2023 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-6891-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICME55011.2023.00159

Robust 3D Craniofacial Landmarks Localization by An End-to-End Regression Network

Xianhe Jiao[a], Junli Zhao[a*], Chenlei Lv[b*], Fuqing Duan[c], Zhenkuan Pan[a], Xin Li[d]
[a] Qingdao University, China [b] Shenzhen Univetsity, China
[c] Beijing Normal University, China [d] Texas A&M University, America
jiaoxianhe@gmail.com, zhaojl@yeah.net, chenleilv@mail.bnu.edu.cn,
fqduan@bnu.edu.cn, zkpan@qdu.edu.cn, xinli@tamu.edu

***Abstract*—Landmark localization plays a significant role in craniofacial registration, reconstruction, and authentication. The key challenges for localizing landmarks on point cloud craniofacial models include irregular structures, non-uniform densities, and uncertain local regions. In this paper, we propose an end-to-end regression network that can directly estimate craniofacial landmarks on point cloud models. The proposed network utilizes edge convolution to extract local features and pooling layers to aggregate global features. It realizes the end-to-end regression for landmark localization. Experimental results demonstrate that our method is robust on point clouds with sparse and unevenly distributed sampling. It can produce accurate, controllable, and efficient 3D landmarks.**

***Index Terms*—Craniofacial landmark localization, regression network, edge convolution, sparse point cloud**

I. Introduction

Craniofacial landmarks are defined by physiological features according to the structural characteristics of the skull and muscular tissues of the head, such as eye corners, nose tip, mouth corners and facial contour points. The localization is to automatically identify skull and skin landmarks on the craniofacial model according to the basic theory of anatomy and forensic anthropology. Such landmarks lay a foundation for the subsequent craniofacial morphological research and take a significant role in related applications, including craniofacial registration, reconstruction, and authentication.

2D landmark localization has achieved high accuracy and robustness with the development of deep learning [2]. To locate 3D landmarks, some recent studies try to convert 3D model into 2D domains [7], [14], [16] which can utilize the advantages of 2D landmark localization. However, dimensional reduction usually results in information loss. Some points of the 3D model would be mapped into the same 2D region that produces inverse projection ambiguous [17]. Therefore, directly locating landmarks on 3D models is becoming a research trend that can avoid the problem.

There are some challenges for direct 3D landmark localization, including irregular structures of 3D models, non-uniform point densities, and uncertain local regions. The challenges make the effective conduction of spatial convolution difficult. Some physiological salient feature landmarks locate in uncertain local regions that are not geometrically salient. Consequently, most existing 3D craniofacial landmark localization methods detect feature points based on local geometric information. They are sensitive to point densities and can only locate geometrically salient landmarks (such as nose tip and eye corners) but often fail to identify other landmarks [12] without significant geometric features.

In order to solve the problems, we propose a novel regression network that can estimate landmarks on 3D craniofacial point cloud models directly. By performing edge convolution and pooling layers to aggregate the local and global features on the models, the network ensures the accuracy of prediction for craniofacial landmarks. It can directly regress seventy-eight (78) craniofacial physiological landmarks (that are widely used in forensic tasks) from the point cloud model with non-uniform density. Furthermore, our method can also allow the users (forensic specialists) to specify different numbers of other landmarks relevant to their requirements (e.g., for different craniofacial registration or reconstruction tasks). The main contributions of this paper are as follows:

- We present an end-to-end regression network that can automatically locate landmarks on 3D craniofacial point clouds in order to avoid the problem of information loss.
- We employ the edge convolution structure and pooling layers to extract local and global features from the point cloud to quickly locate standard or customized feature landmarks.
- We will release a pre-trained model that can be used in many forensic tasks that need accurate landmark localization from sparse and unevenly distributed point cloud models. It is robust to non-uniform densities.

II. Related Work

Landmark localization has always been a hot topic in the field of computer vision and computer graphics. Many effective methods have been summarized by Kostiantyn [1] in recent years. In this paper, we mainly focus on the works of 3D landmark localization, which can be discussed from two aspects: geometric methods and learning methods.

* Junli Zhao and Chenlei Lv are corresponding authors.

