1-1-2021

Deep learning for 3D ear detection: A complete pipeline from data generation to segmentation

Md Mursalin
Edith Cowan University

Syed Mohammed Shamsul Islam
Edith Cowan University

10.1109/ACCESS.2021.3129507
Mursalin, M., & Islam, S. M. S. (2021). Deep learning for 3D ear detection: A complete pipeline from data generation to segmentationn. IEEE Access, 9, 164976-164985.
https://doi.org/10.1109/ACCESS.2021.3129507

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 14, 2021, accepted November 16, 2021, date of publication November 19, 2021,
date of current version December 21, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3129507

Deep Learning for 3D Ear Detection: A Complete Pipeline From Data Generation to Segmentation

MD. MURSALIN AND SYED MOHAMMED SHAMSUL ISLAM, (Senior Member, IEEE)
Centre for AI & ML, Discipline of Computing and Security, School of Science, Edith Cowan University, Joondalup, WA 6027, Australia

Corresponding author: Md. Mursalin (m.mursalin@ecu.edu.au)

This work was supported in part by Edith Cowan University Higher Degree by Research Scholarship (ECUHDRS), and in part by the School of Science Research Incentive Grant.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT The human ear has distinguishing features that can be used for identification. Automated ear detection from 3D profile face images plays a vital role in ear-based human recognition. This work proposes a complete pipeline including synthetic data generation and ground-truth data labeling for ear detection in 3D point clouds. The ear detection problem is formulated as a semantic part segmentation problem that detects the ear directly in 3D point clouds of profile face data. We introduce EarNet, a modified version of the PointNet++ architecture, and apply rotation augmentation to handle different pose variations in the real data. We demonstrate that PointNet and PointNet++ cannot manage the rotation of a given object without such augmentation. The synthetic 3D profile face data is generated using statistical shape models. In addition, an automatic tool has been developed and is made publicly available to create ground-truth labels of any 3D public data set that includes co-registered 2D images. The experimental results on the real data demonstrate higher localization as compared to existing state-of-the-art approaches.

INDEX TERMS 3D point clouds, deep neural network, data generation, ear detection.

I. INTRODUCTION

The external shape of the human ear has distinguishing features that differ significantly from person to person. Research shows that even the ears of identical twins are different [1]–[3]. Importantly, the ear shape of a person remains steady between the ages of 8 to 70 [4]–[6]. These two factors have attracted the research community to investigate using images of the ear for numerous applications, including biometric identification; 3D ear reconstruction from partially occluded ear images [7] or from a single 2D ear image [8]; gender recognition; genetic study; and asymmetry analysis for clinical purposes [9]–[13].

In ear-based biometrics, one of the significant steps is to localize ears in profile face images. Most ear detection approaches have used 2D images for ear region localization as they require fewer computations [11], [12], [14]. Due to the importance of being able to handle unconstrained images for object detection and segmentation, recently numerous deep learning-based methods have been proposed, including simple convolutional neural network (CNN) based methods [15]–[18], landmark-based methods [19], Faster R-CNN based methods [20], pixel-wise methods [21], and geometric-based methods [22]. However, 2D image-based approaches are limited to constrained scenarios due to their sensitivity to lighting conditions and pose variations. Therefore, 3D images can be used to overcome the limitations of 2D images [23].

Recent developments in 3D imaging techniques have fast-tracked 3D image-based applications, including biometrics, robotics, medical diagnosis, and autonomous driving [23], [24]. Generally, 3D data can be represented in various forms, such as point clouds, volumetric grids, depth images, and meshes. Point cloud representation is becoming more popular as it reserves original geometric information in 3D domains without discretization. However, conventional convolutional neural networks cannot be applied directly to point clouds due to the irregular order of these points. Therefore, most work using 3D images has generally converted point cloud data to

The associate editor coordinating the review of this manuscript and approving it for publication was Kaustubh Raosaheb Patil.

Euclidean structured format before sending it to CNN architectures. This representation conversion introduces unnecessarily voluminous data and wraps natural invariances of the data due to the generation of quantization artefacts. Recently, Qi *et al.* have introduced two novel deep learning architectures named PointNet [25] and PointNet++ [26] that can identify features directly on 3D point clouds. These two networks provide solutions to the aforementioned problem and open the door to solve many other research questions in classification and semantic segmentation, including Engelmann [27], PointSIFT [28], 3DContextNet [29], ShellNet [30], LSANet [31], PointCNN [32], PCCN [33], ConvPoint [34], KPConv [35], InterpCNN [36]), RSNet [37], G+RCU [38], 3P-RNN [39], DGCNN [40], SPG [41], GACNet [42], and DPAM [43]. For more details readers are directed to the comprehensive survey on point cloud data presented by Guo *et al.* [44].

In this work, the ear detection problem is expressed as a semantic part segmentation problem where the profile face data (3D point clouds) is divided into two parts: ear and non-ear. As the problem is formulated in a single class with two parts, we are motivated to use a simpler network. In this work, we propose a deep learning-based approach named EarNet to detect ears directly on 3D point clouds of profile face data by modifying PointNet++ [26] architecture. To handle pose variations in the test data sets, we include a rotation augmentation block during the transfer learning of the EarNet.

Conventionally, a large set of training data is required to train a deep neural network efficiently. To the best of our knowledge, however, labeled 3D point cloud data for ear detection is not available. Therefore, we propose a novel approach for generating a large 3D synthetic profile face data set using two publicly available statistical 3D face models to train the proposed EarNet. Three public data sets are utilized to evaluate the performance of the trained model. Moreover, to examine the robustness of this approach, we also use a challenging 3D profile face data set from the University of Western Australia (UWA) that contains occlusions due to earphones. The contribution of this work can be summarized as follows:

1) A novel deep learning-based ear detection model named EarNet is proposed. EarNet is a modified version of PointNet++ [26] with a rotation augmentation block addressing pose variation problems in the real data.
2) A novel approach is proposed to synthetically generate a large number of 3D profile face data, which is used to train the proposed EarNet.
3) A novel approach is proposed to create the ground-truth labels on real 3D data where 2D co-registered images are available. The ground-truth data is then used for quantitative evaluation of the EarNet.
4) Comprehensive experiments are conducted demonstrating state-of-the-art performance on the largest publicly available 3D profile face data set.

