Hindawi
Security and Communication Networks
Volume 2021, Article ID 9987792, 9 pages
https://doi.org/10.1155/2021/9987792

Research Article

Craniofacial Reconstruction via Face Elevation Map Estimation Based on the Deep Convolution Neutral Network

Yining Hu,[1,2] **Zhe Wang**,[1] **Yueli Pan**,[1] **Lizhe Xie**,[3,4,5] **and Zheng Wang**[1,2]

[1]*School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China*
[2]*Jiangsu Provincial Key Laboratory of Computer Network Technology, Southeast University, Nanjing 211189, China*
[3]*Institute of Stomatology, Nanjing Medical University, Nanjing 210029, China*
[4]*Jiangsu Key Laboratory of Oral Diseases, Nanjing Medical University, Nanjing 210029, China*
[5]*Affiliated Hospital of Stomatology, Nanjing Medical University, Nanjing 210029, China*

Correspondence should be addressed to Yining Hu; hyn.list@seu.edu.cn and Lizhe Xie; xielizhe@njmu.edu.cn

Received 4 March 2021; Accepted 19 May 2021; Published 8 June 2021

Academic Editor: Beijing Chen

In this study, to achieve the possibility of predicting face by skull automatically, we propose a craniofacial reconstruction method based on the end-to-end deep convolutional neural network. Three-dimensional volume data are obtained from 1447 head CT scans of Chinese people of different ages. The facial and skull surface data are projected onto two-dimensional space to generate a two-dimensional elevation map, and then, use the deep convolution neural network to realize the prediction of skull to face shape in two-dimensional space. The encoder and decoder are composed of first feature extraction through the encoder and then as the input of the decoder to generate the craniofacial restoration image. In order to accurately describe the features of different scales, we adopt an U-shaped codec structure with cross-layer connections. Therefore, the output features are decomposed with the features of the corresponding scales in the encoding stage to achieve the integration of different scales while restoring the feature scales in the compression and decoding stage. Meanwhile, the U-net structures help to avoid the problem of loss of detail features in the downsampling process. We use supervised learning to obtain the prediction model from skull to facial elevation map. Back-projection operation is performed afterwards to generate facial surface data in 3D space. Experiments show that the proposed method in this study can effectively achieve craniofacial reconstruction, and for most part of the face, restoration error is controlled within 2 mm.

1. Introduction

Craniofacial reconstruction is a technique producing a reconstructed face from a human skull. Based on the relationship between the skull and face in forensic medicine, anthropology, and anatomy, this technique has been widely used in criminal investigation and archaeology. The traditional craniofacial reconstruction is mainly implemented manually by experts, based on the anatomical law of the human head and face on the plaster model of the victim's skull and according to the relationship between the soft tissue of the human head and face and the morphological characteristics of the face and skull. The facial appearance of the victim is gradually reproduced with adding rubber clay and other materials. This method usually requires a complicated process, high cost, and time-consuming. In addition, the result largely depends on the practitioner's experience, so its application in criminal investigations based on timeliness and truthfulness is greatly restricted.

With the development of computer visualization and virtual three-dimensional technology, computer-aided craniofacial reconstruction technology has greatly reduced the repair time and work difficulty and reduced the subjective deviation factors, which has attracted widespread attention. The current reconstruction methods are based on either template [1] or feature points [2, 3]. For the template-based methods, a face template set in advance is required. In the reconstruction process, the template is deformed according

to the shape of the skull until the feature points on the face template match with the feature points estimated from the skull. Reconstruction can be based on fixed templates [4–7] or dynamic templates [8–14]. The feature points-based methods first estimate the soft tissue thickness of the facial key points and then restores the facial surface. Although feature points based methods have been practically applied in the field of forensics, there are still limitations, which are mainly reflected in two aspects. First, in the process of recovering complete face surface from sparse feature point information, the loss of facial details will be inevitable. Second, human interaction is often required to ensure the accuracy of feature points positioning, which result in an extra anthropic factor.

Craniofacial reconstruction is essentially a problem of sample generation based on reference data. With the rapid development of deep learning technology, data generation based on the convolutional neural network shows significant advantages, among which the representative technologies are the variational autoencoder (VAE) [15, 16] and generative adversarial network (GAN) [17]. Both VAE and GAN attempt to learn the mapping of hidden space variables to real data distribution through training samples. The difference is that VAE calculates the mean and variance of samples through the neural network, constrains them to obey standard normal distribution, and then samples out hidden variables for reconstruction [18]; while, the GAN adopts the idea of game theory and directly measures the distance between real distribution and generated distribution through the discriminator, forcing the generator to generate a more realistic distribution. In recent years, the GAN has received extensive attention from the industry, and many variants have been derived, such as the WGAN [19], CGAN [20], Pix2Pix [21], and BEGAN [22].