This work was supported by the National Natural Science Foundation of China under Grant Nos.62172247,61702293, Natural Science Foundation of Shandong Province (No.ZR2019LZH002).

1945-788X/23/$31.00 ©2023 IEEE
DOI 10.1109/ICME55011.2023.00159

A. 3D landmark localization with Geometric Analysis

In methods of geometric analysis, landmarks are defined according to geometry features, such as curvature and local shape descriptors. Some methods [9], [10] employ anthropometric statistic tools to code prior knowledge and locate landmarks such as nose tip and eye corners. Gilani et al. [5], [8] proposed an automatic landmark localization method by registering the reference face to the target one. Cheng et al. [11] divided the 3D face landmark localization into two steps, including depth map converting for landmarks coarse localization and shape index-based geometric analysis. Enrico et al. [12] performed landmark localization by calculating 12 geometric descriptors, including Gaussian curvature, mean curvature, and shape index for each point of the 3D model to analyze the facial shape and implement landmark localization. By geometric analysis, most landmarks with salient geometric features can be accurately calibrated without complex training and estimation. However, the computation is complicated, and the number of landmarks can not be controlled. Generally, the localization is limited to the quality of local geometric features.

B. 3D landmark localization with Neural Network

Considering successful experiences of deep learning in the field of computer vision, another kind of localization method utilize mature neural network technologies to detect 2D or 3D landmarks in related 2D representation. Gao et al. [14] proposed a 3D facial landmark localization network using regression networks 3DLLN(3D landmark localization network), which uses the location map as an intermediate representation and detects 3D landmarks coordinates from it. Zhang et al. [16] and Xu et al. [7] converted 3D face into depth image and predicted 2D landmarks with a 2D convolutional regression network. When a single 2D depth map is used to represent the 3D model, many points in the lateral part of a face concentrate in the same pixel grid due to the viewpoint, which results in information loss after the dimensional mapping. To solve the problem, Terada et al. [6] and Zhang et al. [17] improved the 2D depth representation with cylindrical projection to represent the 3D model. Then, they predicted 2D or 3D landmarks by improved Resnet-based networks, respectively. As mentioned before, information loss is inevitable when the 3D model is converted to related a 2D representation. Generally, landmarks calibration directly on the 3D model can obtain more accurate results than 2D representation-based localization. Eimear [15] used PointNet for feature extraction of 3D models and extended convolutional pose machine for 3D landmarks coordinate regression. However, it is sensitive to non-uniform densities and uncertain local regions.

We proposed an end-to-end regression network that can automatically locate 3D landmarks on craniofacial point cloud models directly. It avoids the problem of information loss. The network extracts global and local features of the point cloud for complicated models with skull and skin data while realizing the control of the relative distance between landmarks from the global perspective. It ensures the accuracy of prediction for landmarks and achieves robustness to handle the mentioned challenges.

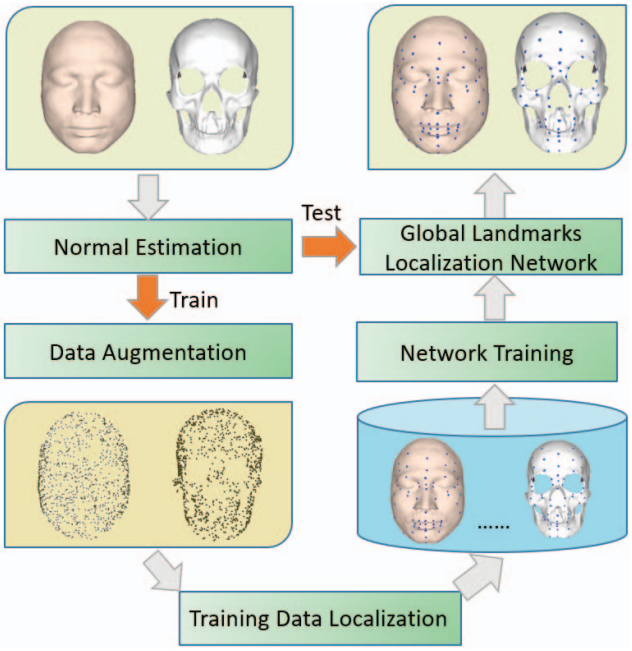


Fig. 1. The pipeline of landmark localization based on regression network

III. Method

The proposed regression network can directly estimate 3D landmarks on craniofacial point cloud models, as shown in Fig.1. Firstly, we introduce the details of model scanning and related calculations as the pre-processing. Then, we introduce the implementation of the end-to-end regression network with edge convolution and pooling layers. Based on the network, we can directly obtain the craniofacial landmarks by the point cloud coordinates and normal vectors.