The rest of the paper is organized as follows. Related work for 3D ear detection is described in Section II. The proposed ear detection pipeline is elaborated in Section III. The performance evaluation is explained in Section IV, followed by a conclusion in Section V.

II. RELATED WORK

The main focus of this work is ear detection in 3D data. Therefore, we only include existing approaches that have used 3D data for ear detection and categorize them into two groups: conventional machine learning-based approaches and deep learning-based approaches. These are summarized below.

A. MACHINE LEARNING-BASED APPROACHES

Existing machine learning-based approaches for ear detection in 3D data are either shape model-based, landmark-based, or graph-based. Chen and Bhanu [45] proposed a shape model-based approach for localizing ears in 3D profile face images. The helix and anti-helix parts of the ear were represented by a shape model consisting of a discrete set of 3D vertices. The authors extracted step edges from the profile images because of strong visibility in the ear helix. The segments of the edges were thinned, dilated, and classified into several clusters. A modified iterative closest point (ICP) algorithm was applied to align the edges and the shape model. Ear detection was obtained by the minimum registration error between the cluster and the shape model. The reported detection accuracy was 92.6% on 312 test images from 52 subjects. The limitation of the approach was the sensitivity of scale and pose variation. Zhou *et al.* [46] presented a 3D shape model to extract a set of shape-based features to train the support vector machine (SVM) classifier. They reported 100% accuracy; however, this result was obtained on only 142 test images.

A landmark-based ear detection technique that achieved 100% detection accuracy on the UND J2 data set was proposed by Lei *et al.* [23]. They presented a tree-based graph (ETG) to represent the ear and a curvedness map for localizing ear landmarks. However, their approach required manual intervention for landmark annotation.

An edge connectivity graph was proposed by Prakash and Gupta [47] for ear detection on 3D images from the UND J2 data set achieving 99.38% detection accuracy. The authors used a connectivity graph technique to extract the initial ear edge image. Their approach handled the influence of the scale and in-plane rotation. However, the authors did not solve off-plane rotation for ear detection. As a result, they had to discard some images from the UND J2 data set because of poor detection quality. Pflug *et al.* [48] proposed a binary mean curvature map for edge detection on 3D profile images and reported 95.65% accuracy on the UND J2 data set. The detected edges were used for semantic analysis to reconstruct the helix contour of the ear. The successful detection was defined by 50% overlapping pixels between ground-truth and the predicted ear region. As a result, their approach included additional pixels as an ear region, including clothes (e.g., scarves and collars).

TABLE 1. Summary of the existing 3D ear detection methods.

Author and publication	Detection method	Test images	Detection accuracy (%)
Chen et al. [45]	Ear helix line	312	92.6
Zhou et al. [46]	Histogram-based	142	100
Prakash et al. [47]	Connectivity graph	1604	99.38
Pflug et al. [48]	Edge-patterns	2414	95.65
Lei et al. [23]	Landmark-based	1800	100
Mursalin et al. [49]	Point cloud	1800	93.09
Zhu et al. [50]	PointNet++	1385	93

B. DEEP LEARNING-BASED APPROACHES

The earliest attempt for ear detection on 3D point clouds was EpNet [49] where network layers of PointNet [25] were customized to detect ear points on 3D profile faces. EpNet mapped the input points into a feature vector by multilayer perceptron networks (MLP). A max-pooling operator was then employed on these feature vectors. This pooling operation resulted in a permutation invariant global feature vector. Finally, using MLP, the point feature vector and the global feature vector were combined and transformed into an output vector. Although EpNet solved permutation and transformation invariance in point clouds, it cannot capture the local structure in the Euclidean space. As a result, detection accuracy was affected by pose variations.

Recently, Zhu and Mu [50] have proposed an ear segmentation approach using PointNet++. The authors trained their network using transfer learning on pre-trained weights from ShapeNet data [51]. They used one 3D data per subject (total of 415 subjects) from the UND J2 data set for fine-tuning their segmentation network. Their approach was tested on the remaining data from the UND J2 data set. However, the authors did not examine the use of data augmentation while training their deep neural network. They also did not examine the effect of pose variation effects on the detection performance. We summarize existing 3D ear detection methods in Table 1. Note that the authors usually reported various performance metrics as dependent on their own evaluation protocols. As a result, direct comparisons between the detection accuracy reported in Table 1 should be avoided.

III. PROPOSED EAR DETECTION PIPELINE

In this work, a large synthetic data set is produced using two publicly available statistical models to train the proposed EarNet. The ground-truths of real data sets are generated by utilizing Mask R-CNN [52]. The complete processing pipeline of ear detection is described below.

A. TRAINING DATA GENERATION

To create an extensive 3D data set for training, two publicly available statistical models, Basel Face Model (BFM) [53] and Liverpool-York Head Model (LYHM) [54] were used. The aim of using two models was to increase variations in the training data. Both models were created using a dimensionality reduction technique named Principal Component Analysis (PCA). By varying shape parameters, different face instances can be generated. Then it is straightforward to label the ear points of these generated data because a known one-to-one correspondence exists amongst the data.

