered face shapes, which shows the superiority of the dual-path design over the other architectures for reconstruction of coarse face shapes. Figure 8 shows the qualitative compari-

Figure 6: Qualitative comparison of 3DDFA [43], PRNet [10], Deng et al. [9], DF^{2}Net [39] and our FineNet on CUFSF by translating the input sketch into a realistic photo. The photos in this figure are translated by DeepFaceDrawing [7].

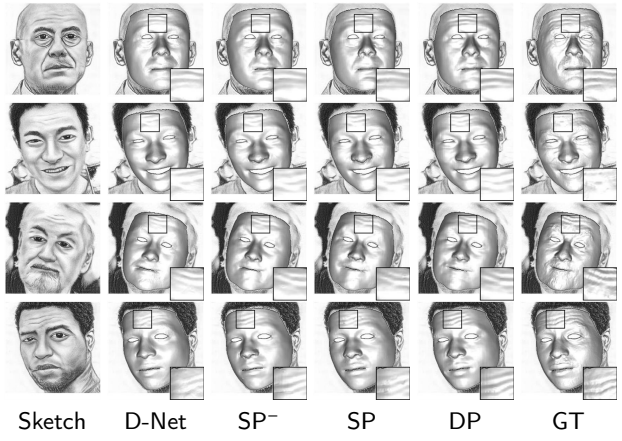


Figure 7: Qualitative comparison results of CoarseNet on FineData, where D-Net stands for the D-Net in [39], SP$^-$ stands for the SP-CoarseNet$^-$, SP stands for the SP-CoarseNet, DP stands for the DP-CoarseNet, GT stands for the ground-truth face models.

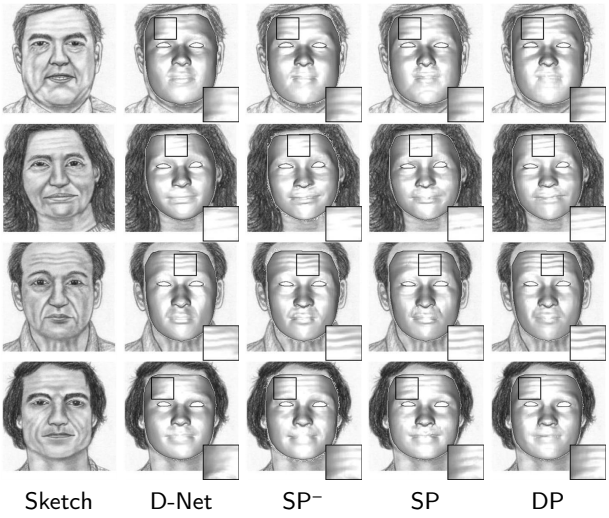


Figure 8: Qualitative comparison results of CoarseNet on CUFSF.

son. Similar to the results on FineData, the reconstructions using DP-CoarseNet have clearer details than those using the other network architectures.

Several observations can be made from these experiments. First, our dual-path network shows the best results both qualitatively and quantitatively, which demonstrates that the synergy of 3D reconstruction from face photos is an effective way to boost the reconstruction from sketches. Second, the qualitative results of SP-CoarseNet and SP-CoarseNet$^-$ are not particularly different, but compared to SP-CoarseNet$^-$, SP-CoarseNet achieves lower depth error and larger shape variance. It demonstrates the effectiveness to decompose the image into content space and style space. Last but not the least, our three types of CoarseNet are superior both qualitatively and quantitatively to D-Net [39]. The possible reason

Method	Var
D-Net [39]	24.3010
SP-CoarseNet$^-$	26.5111
SP-CoarseNet	29.0095
DP-CoarseNet	**30.2642**

Table 3
Quantitative comparison results of CoarseNet on CUFSF.

is that the residual blocks used in our network increase the expressive power of the network.

4.3.2. Ablation Study on CoarseNet Loss Function

To further demonstrate the effectiveness of including the image reconstruction loss L_{rec} and the image conversion loss

Methods	Err
$L_a = \lambda_{3d} \cdot L_{3d}$	16.1782
$L_b = \lambda_{3d} \cdot L_{3d} + \lambda_{rec} \cdot L_{rec}$	16.3843
$L_c = \lambda_{3d} \cdot L_{3d} + \lambda_{cvt} \cdot L_{cvt}$	16.0900
$L = \lambda_{3d} \cdot L_{3d} + \lambda_{rec} \cdot L_{rec} + \lambda_{cvt} \cdot L_{cvt}$	**15.6372**

Table 4
Quantitative comparison results of different DP-CoarseNet loss functions on FineData.