A. Pre-processing

1) Craniofacial Point Cloud Reconstruction from CT: In this study, our craniofacial models came from a database of 208 whole-head CT scans. The raw CT slice images were processed by filtering the noise, and the cranial and facial boundaries were extracted with the Sobel operator. Then, the Marching Cubes algorithm [18] was used to reconstruct the 3D skull and face, which was subsequently simplified to a point cloud of about 40k vertices. Finally, we convert all 3D craniofacial data to a unified Frankfurt coordinate system [19], [20] to eliminate the effects of data acquisition, pose and scale.

2) Ground-truth of Craniofacial Landmarks: Craniofacial landmarks include skull landmarks and facial ones. The skull landmarks are obtained by forensic experts according to the physiological structure of a skull, and the most representative points of each bone structure are used as landmarks. The facial landmarks are mapped from the skull landmarks according to the soft tissue thickness to obtain the corresponding facial

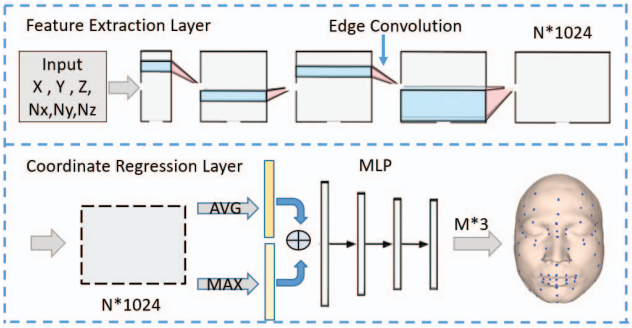


Fig. 2. The End-to-End Regression Network. The upper part is the feature extraction layer, which extracts the local features through edge convolution. The lower part is the coordinate regression layer, which regresses the landmarks by the pooling layers and multi-layer perception.

landmarks. Seventy-eight (78) physiological craniofacial landmarks are used for training and testing in this paper (as shown in Fig.1).

3) Data Augmentation: The normal vector is an indispensable descriptor in traditional geometric analysis methods and plays a significant role in deep learning. To enhance the geometric features contained in the craniofacial point cloud model, we estimate the normal vector for each point based on the local surface.

Since the proposed network has no restriction on the number of points, it is possible to locate the landmarks of an arbitrary number of points after obtaining a pre-trained model. Also, to augment the training data and speed up the network training, we randomly sample each craniofacial point cloud model for k times. The sampling points are set to 1024, as shown in Fig.1. After experimental testing, we set $k = 5$ to balance speed and accuracy.

B. End-to-End Regression Network

We design our network from global and local features. For the local feature-based analysis, each landmark should consist of the most salient geometric features in the neighborhood. We introduce the edge convolution structure [13] to learn the local features contained in the neighborhood for each point. For the global feature-based analysis, the landmarks can be regarded as a sparse set of points to represent the structural information of the whole craniofacial surface. At the same time, the global features can describe the general information of the model by a little feature. Therefore, we extract and utilize the global features of the craniofacial point cloud model to regress the coordinates of the landmarks.So we divide the network architecture into two parts: the feature extraction layer and the coordinate regression layer.As shown in Fig.2 and Table.I.

1) The Feature Extraction Layer: The feature extraction layer is to compute its local features for each point by the four-layer edge convolution, where the edge convolution is divided into two parts: feature aggregation and convolution.

Feature Aggregation. The directed graph G $= (V, E)$ representing the local point cloud structure, as shown in Fig.3.