All 3D profile face images of the UND J2 and UND F data set are left-side profile face images. Our literature review indicates that ear detection is conducted mostly on profile face 3D images. However, both of the above-mentioned statistical models contain full-face images. Therefore, we transformed the full-face image of the statistical models to left-side profile face data. The following steps were conducted to create the left-side profile face data. First, the nose tip was detected using a coarse to fine approach proposed by Mian *et al.* [55], where each 3D face data was sliced horizontally at multiple steps. The location on the slice with the largest altitude triangle was regarded as a possible nose tip and was given a confidence value equivalent to the altitude. This process was iterated for all the slices to get one candidate point per slice corresponding to the nose ridge. Some points that did not correspond to the nose ridge were considered outliers. The outliers were eliminated by using Random Sample Consensus (RANSAC) [56]. The point with the maximum confidence value was selected as a nose tip. Second, the detected nose tip was chosen for the current viewpoint. The full-face image was rotated to a different azimuth ($-45°$, $-60°$, $-90°$) and elevation ($\pm 30°$) angles. This rotation facilitated pose variations in the training data. Third, the hidden point for each rotation angle was deleted by using a hidden point removal algorithm [57]. Finally, the preprocessed data were downsampled. The purpose of this downsampling was to reduce the computation for the EarNet. Three downsampling techniques (random sampling without replacement, uniform box grid, and non-uniform box grid) were applied. The non-uniform box grid method was selected as this shows better sampling quality to retain the overall geometric shape of the 3D face data. The number of points was selected empirically to preserve the overall shape of the face and ear region. We tested 1024, 2048, and 4096 points and chose the 4096 points for better visual quality.

After nose tip detection and downsampling, we used a threshold-based technique proposed by Gautam Kumar [58] to label the ear points. The threshold value was selected empirically. We observed that the distance between ear and nose was around 26 mm. The ear points were present within 20 mm width and 35 mm height. The training data preparation is summarized in Figure 1.

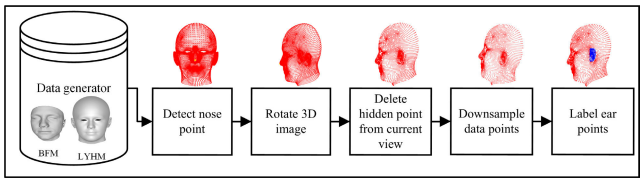

FIGURE 1. Block diagram of training data generation.

B. GROUND-TRUTH LABELING

Data labeling is a crucial task in deep learning. This work proposes a novel approach that involves labeling 3D public data sets where corresponding co-registered 2D images are available. The purpose of the data labeling process is to evaluate the quantitative performance of the proposed ear detection model. Our data labeling process was divided into two stages. Firstly, the ear region was detected on 2D profile face images using the Mask R-CNN [52]. The Mask R-CNN is an extended version of the Faster R-CNN [59], adding a segment to predict an object mask within the detection bounding box detected by Faster R-CNN. The purpose of using Mask R-CNN was to localize the pixels belonging to the ear region instead of just bounding boxes. The output of the Mask R-CNN is a 2D binary mask of a given 2D color profile face image. Secondly, the detected mask was projected to the co-registered 3D data for labeling. We labeled '1' for points that belong to the ear and '0' for points that belong to the non-ear. The block diagram of the ground-truth labeling on real data is illustrated in Figure 2.

The Mask R-CNN implemented by Waleed *et al.* [60] was used in this study. To train the Mask R-CNN, we randomly selected a few sample images from each data set mentioned in Section IV-A. The total number of images for training and testing was 200 and 40, respectively. The VGG Image Annotator (VIA) [61] was used for labeling the 2D color images. We trained the 2D ear detection Mask R-CNN starting from pre-trained COCO weights [60]. The results show an intersection over union score of 90.32% on 40 test images. Therefore, we visually checked all the predicted ear regions and corrected them manually if needed.

C. EAR DETECTION NETWORK (EarNet)

The proposed EarNet is a deep neural network customized to PointNet++ [26] layers for ear detection. The PointNet++ part segmentation network was designed for 16 different classes with 50 parts, whereas we trained our proposed EarNet for 1 class with 2 parts. Therefore, a smaller network with a lower number of parameters can learn the variations. For this reason, we empirically dropped some of the MLP layers in [26]. As a result, the execution time was significantly reduced without decreasing the accuracy. In addition, a data augmentation block was added to rotate the full 3D point cloud with respect to the x and y axes. The purpose of this augmentation was to provide more understanding of a given object. This addition of augmentation improved the performance of ear detection in 3D point

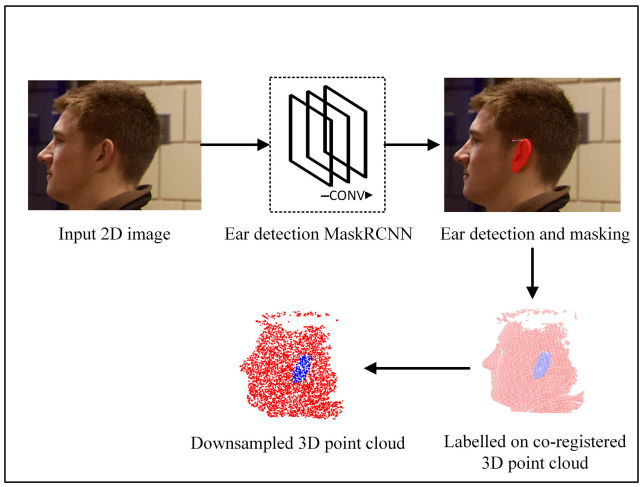

FIGURE 2. Block diagram of ground-truth labeling procedure on real data. Here, the input 2D color image passes through the Mask R-CNN. The output of the network is the detected 2D mask of the ear. This masked image is transferred to the co-registered 3D point cloud.

TABLE 2. The network architecture of the proposed EarNet. Here, SA, FP, DP and FC represent as set abstraction, feature propagation, dropout, and fully connected layers respectively.