Methods	Var
$L_a = \lambda_{3d} \cdot L_{3d}$	25.3564
$L_b = \lambda_{3d} \cdot L_{3d} + \lambda_{rec} \cdot L_{rec}$	25.2439
$L_c = \lambda_{3d} \cdot L_{3d} + \lambda_{cvt} \cdot L_{cvt}$	27.3080
$L = \lambda_{3d} \cdot L_{3d} + \lambda_{rec} \cdot L_{rec} + \lambda_{cvt} \cdot L_{cvt}$	**30.2642**

Table 5
Quantitative comparison results of different DP-CoarseNet loss functions on CUFSF.

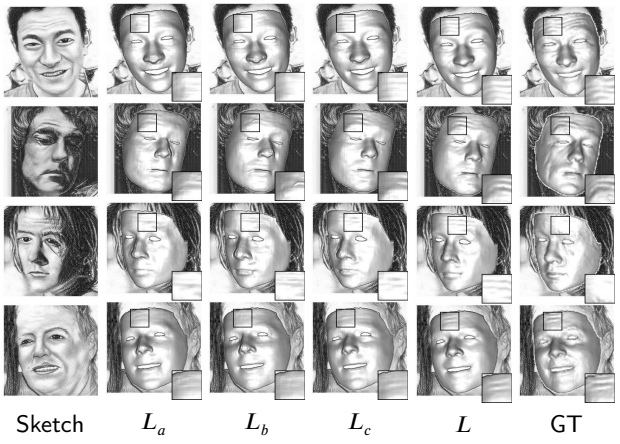


Figure 9: Qualitative comparison results of different DP-CoarseNet loss functions on FineData, where GT stands for the ground-truth face models. The definitions of L_a, L_b, L_c and L are the same as those in Table 4 and Table 5.

L_{cvt} into the DP-CoarseNet loss function, we perform ablation studies on both FineData and CUFSF. For FineData, we quantitatively calculate the average depth errors of the reconstructions obtained using different loss functions. For CUFSF, we quantitatively compare the diversity of the reconstructed coarse face models. As illustrated in Table 4 and Table 5, the loss function with both L_{rec} and L_{cvt} achieves the lowest depth error and the highest variance among the recovered face shapes. Besides, it can be observed that the quantitative result of only adding L_{rec} is worse. The possible reason is that L_{rec} alone cannot well control the reasonable decomposition of the content space and style space, which further necessitates the combination of L_{rec} and L_{cvt}. We also qualitatively compare the reconstruction results of different loss functions. The enlarged views in Figure 9 and Figure 10 indicate that the combination of L_{3d}, L_{rec} and L_{cvt} can restore clearer and more faithful facial details.

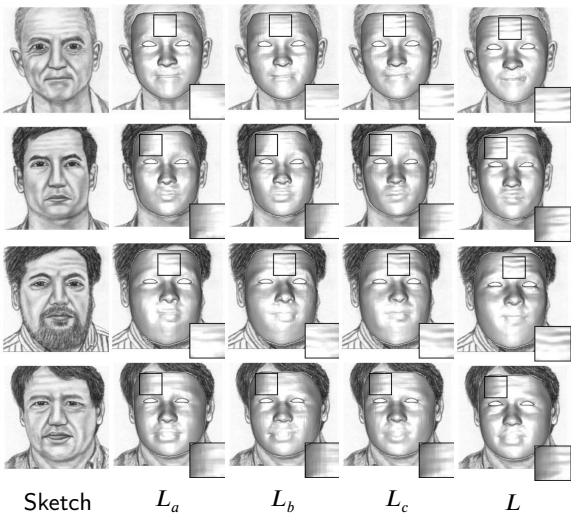


Figure 10: Qualitative comparison results of different DP-CoarseNet loss functions on CUFSF. The definitions of L_a, L_b, L_c and L are the same as those in Table 4 and Table 5.

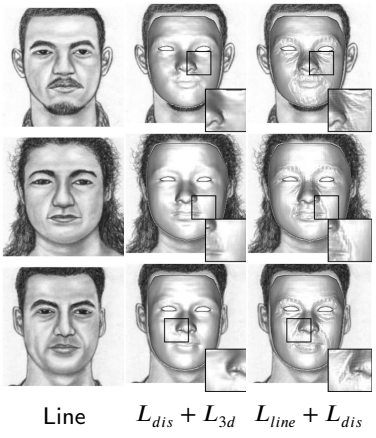


Figure 11: Qualitative comparison results of different FineNet loss functions on CUFSF.

Methods	Shape	Similarity	Detail
$L_{dis} + L_{3d}$	**7.4833**	5.7333	4.3111
$L_{line} + L_{dis}$	7.4056	**7.2167**	**6.7444**

Table 6
User study results of different FineNet loss functions.