The convolution neural networks have also been introduced into the field of craniofacial reconstruction. Li et al. [23] proposed to use a convolutional neural network based on a codec structure, which can well predict the distribution of skeleton soft tissue. The method is with high computation cost, and high performance hardware requirements are also needed, but the generated results are not satisfying. Yuan et al. [24] used the GAN to reconstruct 3D face images. Limited by the data amount and computing power, the author use sparse representation of 3D data to reduce the computation cost and improve the recovery ability; Liu and Xin [25] proposed a prediction method based on the autoencoder and GAN. Candidate faces are generated through the autoencoder. The human face and skull are superimposed to determine the best face. The GAN is used afterwards to optimize the results. Such scheme is essentially a deep learning version of the template-based method. Although the reconstruction accuracy is relatively high, the common problem of the template method is inevitable, that is, the generation process is cumbersome, and the network structure is complex.

Based on the above research, we propose an end-to-end facial morphology prediction method based on the deep convolutional neural network to automatically estimate face information from skull data. For the proposed method, named cylindrical facial projection residual net (CFPRN), it needs neither preset face template nor feature point detection. In order to ignore unnecessary calculations, we do not reconstruct the face data directly in 3D space but try to estimate the face elevation map in 2D cylindrical projection space, and back-projection operation is performed afterwards to get the 3D face surface. We use U-shape network structure so as to adapt with features of different scales. The CFPRN is easy to implement, and experiments have verified the robustness and accuracy of the proposed method.

2. Data Preprocessing

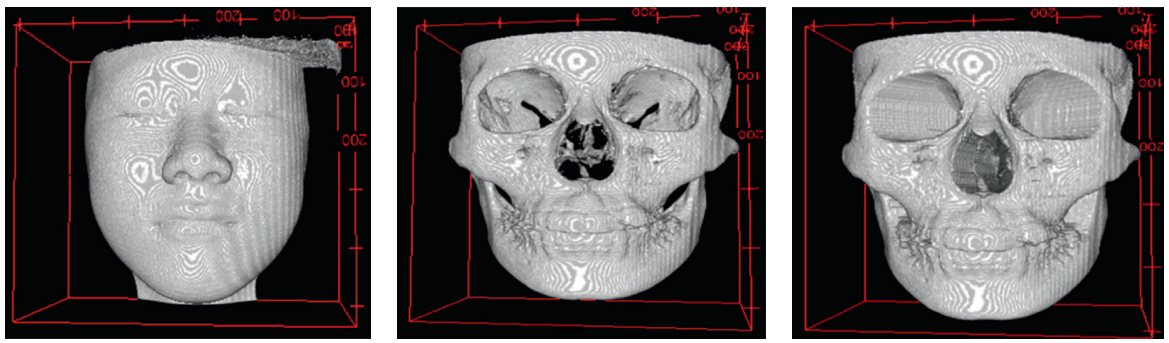
2.1. Data Segmentation. The objective of craniofacial reconstruction is to recover the 3D face surface data from 3D skull data. Both data are obtained from 3D head CT scan. The face surface can be simply retrieved via threshold segmentation, as shown in Figure 1(a); however, due to the complexity distribution of soft tissue and cartilage, threshold segmentation is not suitable for the skull. In order to obtain a clean skull structure, we choose to use adaptive threshold segmentation with a sliding window. The size of the sliding window is set to be 7*7*7. The comparison of global and adaptive thresholding is shown in Figures 1(b) and 1(c).

2.2. Projection and Back-Projection. For craniofacial reconstruction task based on convolutional neural networks, the 3D volume data obtained via head CT scans are usually with excessive data volume [26]. The existing hardware conditions are difficult to meet the problem of constructing a feature network directly for 3D data under the original resolution. In fact, during the reconstruction, only the surface of the skull and the face needs to be considered. Therefore, we use projection operations to map the 3D data to the 2D space for calculation. Considering that the human head is close to a circle in the cross-section and in order to avoid the inconsistency of the resolution in the vertical axis, we use a cylindrical projection surface. The plane projection and sphere projection are not considered because the former leads to inconsistent resolution in vertical direction, and the latter results in inconsistent resolution in different horizontal slices. As shown in Figure 2(a), the cross-section of the CT scan is the XOY plane, and the Z-axis is perpendicular to the cross-section. The coronal plane and the sagittal plane are the XOZ plane and the YOZ plane, respectively. Figure 2(b) shows the projected plane coordinate system.