Layers	Parameters
SA1	(512,[0.1,0.2,0.4],[32,64,128], [[32,32,64],[64,64,128],[64,96,128]])
SA2	(128,[0.4,0.8],[64,128], [[32,32,64],[128,128,256],[128,196,256]])
SA 3	(none, none, none, [256,512,1024])
FP1	[256,128]
FP2	[128,128]
FP3	[128,128]
DP1	0.5
FC	[128,2]

clouds. We also demonstrated that in the presence of pose variations on 3D profile face data, EarNet performed better than PointNet++. The architecture of our ear detection model is shown in Table 2.

The EarNet consists of several layers, including set abstraction, feature propagation, and segmentation layers. In the set abstraction layers, sampling and grouping were conducted using point convolutions and the furthest point sampling method. Skip link concatenation was used for feature propagation to the next layers of the network. In the segmentation block, a fully connected layer was utilized to estimate the per-point class scores for every point in the input data. We used the same notation as [26] for describing the architecture of EarNet, where set abstraction is $SA(K, r, [l_1, \ldots, l_d])$ K number of local regions in radius r and $[l_1, \ldots, l_d]$ is the fully connected layers with $l_i (i = 1, \ldots, d)$ output channels. $FP([l_1, \ldots, l_d])$ represents the feature propagation layer that has d fully connected layers. Further, $MLP([l_1, \ldots, l_d])$ is the multi-layer perceptron. Three consecutive *MLPs* of size (128, 128), (128, 128), and (128, n) were then used to propagate the output of the feature extraction block. Here, n is the number of parts, which is two in this case. In all layers, the ReLU activation was executed. In the last two layers, dropout was applied.

D. EVALUATION METRICS

The performance of the proposed ear detection model was evaluated using two commonly used metrics, namely *Accuracy*, and intersection over union (*IoU*). These metrics were calculated using five different variables: true positive (*TP*), false positive (*FP*), true negative (*TN*), false negative (*FN*), and *Total*. Here, *TP* represents ear points correctly classified as part of an ear, while *TN* represents non-ear points classified as non-ear points. *FP* represents non-ear points classified as ear points, and *FN* represents ear points classified as non-ear points. *Total* is the number of points that exist in a given point cloud data. *Accuracy* is estimated using the following equation,

$$Accuracy = \frac{TP + TN}{Total} \tag{1}$$

The intersection over union (*IoU*) is calculated as follows:

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

IV. EXPERIMENTS

A. DATA

Three publicly available 3D profile face data sets, namely UND F [62], UND G [63], and UND J2 [64] were used to evaluate the performance of the proposed ear detection model. These data sets were developed by the University of Notre Dame and have been used as benchmark data sets within the ear biometrics community. The 3D scans were captured at different times, and the number of subjects (UND J2 is the largest and UND G is the smallest) varies among these data sets. The UND G data set was comprised of images with significant pose variations compared to the other two. A brief description of these data sets is explained below.

The UND F data set consists of 942 3D profile face scans with co-registered 2D color images. The total number of subjects is 302 (176 males and 126 females). The distribution of scans for each subject is not uniform. There are 562 scans of male subjects, and 380 scans of female subjects.

The UND J2 data set comprises a total of 1800 3D scans from 415 different subjects (178 females and 237 males). Each subject has a different number of images with scale and pose variations. Some images include occlusion by hair and earrings. In this study, a set of randomly selected 415 scans was kept separated for transfer learning of the proposed model and other purposes (see Sections IV-B and IV-C4), and the remaining 1385 scans were used for model evaluation.

The UND G data set includes 738 3D profile face scans with yaws of 45, 60, 75, and 90 degrees. There are 437 left-side and 301 right-side profile face scans. In this work, we only used the left side profile face scans.

Apart from the UND data set, we also acquired 3D profile face data from the University of Western Australia (UWA) [65]. The authors collected data from 50 subjects using a Minolta Vivid 910 range scanner. All these images contain earphones.

B. TRAINING

Firstly, we trained the proposed EarNet from scratch using 20,000 synthetic data (80% training and 20% testing). The hyperparameters were selected empirically. The optimal batch size was 16. We observed that the model failed to run if the batch size was greater than the chosen size. The number of data points of each scan was selected as 4096. The optimizer was selected as Adam [66] with a momentum of 0.9. The initial learning rate was set to 10^{-3}. This work used the default values for optimizer and learning rate from [26].

Secondly, we applied transfer learning to the network using 150 real 3D scans. These scans were randomly selected and separated from the total 415 scans in the UND J2 data set. In addition, we separated another 50 scans randomly from the remaining scans to evaluate the transfer learning technique. During transfer learning, we applied rotation augmentation. The total number of data became 3000 (each image contains 20 rotations) after the augmentation. In this work, all experiments were performed in the Lamda Balde machine with GPU 8× 1080 Ti GeForce GTX 1080 Ti. The cod[1] was implemented in PyTorch version 1.3.1.

C. RESULTS AND DISCUSSIONS

1) DETECTION ACCURACY

The average accuracy of our proposed EarNet on different public data sets is reported in Table 3. We obtained consistent accuracy throughout the data sets. Sample ear detection results on each data set are illustrated in Figure 3. The ear points are shown in blue, and the non-ear points are shown in red.

TABLE 3. The mean accuracy of the proposed EarNet on different publicly available real (non-synthetic) profile face data sets.

Data set	Mean Accuracy (%)
UND J2	98.62 ±0.60
UND G	99.01 ±0.04
UND F	99.00 ±0.01
UWA	98.89 ±0.50

We also examined the cases where no ear points were present in the profile face point clouds. The results demonstrated the correctness of our network, which does not detect any ear points in this test data. A sample outcome is shown in Figure 4.

The robustness to occlusions due to earrings is illustrated in Figure 5. The results show that ear points were detected correctly even in the presence of earrings. We also demonstrated the robustness of our model in the presence of earphones. Our ear detection model achieved 98.89% accuracy on the UWA ear data set. A sample result is shown in Figure 6.