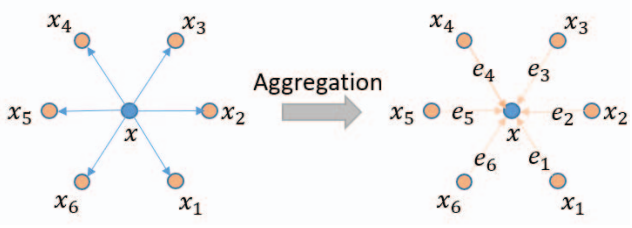


Fig. 3. Feature Aggregation

The $V = \{1, 2, 3, ..., n\}$ and $E \subseteq V \times V$ are the vertices and edges of the local point cloud of the directed graph, respectively. We use the k-nearest neighbor (KNN) graph as a local point cloud structure directed graph, where k is set to 6. The edge feature is defined as

$$e_{ij} = h_\theta(x_i, x_j - x_i), \tag{1}$$

where h_θ is a nonlinear function with a set of learnable parameters θ. After solving all the edge features E for the local point cloud directed graph G, we choose the symmetric aggregation function Max as the edge feature integration function, as shown in Eq.2, and use the extracted features as the features of the local point cloud directed graph centroids x_i, represented as

$$x_i = \underset{j \in (1,n)}{Max} (e_{ij}). \tag{2}$$

Convolution. We obtain n (the number of points) local features after aggregation. Next, we increase the dimension of these features by convolution. The parameters are shown in Table.I. Since we have done the feature aggregation for each point before increasing dimension, so that the local features of each point contain the neighborhood information and lay the foundation for the subsequent global feature aggregation.

TABLE I
THE END-TO-END REGRESSION NETWORK STRUCTURE

Layer	Filter	Input Shape	Output Shape
Conv1	1×1	$1024 \times 1 \times 6$	$1024 \times 1 \times 64$
Conv2	1×1	$1024 \times 1 \times 64$	$1024 \times 1 \times 256$
Conv3	1×1	$1024 \times 1 \times 256$	$1024 \times 1 \times 512$
Conv4	1×1	$1024 \times 1 \times 1024$	$1024 \times 1 \times 1024$
MaxPooling	-	$1024 \times 1 \times 1024$	1×1024
AvgPooling	-	$1024 \times 1 \times 1024$	1×1024
Concat	-	$1 \times 1024 \times 2$	1×2048
Linear1	-	1×2048	1×1024
Linear2	-	1×1024	1×512
Linear3	-	1×512	1×256
Linear4	-	1×256	1×234

2) Coordinate Regression Layer: The coordinate regression layer realizes the aggregation of the extracted local features into global features and the regression of global features to landmarks. After the feature extraction layer, we obtain $n \times 1024$ local feature. Next, we use MaxPooling and AvgPooling layers for feature extraction to form a global description feature of 1×2048. We designed a Multi-Layer Perception

structure with layer-by-layer feature extraction to form $M \times 3$ (M is the number of feature points) features as the coordinates of the landmarks for regression.

The global features can describe the general information of the model with a global view. The features are aggregated from the local features after edge convolution and contain complete neighborhood information. Based on the features, the proposed network can accurately locate landmarks. At the same time, we can adjust the parameter M (number of landmarks) of the output layer to realize the localization of an arbitrary number of landmarks.

3) Loss Function: The loss function design of our network divides into three parts: landmarks coordinate loss, landmarks distance surface loss, and regularization loss, and the weights of different module losses are adjusted according to the experiment.

Landmarks coordinate loss. Calculate the Euclidean distance from the regression landmarks to the ground truth landmarks, as shown in Eq.3, to control the regression landmarks' coordinates to be as close to the ground truth as possible.

$$\mathrm{L}_{dist} = \frac{1}{M} \sum_{m_i \in M} d_G^2(m_i, g_i) \tag{3}$$

Where M is the number of landmarks, m_i is the predicted landmarks, g_i is the ground truth landmarks on the craniofacial surface, and d_G^2 is the distance between the predicted landmarks and the ground truth.