The coordinate transform between 3D space and the cylindrical projection plane is defined as follows. For projection,

$$\begin{aligned} u &= \arctan\left(\frac{x'}{y'}\right) * \frac{n}{\pi} = \alpha * \frac{n}{\pi}, \\ v &= z', \\ r &= \sqrt{\left(x'^2 + y'^2\right)}. \end{aligned} \tag{1}$$

For back-projection,



(a) (b) (c)

Figure 1: Result of adaptive threshold segmentation. (a) 3D face image. (b) 3D skull via thresholding. (c) 3D skull via adaptive thresholding.

$$x' = r\ \sin\left(\frac{u\pi}{2n}\right),$$
$$y' = r\ \cos\left(\frac{u\pi}{2n}\right), \quad (2)$$
$$z' = v,$$

where x', y', z' are the coordinates from the 3D space, and u, v are the coordinates from the projection plane. r is the pixel value of the 2D projected altitude map which represents the distance from the point to the projection axis in 3D space. Thus, the depth information in 3D facial and skull surface is preserved in the projection and back-projection steps. $2n$ is the total sample number in U axis.

3. Network Architecture

The network structure refers to the encoder-decoder structure of U-net [27] and draws on the relevant ideas of the CGAN [20], Pix2Pix [21], and other networks to realize an end-to-end network.

In the encoder-decoder structure, the first half of the network acts as an encoder, which successively is downsampling through pooling, convolution with strides, to extract deep features from the input image. The second half of the network acts as a decoder, which successively is upsampling through deconvolution, interpolation, to map the feature output by the encoder back to the size of the previous level. In the meantime, cross-layer connection is considered, so that the high-level feature map after being upsampled by the decoder and the low-level feature map of the same scale in the encoder are connected in the channel dimension, and feature information of different scales are merged to make the prediction result more accurate and stable. Figure 3 shows the specific structure of the proposed network.

The network is generally divided into two parts: encoder module and decoder module.

The encoder module is mainly composed of a convolutional layer and five convBlocks; each convBlock, as shown in the bottom right of Figure 3, contains a leaky Relu activation layer, a 3*3 convolutional layer, and a group normalization layer. The encoder module performs 6 downsampling in total, and the pooling operation is replaced by a convolution operation with a step size of 2 so as to retain more feature information.

The decoder module is composed of five deconvBlocks and a convolutional layer. Each deconvBlock, as shown in the bottom right of Figure 3, contains a leaky Relu activation layer, an upsampling layer, a 3*3 convolutional layer, and a group normalization layer. The decoder module performs upsampling 6 times in total; bilinear interpolation is considered for upsampling, expanding the height and width of the feature map by 2 each time. The feature map after each upsampling is connected in the channel dimension with the feature map of the corresponding scale in the encoder. Through such a cross-layer connection, the deep and shallow features can be effectively merged.

In the meantime, we use some tricks to improve the performance of the entire network. (i) Replace deconvolution with a structure of upsampling using bilinear interpolation and convolution, which can effectively avoid the checkerboard effect [28]. (ii) Replace Relu with leaky Relu, which can effectively reduce the dead neurons. Replace pooling operation with convolution operation with a step size of 2 to retain more features. (iii) Use group normalization [29] instead of batch normalization which can effectively avoid the impact of batch size on the training results.

We use normalized skull elevation map as network input. The data range is limited to (−1, 1). The normalization can speed up the convergence of the network and increase the generalization ability of the model. For the supervised data, we have 2 options: one is to use the face elevation map directly and the other is to use the residual between the face and skull surface (mentioned as "face" and "res," respectively, in the experiment section).

The loss is defined as the distance between predicted and real face elevation map. We use mean square error

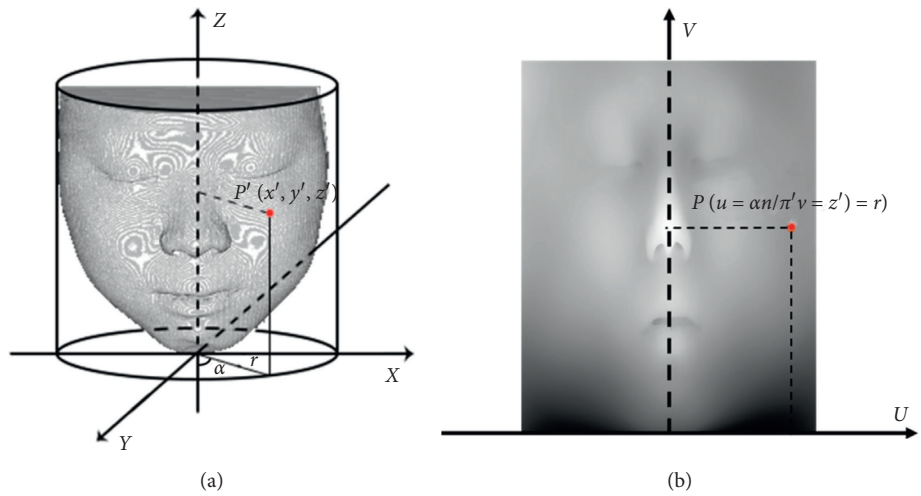


Figure 2: Cylindrical projection. (a) 3D face image before projection. (b) 2D face image after projection.