4.4. Experiments on FineNet

The aim of the FineNet is to refine the face model so that it can preserve all the characteristic details depicted in the sketch. We conduct experiments to demonstrate the effectiveness of the proposed line loss function, and show that it also enables easy editing of details on the reconstructed face models. We also conduct comparative experiments on line drawing sketches to show the versatility of the method.

4.4.1. Ablation Study on Line Loss Function

We first evaluate the effect of the line loss function on FineNet results. Specifically, we remove the line loss L_{line} term in the loss function of FineNet (Eq.(11)). As the displacement loss L_{dis} alone cannot guide an effective training, we replace the line loss term with a 3D face reconstruction loss term $L_{3d} = \frac{1}{\mathcal{P}}\|d_{gt} - d_{pre}\|_2$, where d_{gt} is the ground-truth depth map, d_{pre} is the reconstructed face depth map, and $\mathcal{P}$ is the number of pixels in the sketch. Comparing the two loss functions: $L_{line} + L_{dis}$ and $L_{dis} + L_{3d}$, the former emphasizes the characteristic features depicted by lines, while the latter tries to reconstruct faithful shapes. Figure 11 confirms their effects, where using the loss $L_{line} + L_{dis}$ reconstructs face models with more emphasized details than using $L_{dis} + L_{3d}$. We also conduct a user study for quantitative comparison between the two loss functions. The user study involves 15 participants. 10 sketches are randomly chosen from the test set of CUFSF. For each sketch, two reconstructions are obtained using the two loss functions respectively. The sketches and the reconstructions thus form the 20 pairs (a sketch and one of the two reconstructions) to be presented to each participant. Three aspects are evaluated in the user study: the overall shape of the face model (Shape), the similarity between the face model and the original sketch (Similarity), and the restoration of the characteristic details on the face model (Detail). The score for each aspect ranges from 1 to 10. The average scores are shown in Table 6, which shows that the participants agree that using the line loss better reconstructs the details from the input sketches and gives more similar appearance. However, in terms of the overall shape, using the 3D loss scores slightly better than using the line loss. The reason may lie in the users' preference to a smoother shape when evaluating the overall shape aspect. Overall, both the qualitative and quantitative comparisons demonstrate the effectiveness of the proposed line loss function in guiding the reconstruction to preserve characteristic details in face sketches.

We further conduct experiment on CUFSF with varying λ_{line} and λ_{dis}, where we set $\lambda_{line} + \lambda_{dis} = 1$. Figure 12 shows a qualitative comparison. We can see that varying λ_{line} and λ_{dis} controls the level of details on the reconstructed face model. The larger λ_{line}, the higher emphasis on preserving characteristic lines, and thus more obvious details generated. We use the setting $\lambda_{line} = 0.85$ and $\lambda_{dis} = 0.15$ by default in the following experiments.

4.4.2. Comparison with State-of-the-Art

We qualitatively compare FineNet with DP-CoarseNet, DeepSketch2Face [15] and DF^{2}Net [39] on CUFSF. DeepSketch2Face enables the modeling of exaggerated caricature faces, which are composed of sparse contours, as shown in the second column of Figure 13. DF^{2}Net is the state-of-the-art 3D face reconstruction method based on photos, which can recover facial details. As shown in Figure 13, DeepSketch2Face does not emphasize recovering faithful overall shape as ours. DF^{2}Net can recover 3D faces with details, but the unrealistic shading in sketches distorts the generated overall shape. Besides, compared with the results of DP-CoarseNet, FineNet can reconstruct the details of the face through characteristic lines. More results are shown in Figure 16.

Methods	Shape	Similarity	Detail
DeepSketch2Face [15]	6.3030	2.9879	3.0000
DF^{2}Net [39]	4.9636	6.8788	6.2970
DP-CoarseNet	**7.6424**	5.9697	4.8061
FineNet	7.6303	**7.2061**	**6.8606**

Table 7
User study results of different methods.

We also conduct a user study to compare the results quantitatively. The setting of the user study is the same as that in the ablation study on line loss function. The average scores are shown in Table 7. FineNet achieves the best scores in terms of similarity to the input sketches and detail preservation, and DF^{2}Net ranks the second in these two aspects. As both methods specifically consider the reconstruction of facial details, the ranking is as expected. On the overall shape, DF^{2}Net scores the least with the visually obvious distortions. FineNet scores slightly less than DP-CoarseNet, for which the reason again may lie in the users' preference to a smoother shape when evaluating the overall shape aspect. Overall, FineNet has evident strengths in reconstruction from sketches comparing to other methods.