Loss of landmarks from the surface. The distance from the regression landmarks to the craniofacial surface to ensure that the regression landmarks are as close to the craniofacial surface as possible, as shown in Eq.4,

$$\mathrm{L}_{surf} = \frac{1}{M} \sum_{m_i \in M} d_S^2(m_i, S) \tag{4}$$

where m_i is the predicted landmarks, S is the craniofacial surface, and d_S^2 is the nearest distance between the predicted landmarks and the craniofacial surface.

L2 regularization loss. To prevent overfitting, we add L2 regularization loss, as shown in Eq.5

$$L_{regular} = ||\omega||_2^2 \tag{5}$$

Thus, the loss function of the regression network as shown in Eq.6.λ_1,λ_2 and λ_3 are the weights of L_{dist},L_{surf} and $L_{regular}$,respectively.

$$L = \lambda_1 L_{dist} + \lambda_2 L_{surf} + \lambda_3 L_{regular} \tag{6}$$

IV. Experiments

In this section, we conducted an ablation study on our proposed approach to fully understand and evaluate the role and effect of each part for landmarks localization. We test on craniofacial with different sampling to verify the robustness of our method. Finally, we compared the proposed landmarks localization method with other deep learning methods.

TABLE II
LOCALIZATION ERROR WITH DIFFERENT POINTS NUMBERS.

	N=512	N=1024	N=2048	N=16384
NME	0.02712	0.02487	0.02559	0.02610

A. Normalized Mean Error

In this paper, we use the Normalized Mean Error (NME) [21]. The normalized mean error is used to determine the accuracy of landmarks prediction, and the lower the NME, the higher the accuracy. The NME is defined as shown in Eq. 7,

$$NME = \frac{1}{M} \sum_{i}^{M} \frac{||m_i - g_i||_2}{d} \tag{7}$$

where m_i and g_i denoted the ground truth landmarks and predicted landmarks, respectively. $m_i = (x_i, y_i, z_i)$, and the NME is set d as the distance between the outer eye corners, i denotes the ordinal number of the point, and M denotes the number of landmarks.

B. Ablation Study

In this chapter, to more clearly understand the effectiveness of different parts of our method on landmark localization, we conducted an ablation study for our method.

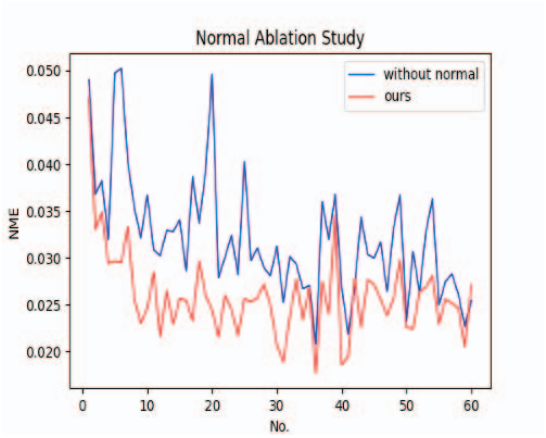
We verify the effectiveness of normal on the landmark localization effect by using only the point coordinates as input to the network and the experimental results as shown in Fig.4(a). In order to verify the effectiveness of edge convolution, we replace the network architecture with the point convolution structure in the PointNet [4] architecture for comparison experiments, and the results are shown in Fig. 5. In summary, the error in landmark localization can be reduced by adding the normal and using the edge convolution module.

C. Robustness on Point Number of 3D Model

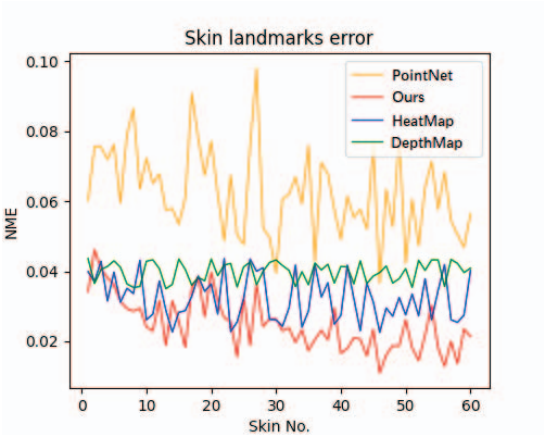
After obtaining the network pre-training model, we can test a point cloud of an arbitrary number of points. We randomly sampled the point cloud with different numbers of points as shown in Fig.6, and by calculating the NME of 60 sets of models, the results are shown in Table.II. We find that the landmark localization can still achieve an accurate result even though the point cloud has been very sparse and non-uniform, which can show that we have good robustness in landmark localization by using global features.