We also compared the performance of our approach on the UND J2 data set with recently published work by Zhu *et al.* [50]. They used 415 scans (one scan per all 415 subjects) for transfer learning of the basic

[1] https://github.com/doctormachine/EarDetection

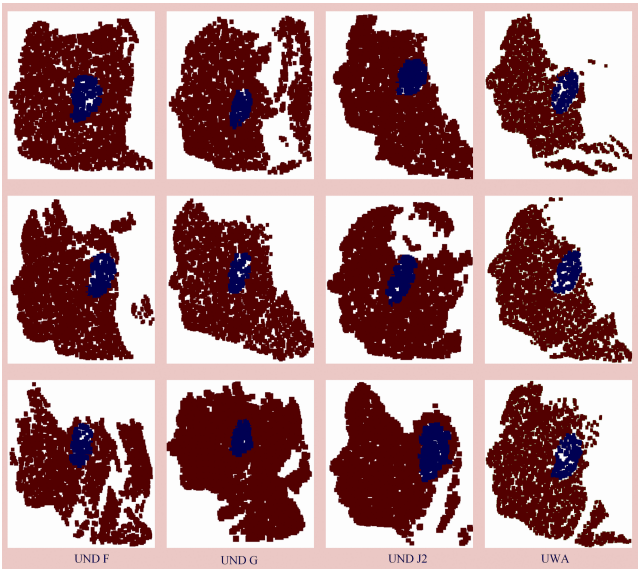

FIGURE 3. Sample ear detection results on various real (non-synthetic) data sets (best seen in color).

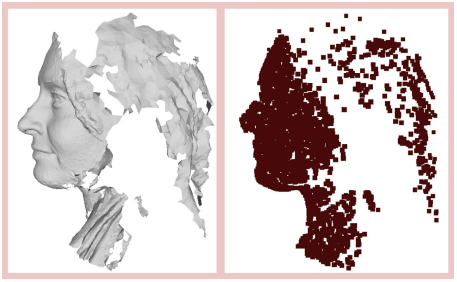

FIGURE 4. A test sample with a missing ear and corresponding detection result using our proposed approach demonstrating robustness to false positive.

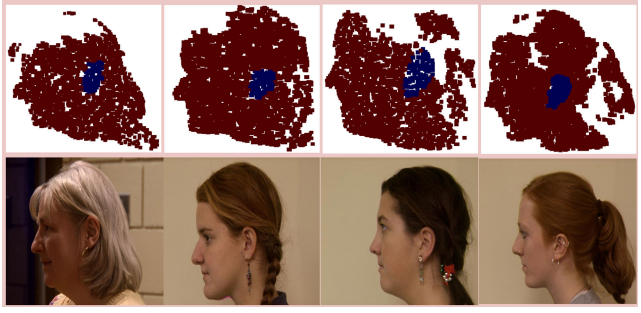

FIGURE 5. Sample prediction results in the presence of earrings. The 2D color images are presented to illustrate the occlusion (best seen in color).

PointNet++ network, and reported an accuracy of 93% on the remaining 1385 scans. On the other hand, our approach achieved an accuracy of 98.62% by using only 150 scans in the transfer learning. The better performance of our approach may be explained as follows. The learned weights of the basic PointNet++ were established by training 16 different objects, where each object contained 50 parts. However, the UND J2 data set is entirely different from the data used to

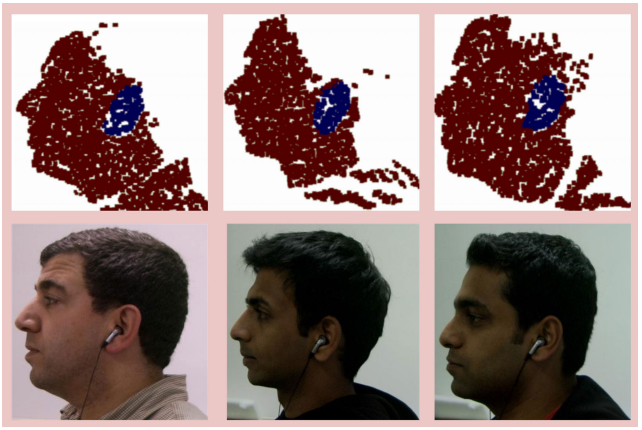

FIGURE 6. Sample prediction results in the presence of earphones. The 2D color images are presented to illustrate the occlusion (best seen in color).

train the PointNet++. Our proposed EarNet also outperforms EpNet [49] which was based on PointNet. The performance comparison is summarized in Table 4.

TABLE 4. The ear detection accuracy of our proposed approach compared to the existing ear detection methods on the UND J2 profile face data set (non-synthetic).

Author	Approach	Accuracy (%)
Mursalin et al. [49]	EpNet	93.09
Zhu et al. [50]	PointNet++	93
This paper	Proposed EarNet	98.67

To validate our detection accuracy, we compared our approach with the ear detection approach proposed by Islam *et al.* [65] using bounding boxes. Figure 7 shows that even with the lowest (<60%) IoU value, our detection is inside the bounding box, where the ear shape is significantly visible.

2) MEAN IoU

The mean IoU results on different data sets achieved by our model and those by PointNet and PointNet++ models are reported in Table 5. Our approach shows higher mIoUs for all data sets. We observed five failure cases on the UND G data set using both PointNet and PointNet++, as illustrated in Figure 8. These images contain significant pose variations. The first column is the ground-truth labels, whilst the remaining columns are the prediction of different models. It is worth noting here that all these images have significant pose variations. We see that PointNet captures the global shape but misses the local understanding of the shape. Although PointNet++ captures the local shape, it still lacks an understanding of the global structure. Therefore, our model captures both global and local shapes and does not have any complete failure cases.