4.4.3. Reconstruction from Line Drawing

To demonstrate the versatility of our method to different types of sketches, we conduct an experiment on line drawing sketches from the ZJU-VIPA Line Drawing Face Database [22]. We qualitatively compare the reconstruction results from DP-CoarseNet, DeepSketch2Face [15], DF^{2}Net [39], and FineNet. For DeepSketch2Face [15], we pre-processed the line drawings to adapt them to the input form required (the second column in Figure 14). Figure 14 shows a qualitative comparison. Similar to the results in Figure 13, DeepSketch2Face, because of the different purposes, does not recover faithful overall shapes or preserve facial details. DF^{2}Net preserves the details well but distorts the overall shapes. In contrast, our method can reconstruct the overall shapes well through the DP-CoarseNet, and also preserves the details through the following FineNet.

4.4.4. 3D Face Editing

FineNet uses a characteristic line drawing to guide the refinement of face models. Editing on the line drawing will be reflected in the reconstructed face model. We demonstrate that FineNet can be used for 3D face editing by allowing users to add or remove lines. Examples are shown in Figure 15, where wrinkles and scars can be easily introduced into or removed from the reconstructed face models.

5. Conclusions

While photos record facial appearance from realistic shading, sketches are more subjective recordings with empha-

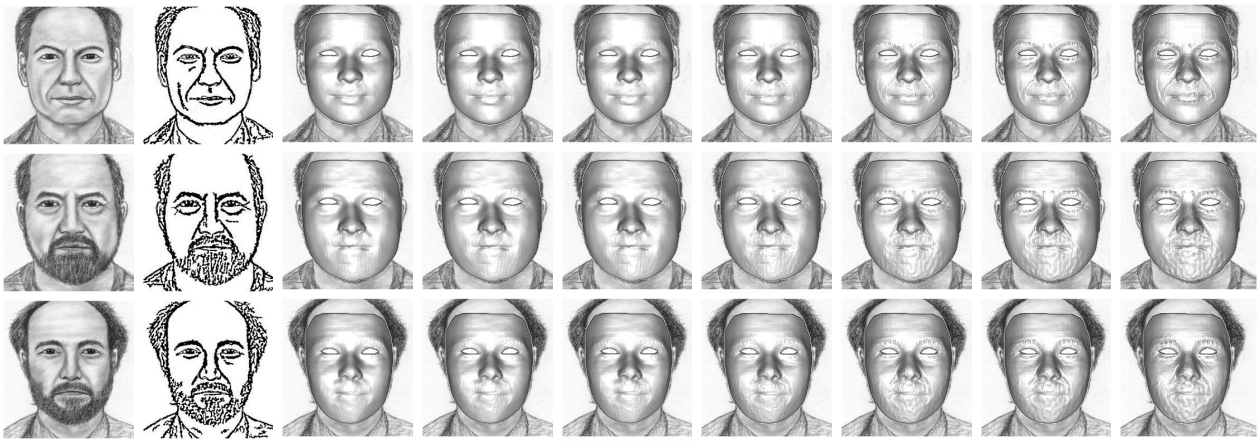


Figure 12: Qualitative comparison results of different λ_{line} and λ_{dis} on CUFSF, where λ_{line} and λ_{dis} are represented in the form of $(\lambda_{line}, \lambda_{dis})$.