D. Comparison Experiments

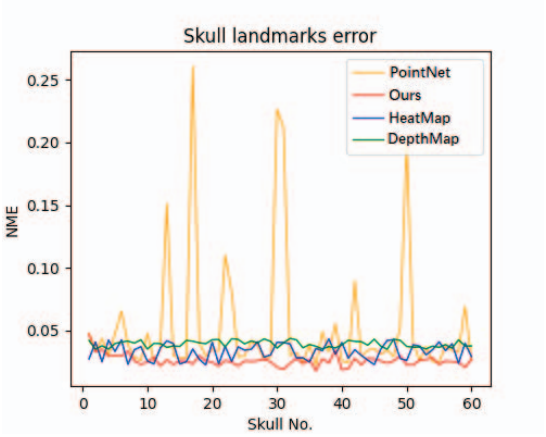
We compared our method with two recent 3D landmark localization approaches, DepthMap [7] and HeatMap in [15]. As shown in Fig. 5, our method can more accurately locate the seventy-eight (78) physiological landmarks. The comparison methods cannot detect accurate landmarks without salient geometric features (landmarks located in the middle of foreheads and cheeks). Benefited from the combination between local and global features, our method can solve the problem. The landmarks without salient geometric features can be detected



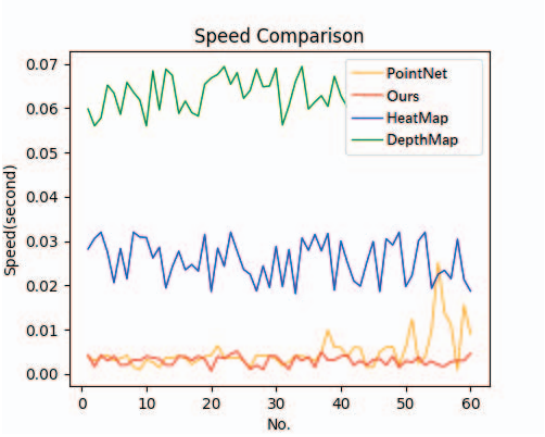
(a) Normal Ablation



(b) Skin Comparison



(c) Skull Comparison



(d) Speed Comparison

Fig. 4. The normalized mean error and mean speed

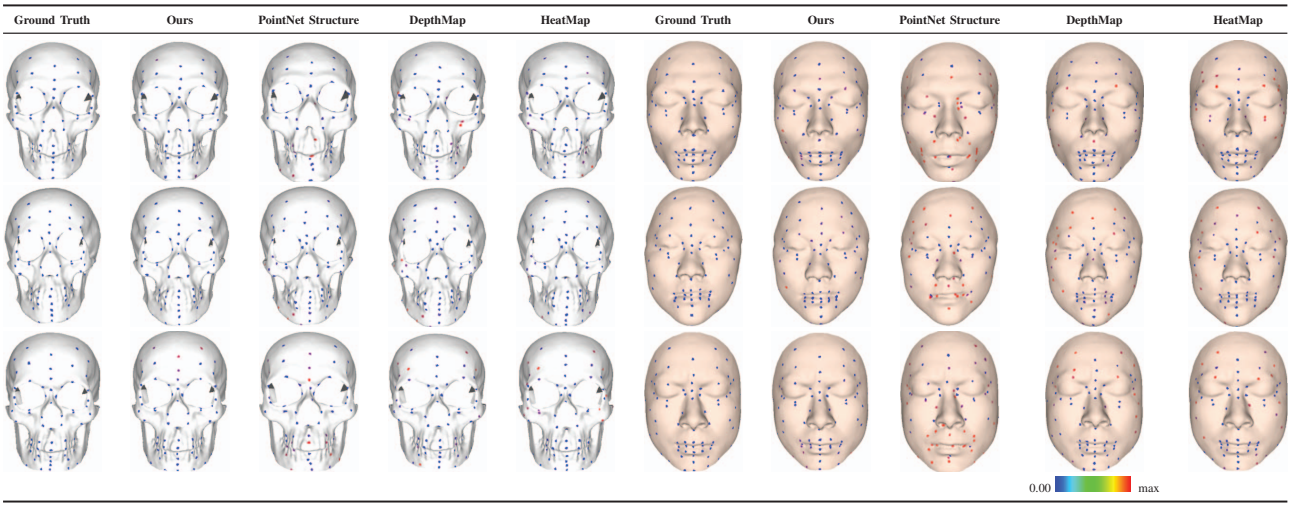
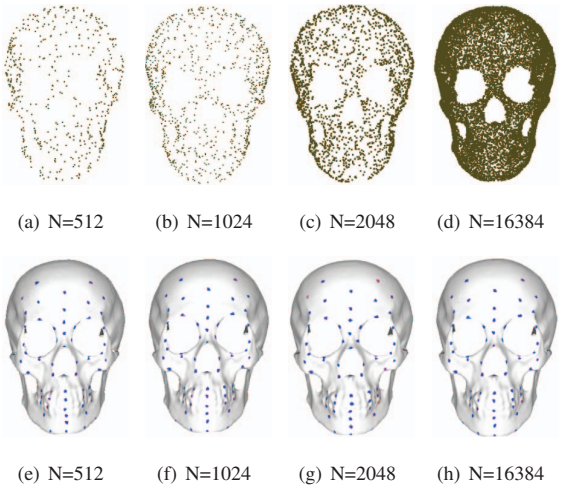


Fig. 5. Skin and skull localization result quality comparison. From left to right are: the ground truth, the localization results of our method, the localization results of PointNet, the localization results of DepthMap, and the localization results of HeatMap.Blue landmarks represent the smaller error.

TABLE III
LOCALIZATION ERROR AND ALGORITHM SPEED (IN SECONDS) COMPARISON WITH POINTNET [4], DEPTHMAP [7] AND HEATMAP [15]. LOWER NUMBER (ERROR AND SPEED) IS BETTER.

	Ours	Pointnet	DepthMap	HeatMap
NME	**0.02487**	0.06257	0.03883	0.03110
Speed	**0.00303**	0.00492	0.06318	0.02561



(a) N=512 (b) N=1024 (c) N=2048 (d) N=16384

(e) N=512 (f) N=1024 (g) N=2048 (h) N=16384

Fig. 6. Different Points Number Comparision