To demonstrate robustness to the data point resolution of our proposed ear detection model, we conducted experiments on the test images with various resolutions. A sample result is illustrated in Figure 9, where the point cloud resolution is shown in descending order (top to bottom).

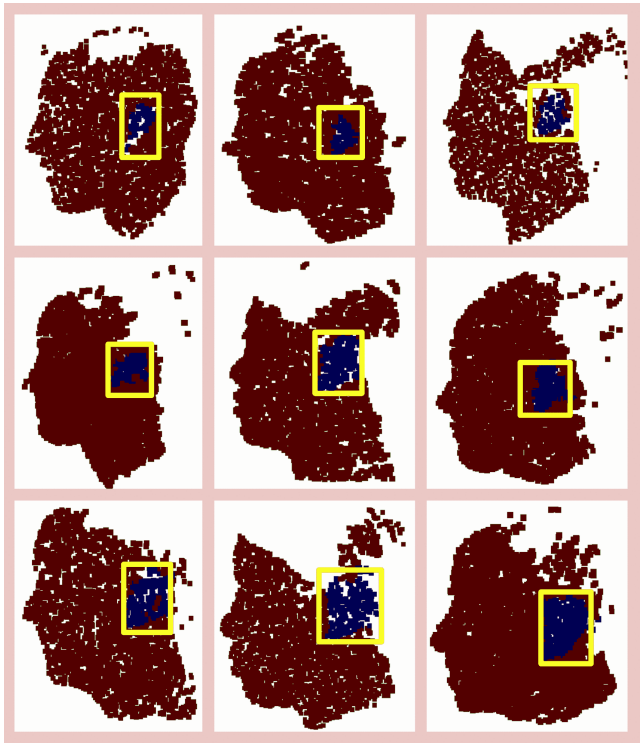

FIGURE 7. Sample test results while comparing the ear data points detected by our model with those in a bounding box proposed by Islam *et al.* [65] (best seen in color).

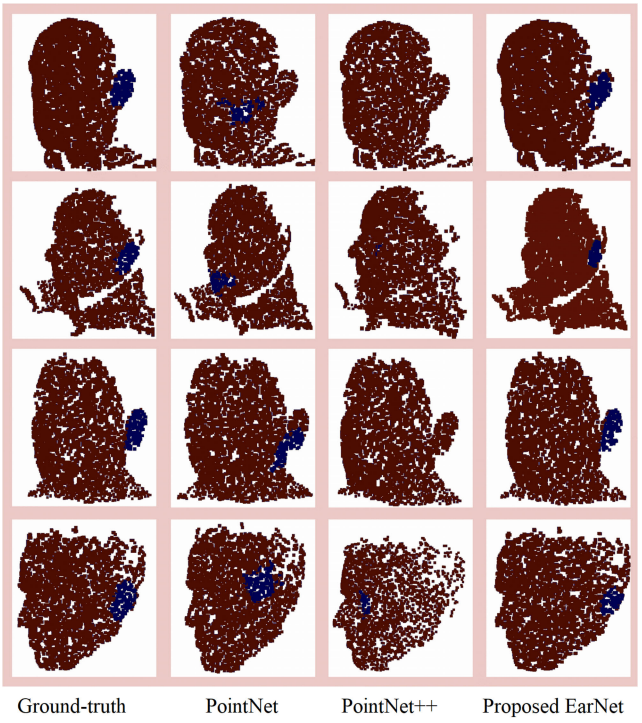

FIGURE 8. Sample test results of our EarNet compared to the original PointNet and PointNet++ networks. Notice that our model is able to detect the ear where other models fail (best seen in color).

TABLE 5. Quantitative comparison of ear detection on different data sets (non-synthetic).

Model	Data set	mIoU(%)				
		Mean	Std	Max	Min	Median
PointNet	UND J2	74.40	12.12	94.89	2.46	77.12
	UND G	70.17	18.87	93.05	**0**	76.04
	UND F	75.88	10.88	93.31	11.68	78.36
	UWA	71.99	17.31	90.11	8.29	78.24
PointNet++	UND J2	79.46	8.98	94.27	7.09	81.32
	UND G	70.90	19.25	93.70	**0**	77.00
	UND F	80.66	7.70	93.90	37.68	82.02
	UWA	72.41	21.71	92.45	5.98	81.59
Proposed EarNet	UND J2	**81.78**	**7.31**	**95.79**	**46.58**	**83.08**
	UND G	**82.16**	**7.69**	**96.51**	**21.65**	**83.57**
	UND F	**82.39**	**6.41**	**96.98**	**46.71**	**83.49**
	UWA	**84.04**	**5.82**	**95.38**	**63.29**	**85.08**

Although our model demonstrates considerable robustness against occlusion due to hair, we noticed that in the four cases where our model obtained mIoU less than 50%, a portion of the ear was covered by hair. These cases are illustrated in Figure 10. The corresponding 2D images (bottom row) show that these cases have hair over the ear (last two images) along with significant pose variations (the first two images).

3) DETECTION SPEED

The proposed EarNet achieved faster detection speed compared to PointNet++. The average inference time per 3D real scan (non-synthetic) was 0.11 s on a GPU GeForce GTX 1080 Ti machine. The detection speed between different models is reported in Table 6. Although PointNet shows a faster detection speed, it has less accuracy than the other two models.

TABLE 6. The mean detection speed comparison between different models. Here, we tested on 433 left side profile face from the UND G set (non-synthetic).

Approach	Detection Speed (second)
PointNet	0.09
PointNet++	0.15
Proposed EarNet	0.11

4) OTHER EXPERIMENTS

We performed experiments to evaluate the effects of training data size (synthetic) on network performance. First, we trained our network with 35,000 synthetic data. Then we retrained the network multiple times, dropping 5,000 data each time. Our experiments demonstrated that 20,000 data is optimal for training (Figure 11).