(MSE) to define the loss function, which represents the average value of the square of the difference between the predicted and the real elevation map. The expression is as follows:

$$\mathrm{MSE}(x, y) = \frac{1}{m}\sum_{i=1}^{m}(x_i - y_i)^2, \tag{3}$$

where m denotes the number of pixels, and the terms x, y denote the predicted and the label value, respectively.

4. Experiment

4.1. Data Description. The dataset used for experiment is acquired from the head cone-beam CT scan from NewTom 5G. The dataset contains CT data of 1447 participants from Affiliated Hospital of Stomatology, Nanjing Medical University. Each sample has 540 CT slices, the resolution for each slice is 610×610, and the pixel size is 0.3 mm*0.3 mm. 1310 samples were randomly selected as training set, and the validation set is composed of the rest 137 samples.

4.2. Evaluation Indices. Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [30] is chosen as evaluation indices for the experiments. The peak signal-to-noise ratio (PSNR) measures the ratio between the energy of the peak signal and the average energy of the noise, which is commonly used for signal recovery quality. The PSNR is defined as

$$\mathrm{PSNR} = 10\log_{10}\frac{\mathrm{MAX}^2}{\mathrm{MSE}}, \tag{4}$$

where MAX denotes the maximum pixel value in the data, and MSE is the mean square error. Besides PSNR, SSIM is considered for the similarity measure between the ground truth and the predictions. The SSIM measures image similarity from three aspects of brightness, contrast, and structure, with value range (0, 1), and larger value stands for smaller image distortion.

5. Results and Discussion

5.1. Result. We intuitively visualized the experimental results. Figure 4 shows the input elevation map of the skull. Figure 5 shows that the prediction results of the face elevation map correspond to skulls in Figure 4 and the corresponding ground truth. We can see the predicted result is very close to the ground truth. Pseudocolour maps shown in Figure 6 visualize the difference between the output and the ground truth (in percentage), from which we may see that the error mainly occurs in the eyes, nose, and mouth area. Obviously, because of the cavity in the skull, it is impossible to accurately predict the eyes and nose.

We use the predicted elevation map to generate 3D facial data through back-projection. The generated 3D face is compared with ground truth. The difference map is shown in Figure 7, from which we may see that for most part, the error is limited to 1 mm.

5.2. Comparison. We have repeated experiments on different network architectures and different image sizes. The specific results are given in the table, and the bold line is our proposed one. Table 1 indicates that the proposed CFPRN is with high accuracy and shows best performance among all the candidates. The abbreviation "Res" means the network output is the residual of the face and skull, and the abbreviation "Face" means the network output is the face surface directly. Table 2 indicates that the CFPRN works well under different resolution settings.

5.3. Error Analysis

(1) In order to simplify the network and improve efficiency, we reduce the dimension of the input, which causes a partial loss of data accuracy. After the prediction is