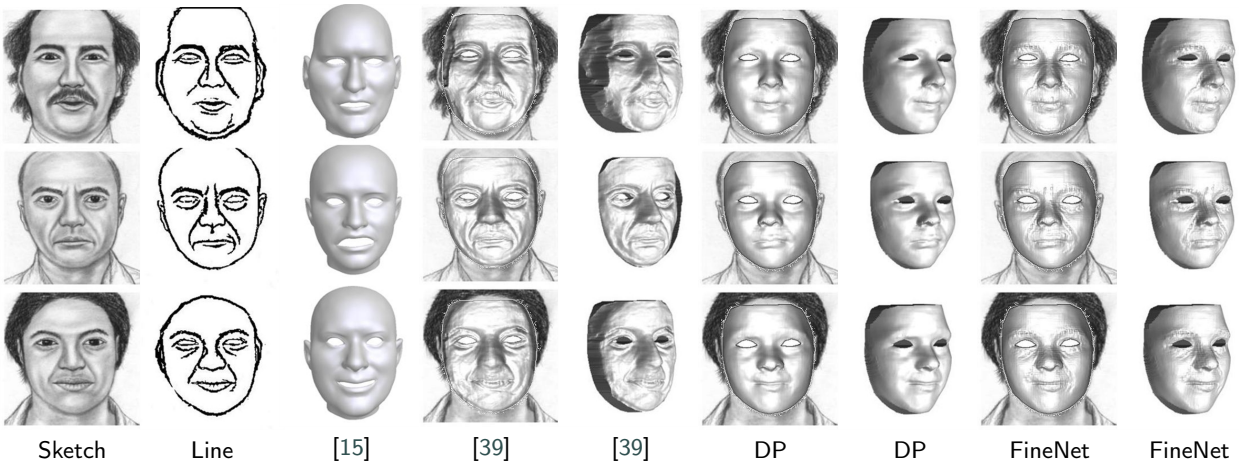


Figure 13: Qualitative comparison between DeepSketch2Face [15], DF^{2}Net [39], our DP-CoarseNet and our FineNet on CUFSF. Please note that Line is the input of [15], Sketch is the input of [39] and our DP-CoarseNet, while Sketch and Line are the inputs of our FineNet. More results are shown in Figure 16.

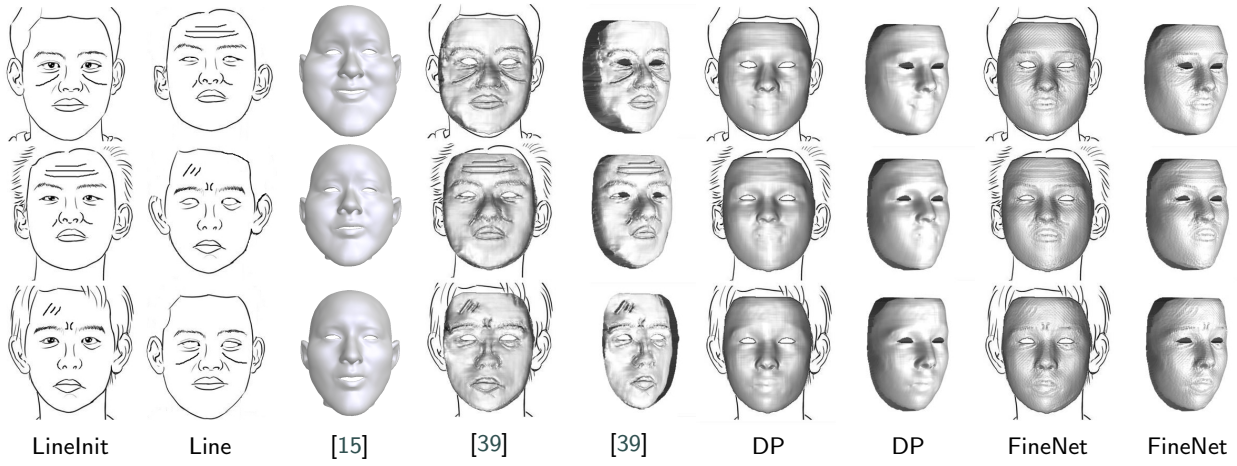


Figure 14: Qualitative comparison between DeepSketch2Face [15], DF^{2}Net [39], our DP-CoarseNet and our FineNet on line drawings. Please note that Line is the input of [15], LineInit is the input of [39], our DP-CoarseNet and our FineNet.

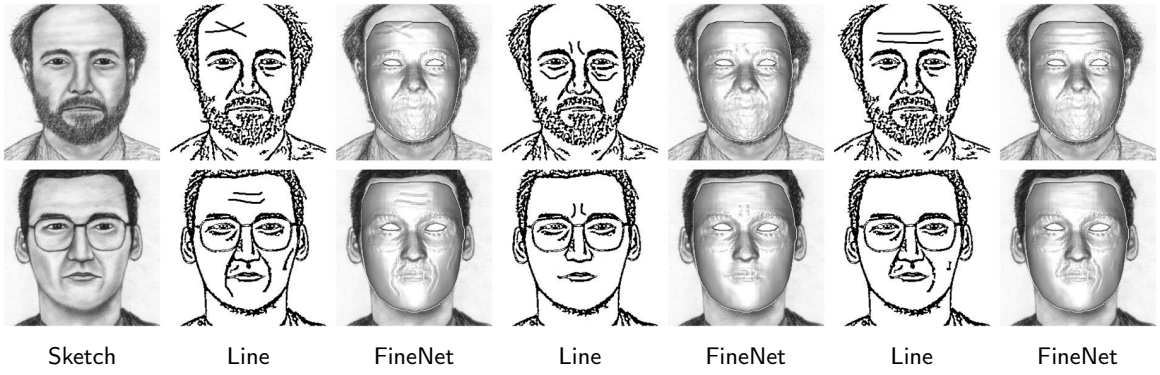


Figure 15: Qualitative results of 3D face editing using FineNet on CUFSF.

sized characteristic features. In this paper, we propose the first sketch 3D face reconstruction framework that can generate a 3D face model with faithful overall shape and preserved or even enhanced characteristic features. We use photo-to-sketch synthesis and a dual-path design to mitigate the demand of paired sketch-3D training data, and to enable the reconstruction process to take advantage of complementary information in face photos. We also design a novel line loss function and demonstrate its ability to preserve the characteristic details presented in the sketch. Comprehensive experiments have shown that our method outperforms previous methods in 3D face reconstruction from sketches. We also demonstrate our method can be used for easy editing of 3D face models with added or removed characteristic details.