The quantitative results of our ear detection model trained on synthetic data only are presented in Table 7. The overall accuracy on three public data sets (UND J2, UND G, and UND F) is roughly 90%. The real data has more variability, which our trained model was not able to capture. As a result, we see a lower mIoU value. This result indicates that there is a possibility to improve the model's performance. Therefore, we also conducted experiments to see the effects of adding real data from the UND J2 data set using

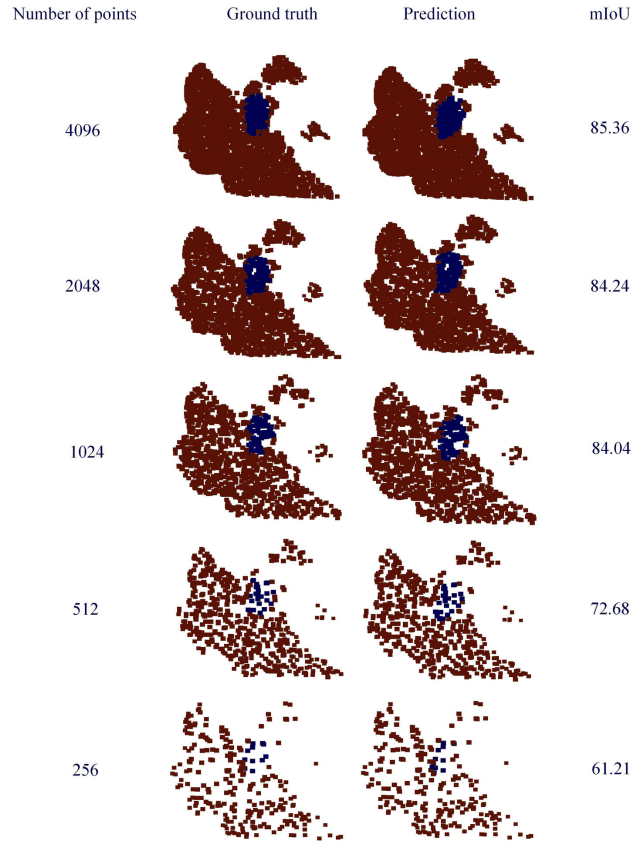

FIGURE 9. Sample ear detection results with respect to different point cloud resolutions (best seen in color).

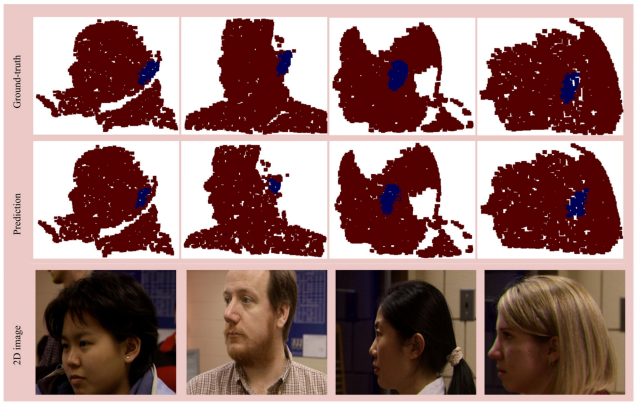

FIGURE 10. Sample prediction results with lower mIoU (mostly due to hair covering the ears). The top row is the ground-truth, the middle row is the prediction, and the bottom row is the corresponding 2D images (best seen in color).

transfer learning. A total of 415 scans from each subject were separated from the UND J2 data set. We performed three experiments selected from the 415 scans: first 50 subjects, randomly selected 50 subjects, and 50 subjects that seem hard to detect visually. The data for 50 subjects that were hard to detect were selected manually as they were visually challenging data in terms of occlusion. We performed the experiment three times for the randomly selected data and reported the average result. As illustrated in Figure 12, the

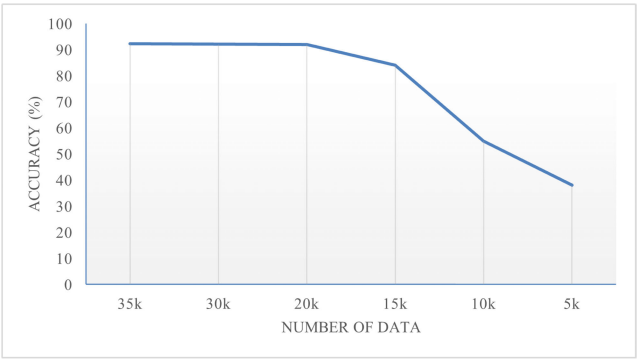

FIGURE 11. Effect of dropping the number of training data (synthetic data).

TABLE 7. The ear detection results on the three public data sets before transfer learning. The model was trained using synthetic data only and reported results on non-synthetic data.

Data set	Accuracy (%)	mIoU(%)				
		Mean	Std	Max	Min	Median
UND J2	89.18	59.73	11.42	79.35	10.22	60.52
UND G	90.9	43.86	13.08	71.70	0	58.90
UND F	89.58	60.47	11.48	81.25	17.12	65.49

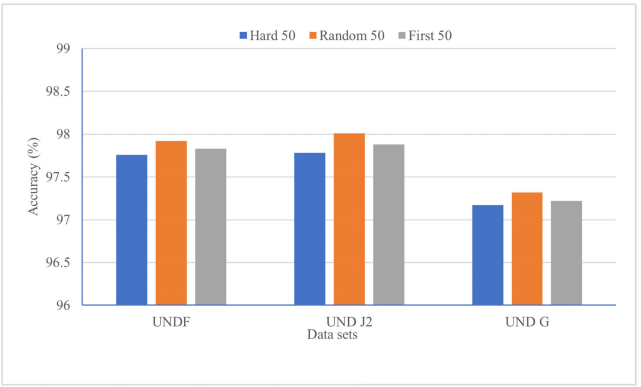

FIGURE 12. Effect of data selection for adding 50 real (non-synthetic) data while performing transfer learning (best seen in color).

performance of the network does not depend on how the set of 50 real data is selected for transfer learning. On the other hand, although a small number of real data contributes to a significant improvement in the accuracy, no significant changes in performance were observed by conducting transfer learning using more than 50 real data (see Figure 13). Therefore, we used 150 real data randomly selected from the pool of 415 scans (kept separated from the testing set) for transfer learning.