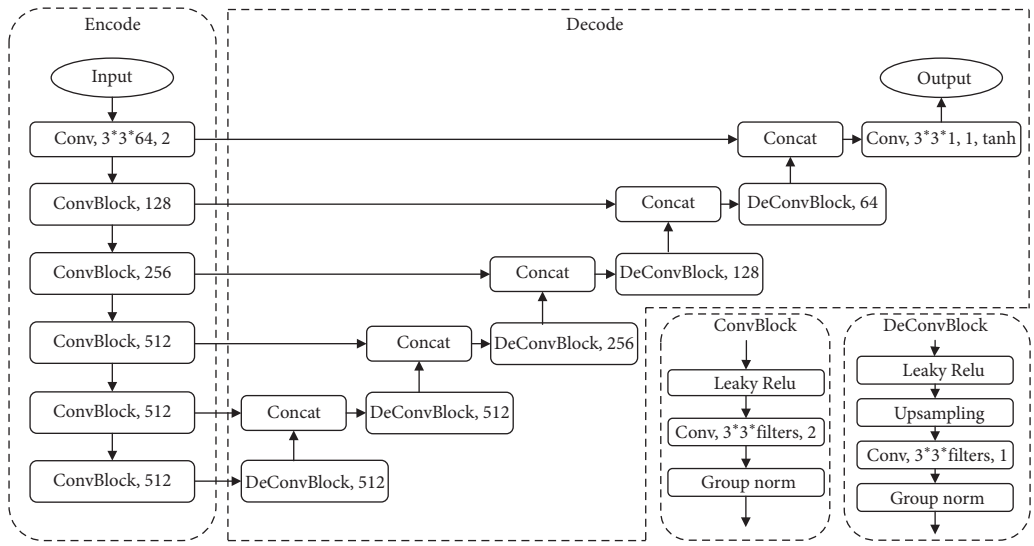
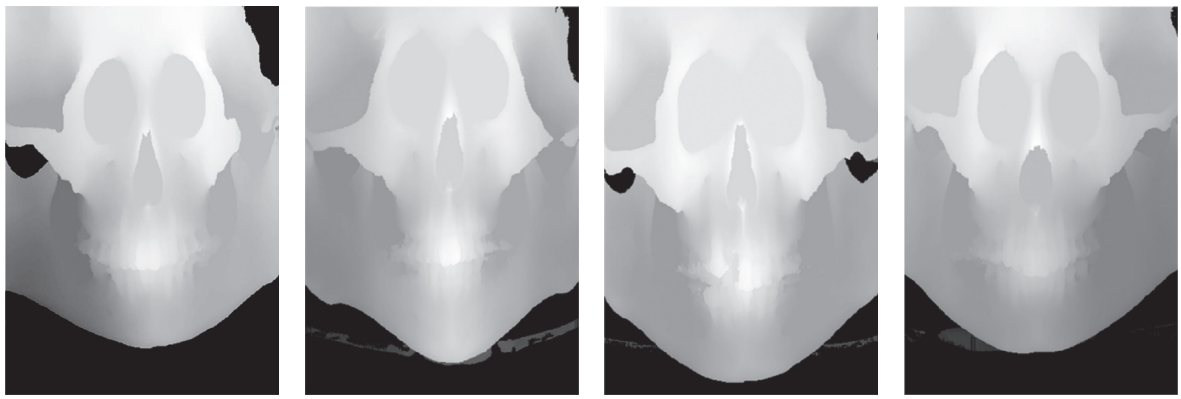
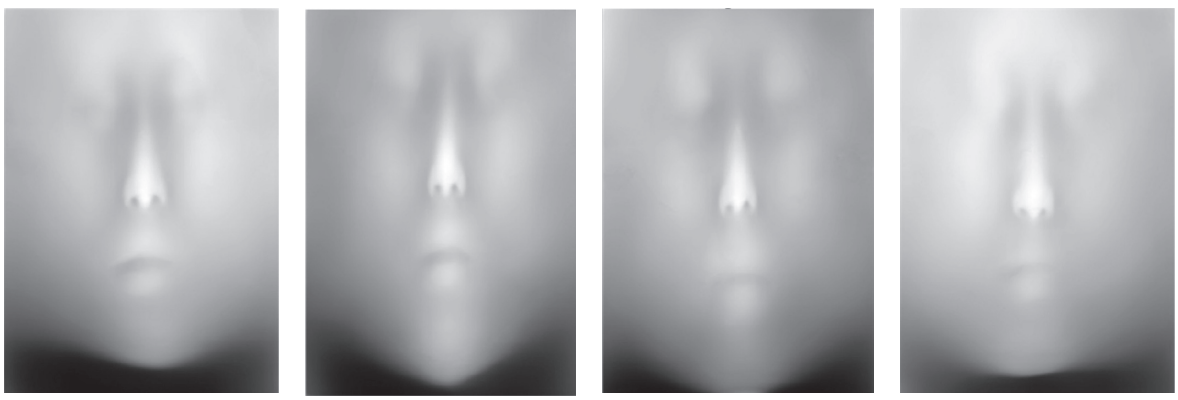


Figure 3: Network structure.



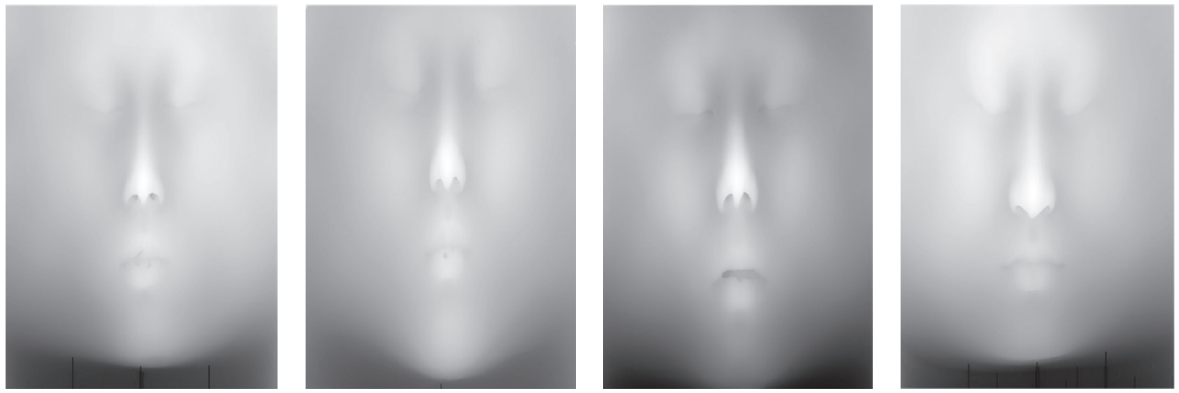
(a) (b) (c) (d)

Figure 4: Input skull images. (a) Skull 1. (b) Skull 2. (c) Skull 3. (d) Skull 4.