Our work shows some initial results that different portrait forms can benefit the 3D reconstruction of faces from different aspects. It may also inspire further exploration of multi-modal learning for 3D face reconstruction. As an initial exploration, the current work inevitably has limitations and open directions for future improvements. An immediate future work is to improve the detail geometry. The details of the reconstructed face models are not always visually pleasing. It is mainly because facial details have different types, e.g., the different geometries of forehead wrinkles and smile lines. This will require different reconstruction strategies/parameters in dealing with different facial regions and detail types. Possible solutions include segmenting faces into different regions for separate training, and extracting non-binary lines to better represent the detail geometries. Another direction for future work is to incorporate real sketches into the training of the dual-path network to reduce the artifacts caused by the gap between real and synthetic data.

References

[1] Bas, A., Smith, W.A., Bolkart, T., Wuhrer, S., 2016. Fitting a 3D morphable model to edges: A comparison between hard and soft correspondences, in: Asian Conference on Computer Vision, Springer. pp. 377–391.

[2] Blanz, V., Mehl, A., Vetter, T., Seidel, H.P., 2004. A statistical method for robust 3D surface reconstruction from sparse data, in: Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 2004. 3DPVT 2004., IEEE. pp. 293–300.

[3] Blanz, V., Vetter, T., 1999. A morphable model for the synthesis of 3D faces, in: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, pp. 187–194.

[4] Booth, J., Antonakos, E., Ploumpis, S., Trigeorgis, G., Panagakis, Y., Zafeiriou, S., 2017. 3D face morphable models "in-the-wild", in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE. pp. 5464–5473.

[5] Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K., 2013. FaceWarehouse: A 3D facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics 20, 413–425.

[6] Chang, L., Zhou, M., Han, Y., Deng, X., 2010. Face sketch synthesis via sparse representation, in: 2010 20th International Conference on Pattern Recognition, IEEE. pp. 2146–2149.

[7] Chen, S.Y., Su, W., Gao, L., Xia, S., Fu, H., 2020. DeepFaceDrawing: Deep generation of face images from sketches. ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2020) 39, 72:1–72:16.

[8] DeCarlo, D., Finkelstein, A., Rusinkiewicz, S., Santella, A., 2003. Suggestive contours for conveying shape, in: ACM SIGGRAPH 2003 Papers, pp. 848–855.

[9] 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: IEEE Computer Vision and Pattern Recognition Workshops.

[10] 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.

[11] Gao, X., Wang, N., Tao, D., Li, X., 2012. Face sketch–photo synthesis and retrieval using sparse representation. IEEE Transactions on circuits and systems for video technology 22, 1213–1226.

[12] Gao, X., Zhong, J., Tao, D., Li, X., 2008. Local face sketch synthesis learning. Neurocomputing 71, 1921–1930.

[13] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

[14] Guo, Y., Cai, J., Jiang, B., Zheng, J., et al., 2018. CNN-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. IEEE transactions on pattern analysis and machine intelligence 41, 1294–1307.

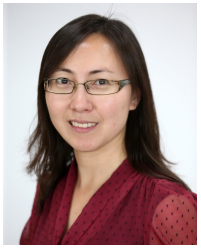
[15] Han, X., Gao, C., Yu, Y., 2017. DeepSketch2Face: A deep learning based sketching system for 3d face and caricature modeling. ACM Trans. Graph. 36, 126:1–126:12.

[16] Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510.

[17] Huang, X., Liu, M.Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image-to-image translation, in: Proceedings of the Eu-