by global feature analysis. Some instances are shown in Fig. 4. The quantitative analysis results are reported in Tables II and III. It is clear that our method achieves more accurate landmarks at a faster speed.

V. CONCLUSION

In this paper, we propose a regression network that can robustly and automatically estimate craniofacial landmarks on point cloud models. Our method overcomes the main challenges of 3D craniofacial landmark localization. It improves accuracy with a controllable point number. Experimental results demonstrate the effectiveness of our method on various point cloud models with non-uniform densities or sparse distributions.

ACKNOWLEDGMENT

The authors gratefully appreciated the anonymous reviewers for all of their helpful comments. We also thank the support of Xianyang Hospital for providing craniofacial data.

REFERENCES

[1] Kostiantyn Khabarlak and Larysa Koriashkina, "Fast facial landmark detection and applications: A survey," arXiv preprint arXiv:2101.10808, 2021.

[2] Xu, Zixuan and Li, Banghuai and Yuan, Ye and Geng, Miao, "Anchorface: An anchor-based facial landmark detector across large poses," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 3092–3100, 2021.

[3] Xiang, Mingcan and Liu, Yinglu and Liao, Tingting and Zhu, Xiangyu and Yang, Can and Liu, Wu and Shi, Hailin, "The 3rd grand challenge of lightweight 106-point facial landmark localization on masked faces," 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–6, 2021.

[4] Qi, Charles R and Su, Hao and Mo, Kaichun and Guibas, Leonidas J, "Pointnet: Deep learning on point sets for 3d classification and segmentation," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 652–660, 2017.