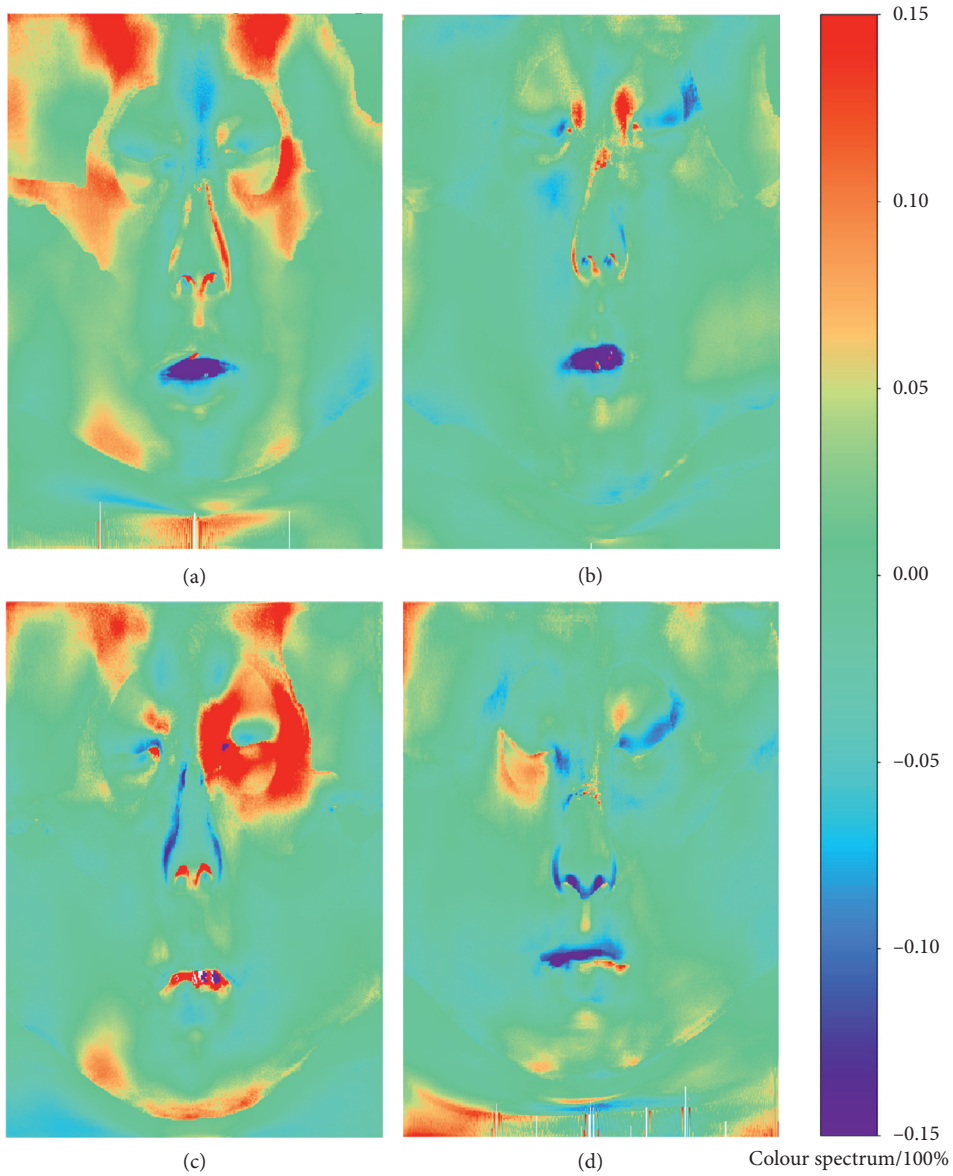
(a) (b) (c) (d)

Figure 5: Continued.



(e) (f) (g) (h)

Figure 5: Comparison between label face and predicted face. (a) Prediction face 1. (b) Prediction face 2. (c) Prediction face 3. (d) Prediction face 4. (e) Label face 1. (f) Label face 2. (g) Label face 3. (h) Label face 4.



(a) (b) (c) (d)

Figure 6: 2D difference maps. (a) Difference map of face 1. (b) Difference map of face 2. (c) Difference map of face 3. (d) Difference map of face 4 colour spectrum/100%.