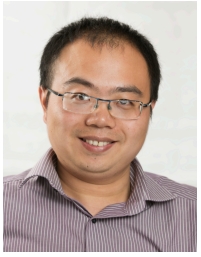
ropean Conference on Computer Vision (ECCV), pp. 172–189.
[18] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134.
[19] Jiang, L., Zhang, J., Deng, B., Li, H., Liu, L., 2018. 3D face reconstruction with geometry details from a single image. IEEE Transactions on Image Processing 27, 4756–4770.
[20] 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. pp. 261–273.
[21] Kemelmacher-Shlizerman, I., Basri, R., 2010. 3D face reconstruction from a single image using a single reference face shape. IEEE transactions on pattern analysis and machine intelligence 33, 394–405.
[22] Liang, Y., Song, M., Xie, L., Bu, J., Chen, C., 2012. Face sketch-to-photo synthesis from simple line drawing, in: Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE. pp. 1–5.
[23] Liu, Q., Tang, X., Jin, H., Lu, H., Ma, S., 2005. A nonlinear approach for face sketch synthesis and recognition, in: 2005 IEEE Computer Society conference on computer vision and pattern recognition (CVPR'05), IEEE. pp. 1005–1010.
[24] 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. pp. 20–27.
[25] Pantic, M., Tzimiropoulos, G., Zafeiriou, S., 2013. 300 faces in-the-wild challenge (300-w), in: ICCV Workshop.
[26] 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. pp. 296–301.
[27] Richardson, E., Sela, M., Kimmel, R., 2016. 3D face reconstruction by learning from synthetic data, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE. pp. 460–469.
[28] Richardson, E., Sela, M., Or-El, R., Kimmel, R., 2017. Learning detailed face reconstruction from a single image, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1259–1268.
[29] Sirovich, L., 1987. Turbulence and the dynamics of coherent structures. Quarterly of applied mathematics XLV, 561–590.
[30] Tang, X., Wang, X., 2002. Face photo recognition using sketch, in: Proceedings. International Conference on Image Processing, IEEE. pp. 257–260.
[31] Tang, X., Wang, X., 2004. Face sketch recognition. IEEE Transactions on Circuits and Systems for video Technology 14, 50–57.
[32] Troje, N.F., Bülthoff, H.H., 1996. Face recognition under varying poses: The role of texture and shape. Vision research 36, 1761–1771.
[33] Wang, N., Li, J., Tao, D., Li, X., Gao, X., 2013. Heterogeneous image transformation. Pattern Recognition Letters 34, 77–84.
[34] Wang, N., Tao, D., Gao, X., Li, X., Li, J., 2014. A comprehensive survey to face hallucination. International journal of computer vision 106, 9–30.
[35] Wang, X., Tang, X., 2008. Face photo-sketch synthesis and recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 31, 1955–1967.
[36] Winnemöller, H., Kyprianidis, J.E., Olsen, S.C., 2012. XDoG: an extended difference-of-gaussians compendium including advanced image stylization. Computers & Graphics 36, 740–753.
[37] Wuhrer, S., Shu, C., 2012. Shape from suggestive contours using 3D priors, in: 2012 Ninth Conference on Computer and Robot Vision, IEEE. pp. 236–243.
[38] Xiao, B., Gao, X., Tao, D., Yuan, Y., Li, J., 2010. Photo-sketch synthesis and recognition based on subspace learning. Neurocomputing 73, 840–852.
[39] Zeng, X., Peng, X., Qiao, Y., 2019. DF2Net: A dense-fine-finer network for detailed 3D face reconstruction, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2315–2324.
[40] Zhang, W., Wang, X., Tang, X., 2011. Coupled information-theoretic encoding for face photo-sketch recognition, in: CVPR 2011, IEEE. pp. 513–520.
[41] Zhong, J., Gao, X., Tian, C., 2007. Face sketch synthesis using E-HMM and selective ensemble, in: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, IEEE. pp. 485–488.
[42] Zhu, J.Y., Park, T., Isola, P., Efros, A.A., 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE international conference on computer vision, pp. 2223–2232.
[43] Zhu, X., Liu, X., Lei, Z., Li, S.Z., 2017b. Face alignment in full pose range: A 3D total solution. IEEE transactions on pattern analysis and machine intelligence 41, 78–92.

Li Yang received the BSc degree from China University of Mining and Technology, China in 2017. She received the MSc degree from Nanjing University, China in 2020. She a member of R&L Group in Nanjing University, which is led by professor Yang Gao. Her research interests are mainly on machine learning, deep learning and their application in 3D face reconstruction.



Jing Wu received the BSc and MSc degrees from Nanjing University, China in 2002 and 2005 respectively. She received the PhD degree in computer science from the University of York, UK in 2009. She is currently a lecturer in the School of Computer Science and Informatics at Cardiff University. Her research interests lie in the area of image-based 3D reconstruction and its application in object modelling and environment mapping.

Jing Huo received the PhD degree from the Department of Computer Science and Technology, Nanjing University, Nanjing, China, in 2017. She is currently an Assistant Researcher with the Department of Computer Science and Technology, Nanjing University. Her current research interests include machine learning and computer vision, with a focus on subspace learning, adversarial learning and their applications to heterogeneous face recognition and cross-modal face generation.