[5] Gilani, Syed Zulqarnain and Mian, Ajmal and Eastwood, Peter, "Deep, dense and accurate 3D face correspondence for generating population specific deformable models," Pattern Recognition, vol. 69, pp. 238–250, 2017.

[6] Terada, Takuma and Chen, Yen-Wei and Kimura, Ryusuke, Peter, "3D facial landmark detection using deep convolutional neural networks," 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 390–393, 2018.

[7] Yali Xu and Junli Zhao, "An Automatic 3D Craniofacial Feature Points Location Based on DepthMap," International Symposium on Artificial Intelligence and Robotics, 2021.

[8] Zulqarnain Gilani, Syed and Shafait, Faisal and Mian, Ajmal, "Shape-based automatic detection of a large number of 3D facial landmarks," Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4639–4648, 2015.

[9] Sukno, Federico M and Waddington, John L and Whelan, Paul F, "3-D facial landmark localization with asymmetry patterns and shape regression from incomplete local features," IEEE transactions on cybernetics, vol. 45, pp. 1717–1730, 2014.

[10] Gupta, Shalini and Markey, Mia K and Bovik, Alan C, "Anthropometric 3D face recognition," International journal of computer vision, vol. 90, pp. 331–349, 2010.

[11] Cheng, Xianghao and Da, Feipeng, "3D Facial landmark localization based on two-step keypoint detection," 2018 International Conference on Audio, Language and Image Processing (ICALIP), pp. 406–412, 2018.

[12] Vezzetti, Enrico and Marcolin, Federica and Tornincasa, Stefano and Ulrich, Luca and Dagnes, Nicole, "3D geometry-based automatic landmark localization in presence of facial occlusions," Multimedia Tools and Applications, vol. 77, pp. 14177–14205, 2018.

[13] Wang, Yue and Sun, Yongbin and Liu, Ziwei and Sarma, Sanjay E and Bronstein, Michael M and Solomon, Justin M, "Dynamic graph cnn for learning on point clouds," Acm Transactions On Graphics (tog), vol. 38, pp. 1–12, 2019.

[14] Gao, Kangkang and Yang, Shanming and Fu, Keren and Cheng, Peng, "Deep 3d facial landmark detection on position maps," Intelligence Science and Big Data Engineering. Visual Data Engineering: 9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part I 9, pp. 299–311, 2019.

[15] O'Sullivan, Eimear, "Extending Convolutional Pose Machines for Facial Landmark Localization in 3D Point Clouds," Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.

[16] Zhang, Jingchen and Gao, Kangkang and Zhao, Qijun and Wang, Daning, "Pose invariant 3D facial landmark detection via pose normalization and deep regression," 2020 2nd International Conference on Image Processing and Machine Vision, pp. 74–78, 2020.

[17] Zhang, Jingchen and Gao, Kangkang and Fu, Keren and Cheng, Peng, "Deep 3D Facial Landmark Localization on position maps," Neurocomputing, vol. 406, pp. 89–98, 2020.

[18] Lorensen, William E and Cline, Harvey E, "Marching cubes: A high resolution 3D surface construction algorithm," ACM siggraph computer graphics, vol. 21, pp. 163–169, 1987.

[19] Hu, Yongli and Duan, Fuqing and Yin, Baocai and Zhou, Mingquan and Sun, Yanfeng and Wu, Zhongke and Geng, Guohua, "A hierarchical dense deformable model for 3D face reconstruction from skull," Multimedia tools and applications, vol. 64, pp. 345–364, 2013.

[20] Duan, Fuqing and Yang, Yanchao and Li, Yan and Tian, Yun and Lu, Ke and Wu, Zhongke and Zhou, Mingquan, "Skull identification via correlation measure between skull and face shape," IEEE transactions on information forensics and security, vol. 9, pp. 1322–1332, 2014.

[21] Kumar, Abhinav and Marks, Tim K and Mou, Wenxuan and Feng, Chen and Liu, Xiaoming, "UGLLI face alignment: Estimating uncertainty with gaussian log-likelihood loss," Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019.