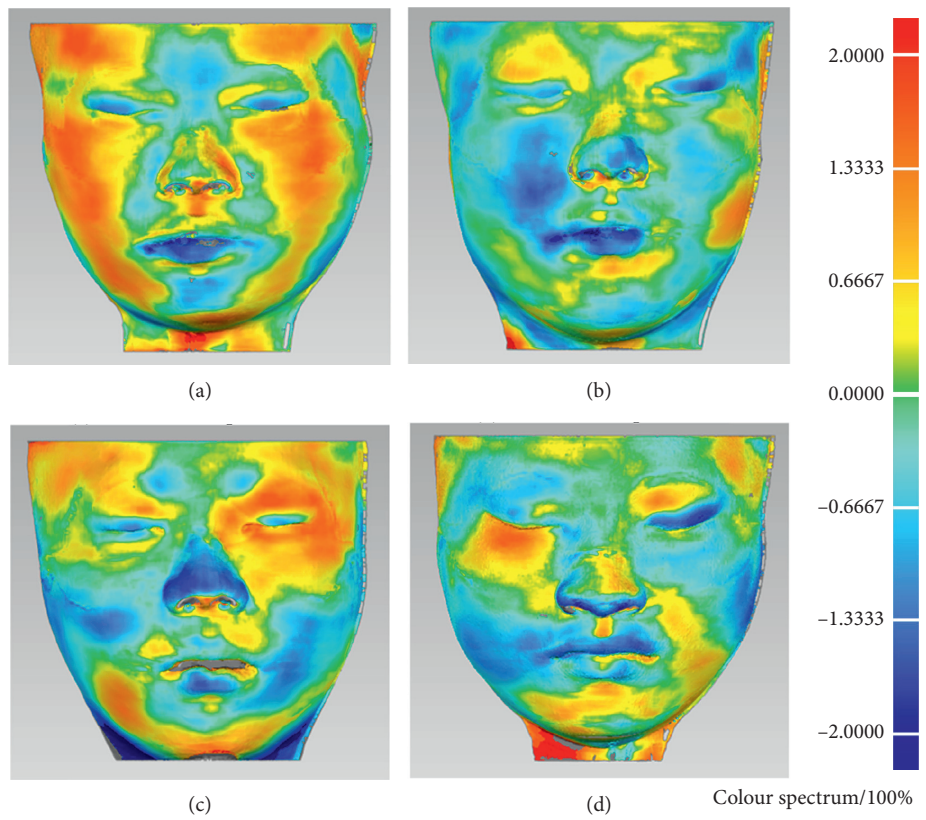


Figure 7: 3D difference maps. (a) Difference map of face 1. (b) Difference map of face 2. (c) Difference map of face 3. (d) Difference map of face 4 colour spectrum/mm.

Table 1: Results of different network architectures.

Network	RMSE	PSNR	SSIM
Inception ResNet U, Res	11.245456	31.477464	0.979727
Res, Pix2Pix, group norm	10.770702	31.938354	0.968041
Face, Pix2Pix, group norm	**10.16162**	**32.424038**	**0.987716**
Face, Pix2Pix, batch norm	10.724989	31.925444	0.986937

Table 2: Results of different network generation sizes.

Size	RMSE	PSNR	SSIM
Pix2Pix, 128*90	10.258215	32.379723	0.990736
Pix2Pix, 256*180	10.348463	32.392857	0.990038
Pix2Pix, 512*360	**10.16162**	**32.424038**	**0.987716**

completed, back-projection is performed, which may also cause extra error transfer.

(2) From the error map, we may see that basically, all samples have large errors in the part of the nose and eyes. This is because the skull has holes in the eyes and nose, which cannot be accurately predicted, and this might be overcome by introducing much more samples.

6. Conclusion and Prospects

In this study, we propose an end-to-end deep learning method for craniofacial reconstruction. The main contribution of the proposed method can be summarized as follows:

(1) We use projection and back projection for the transfer between 3D skull and face data into 2D elevation map. Instead of performing craniofacial

reconstruction in 3D space, the recovery runs in 2D space. The face elevation map is estimated according to the skull elevation map. Such design largely reduces the data size and computation cost, so that the proposed method is available on consumer graphics cards.

(2) We design an U-shaped end-to-end network to fit for the features in different scales. The accuracy and robustness of the prediction are guaranteed according to the experiment results.

According to our experiment results, we can also make further prospects:

(1) We should expand the amount of samples. Divide samples according to gender and age to balance the distribution of sample data.

(2) The eyes and nose of the skull should be hollowed out or filled. Because the specific shape of the face in these parts cannot be inferred from the skull, it is helpful to reduce the impact on the experimental results by hollowing out or filling these parts.

(3) We will try other network architectures, such as the conditional GAN. By introducing more conditions, we may provide subdivided predictions with higher accuracy.

Data Availability

All the experiment data are obtained from Affiliated Hospital of Stomatology, Nanjing Medical University. The access to the original data is restricted due to the patient privacy. However, the data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] P. Vanezis, M. Vanezis, G. McCombe, and T. Niblett, "Facial reconstruction using 3-D computer graphics," *Forensic Science International*, vol. 108, no. 2, pp. 81–95, 2000.

[2] Q. H. Dinh, T. C. Ma, and T. D. Bui, *Facial Soft Tissue Thicknesses Prediction Using Anthropometric Distances*, Springer, Berlin, Heidelberg, 2011.