Yu-Kun Lai received the bachelor's and PhD degrees in computer science from Tsinghua University, in 2003 and 2008, respectively. He is currently a professor in the School of Computer Science and Informatics at Cardiff University. His research interests include computer graphics, geometry processing, image processing, and computer vision. He is on the editorial boards of *Computer Graphics Forum* and *The Visual Computer*.

Yang Gao (M'05) received the Ph.D. degree in computer software and theory from Nanjing University, Nanjing, China, in 2000. He is a Professor with the Department of Computer Science and Technology, Nanjing University. He has published over 100 papers in top conferences and journals. His current research interests include artificial intelligence and machine learning.

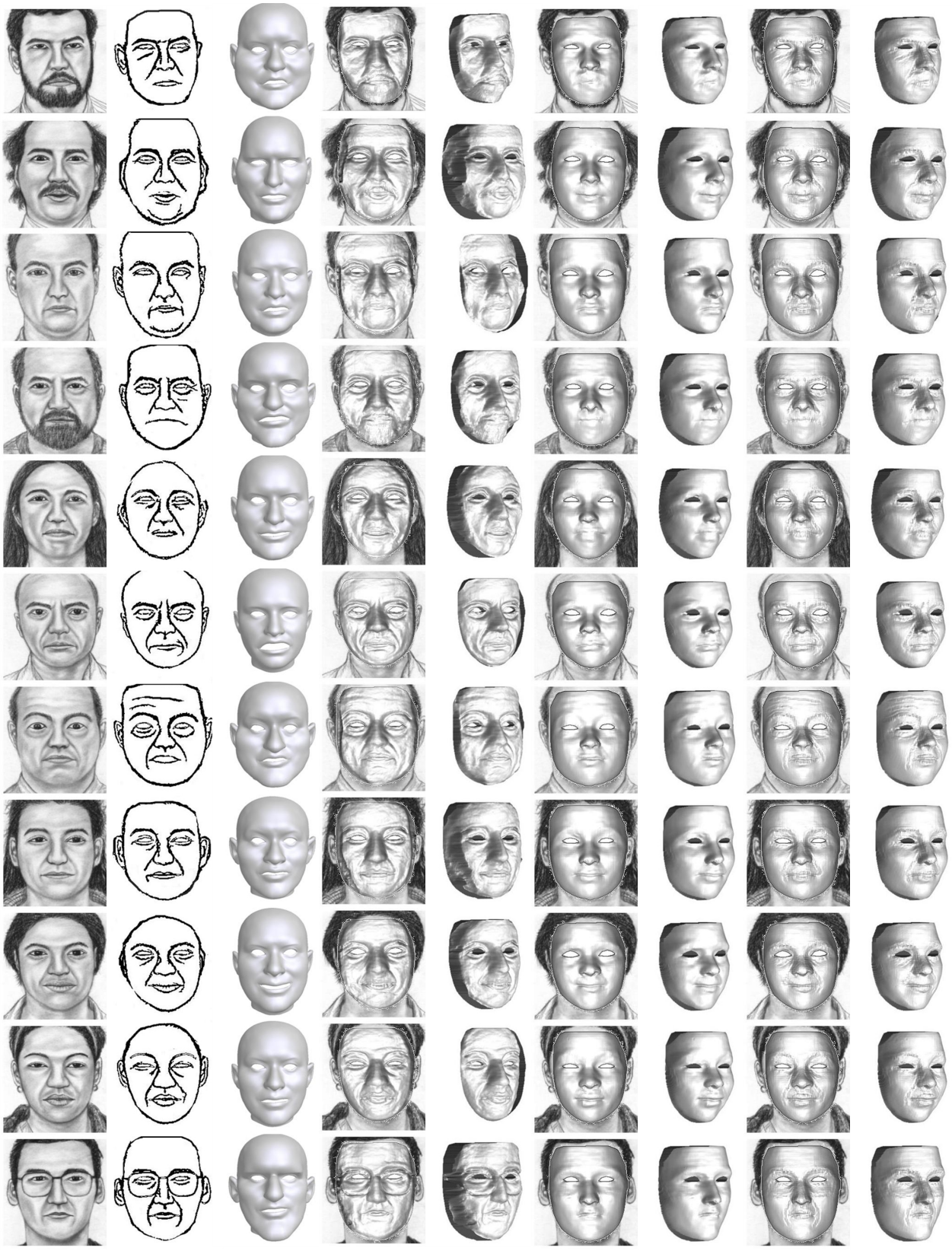


Figure 16: Qualitative comparison between DeepSketch2Face [15], DF^{2}Net [39], our DP-CoarseNet and our FineNet on CUFSF. Please note that Line is the input of [15], Sketch is the input of [39] and our DP-CoarseNet, while Sketch and Line are the inputs of our FineNet.