[3] P. Guyomarc'H, B. Dutailly, C. Couture et al., "Anatomical placement of the human eyeball in the orbit--validation using CT scans of living adults and prediction for facial approximation," *Journal of Forensic Sciences*, vol. 57, no. 5, pp. 1271–1275, 2012.

[4] A. J. Tyrrell, M. P. Evison, A. T. Chamberlain et al., "Forensic three-dimensional facial reconstruction: historical review and contemporary developments," *Journal of Forensic Sciences*, vol. 42, no. 4, p. 653, 1997.

[5] M. W. Jones, *Facial Reconstruction Using Volumetric Data*, Aka GmbH, Augsburg, Germany, 2001.

[6] G. Quatrehomme, S. Cotin, G. Subsol et al., "A fully three-dimensional method for facial reconstruction based on deformable models," *Journal of Forensic Sciences*, vol. 42, no. 4, pp. 649–652, 1997.

[7] L. A. Nelson and S. D. Michael, "The application of volume deformation to three-dimensional facial reconstruction: a comparison with previous techniques," *Forensic Science International*, vol. 94, no. 3, pp. 167–181, 1998.

[8] M. Berar, M. Desvignes, G. Bailly et al., "Statistical skull models from 3D X-ray images," 2006, http://arxiv.org/abs/0610182.

[9] M. Desvignes, G. Bailly, Y. Payan et al., "3D semi-landmarks based statistical face reconstruction," *Journal of Computing & Information Technology*, vol. 14, 2006.

[10] P. Claes, D. Vandermeulen, S. De Greef, G. Willems, and P. Suetens, "Craniofacial reconstruction using a combined statistical model of face shape and soft tissue depths: methodology and validation," *Forensic Science International*, vol. 159, pp. S147–S158, 2006.

[11] P. Claes, D. Vandermeulen, S. D. Greef et al., "Statistically deformable face models for cranio-facial reconstruction," in *Proceedings of the Ispa International Symposium on Image & Signal Processing & Analysis*, IEEE, Zagreb, Croatia, September 2006.

[12] P. Claes, D. Vandermeulen, S. D. Greef et al., "Bayesian estimation of optimal craniofacial reconstructions," *Forensic Science International*, vol. 201, no. 1–3, pp. 146–152, 2010.

[13] P. Paysan, M. Lüthi, T. Albrecht et al., "Face reconstruction from skull shapes and physical attributes," in *Proceedings of the Symposium of the German Association for Pattern Recognition (DAGM 2009)*, pp. 232–241, Springer, Jena, Germany, September 2009.

[14] Y. Hu, F. Duan, B. Yin et al., "A hierarchical dense deformable model for 3D face reconstruction from skull," *Multimedia Tools and Applications*, vol. 64, no. 2, pp. 345–364, 2013.

[15] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," 2014, http://arxiv.org/abs/1401.4082.

[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2014, http://arxiv.org/abs/1312.6114.

[17] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza et al., "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, pp. 2672–2680, 2014.

[18] C. Doersch, "Tutorial on variational autoencoders," 2016, http://arxiv.org/abs/1606.05908.

[19] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," 2017, http://arxiv.org/abs/1701.07875.

[20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *Computer Science*, vol. 6, pp. 2672–2680, 2014.

[21] P. Isola, J. Y. Zhu, T. Zhou et al., "Image-to-Image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*, IEEE, Seattle, WA, USA, June 2016.

[22] D. Berthelot, T. Schumm, and L. Metz, "BEGAN: boundary equilibrium generative adversarial networks," 2017, http://arxiv.org/abs/1703.10717.

[23] X. Li, B. Sheng, L. Ping et al., "Voxelized facial reconstruction using deep neural network," in *Proceedings of the Computer Graphics International*, New York, NY, USA, June 2018.

[24] Y. Yuan, Y. Zhang, S. Wang et al., "Sparse representation-based face object generative via deep adversarial network," in *Proceedings of the 2018 7th International Conference on Digital Home (ICDH)*, Guilin, China, December 2018.

[25] C. Liu and L. Xin, "Superimposition-guided facial reconstruction from skull," 2018, http://arxiv.org/abs/1810.00107.

[26] F. Tilotta, F. Richard, J. Glaunès et al., "Construction and analysis of a head CT-scan database for craniofacial reconstruction," *Forensic Science International*, vol. 191, no. 1–3, 2009.
[27] O. Ronneberger, P. Fischer, and T. Brox, *U-net: Convolutional Networks for Biomedical Image Segmentation*, Springer, Berlin, Germany, 2015.
[28] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, 2016.
[29] Y. Wu and K. He, "Group normalization," *International Journal of Computer Vision*, vol. 14, 2018.
[30] Z. Wang, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 41, 2004.