We also demonstrated the effects of training from scratch (including synthetic and real data) compared to training with synthetic data first and then transfer learning with real data. Our experiments did not show any significant differences in terms of accuracy. However, the transfer learning from the trained network with synthetic data required 2.25 hours less than training from scratch on the same machine.

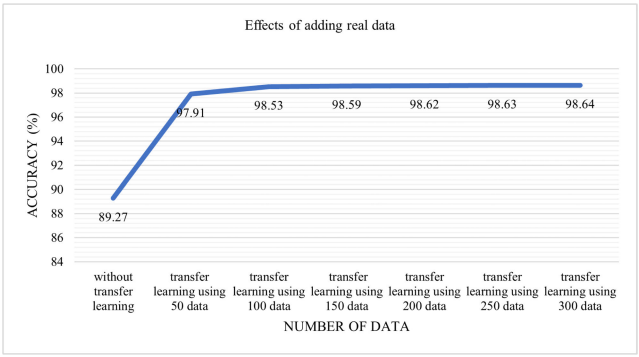

FIGURE 13. Effect of using different number of real (non-synthetic) data on the accuracy of transfer learning.

V. CONCLUSION

This work aims to detect ears directly on 3D point clouds of profile face data by applying a deep neural network named EarNet. A large set of synthetic profile face data was generated for training the proposed EarNet. Additionally, a novel approach is proposed to create ground-truth labels on real 3D data with corresponding co-registered 2D images. The experimental results demonstrate that our model performs significantly better than existing deep learning models for ear detection directly from 3D point clouds. A possible direction for future research is to incorporate the proposed ear detection model into an ear recognition pipeline. In addition, we aim to investigate different deep learning-based 2D segmentation networks for the ground-truth labeling pipeline.

ACKNOWLEDGMENT

The authors are thankful to Professor David Suter for his useful suggestions, support, and feedback.

REFERENCES

[1] H. Nejati, Z. Li, T. Sim, E. Martinez-Marroquin, and D. Guo, "Wonder ears: Identification of identical twins from ear images," in *Proc. Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 1201–1204.
[2] F. N. Sibai, A. Nuaimi, A. Maamari, and R. Kuwair, "Ear recognition with feed-forward artificial neural networks," *Neural Comput. Appl.*, vol. 23, no. 5, pp. 1265–1273, 2013.
[3] Z. Sun, A. A. Paulino, J. Feng, Z. Chai, T. Tan, and A. K. Jain, "A study of multibiometric traits of identical twins," *Proc. SPIE*, vol. 7667, Apr. 2010, Art. no. 76670T.
[4] H. Chen and B. Bhanu, "Contour matching for 3D ear recognition," in *Proc. 7th IEEE Workshops Appl. Comput. Vis. (WACV/MOTION)*, vol. 1, Jan. 2005, pp. 123–128.
[5] A. Iannarelli, *Forensic Identification Series: Ear Identification*, vol. 5. Hollywood, Los Angeles, CA, USA: Paramont, 1989.
[6] S. Tiwari, S. Jain, S. S. Chandel, S. Kumar, and S. Kumar, "Comparison of adult and newborn ear images for biometric recognition," in *Proc. 4th Int. Conf. Parallel, Distrib. Grid Comput. (PDGC)*, Dec. 2016, pp. 421–426.
[7] S. E. Kabbour and P.-Y. Richard, "Human ear surface reconstruction through morphable model deformation," in *Proc. Digit. Image Comput., Techn. Appl. (DICTA)*, Dec. 2018, pp. 1–5.
[8] C. Li, Z. Mu, F. Zhang, and S. Wang, "A novel 3D ear reconstruction method using a single image," in *Proc. 10th World Congr. Intell. Control Autom.*, Jul. 2012, pp. 4891–4896.
[9] M. Burge and W. Burger, "Ear biometrics," in *Biometrics*. Boston, MA, USA: Springer, 1996, pp. 273–285.
[10] V. Z. Emeršič, B. Meden, P. Peer, and V. Štruc, "Evaluation and analysis of ear recognition models: Performance, complexity and resource requirements," *Neural Comput. Appl.*, vol. 33, no. 10, pp. 15785–15800, 2018.
[11] Ž. Emeršić, V. Štruc, and P. Peer, "Ear recognition: More than a survey," *Neurocomputing*, vol. 255, pp. 26–39, Sep. 2017.
[12] S. Islam, M. Bennamoun, R. A. Owens, and R. Davies, "A review of recent advances in 3D ear and expression invariant face biometrics," *ACM Comput. Surv.*, vol. 44, no. 3, p. 14, 2012.
[13] P. Srivastava, D. Agrawal, and A. Bansal, "Ear detection and recognition techniques: A comparative review," in *Proc. Adv. Data Inf. Sci.* Singapore: Springer, 2020, pp. 533–543.
[14] Q. Zhu and Z. Mu, "Local and holistic feature fusion for occlusion-robust 3D ear recognition," *Symmetry*, vol. 10, no. 11, p. 565, Nov. 2018.
[15] P. L. Galdámez, W. Raveane, and A. G. Arrieta, "A brief review of the ear recognition process using deep neural networks," *J. Appl. Log.*, vol. 24, pp. 62–70, Nov. 2017.
[16] M. Moniruzzaman and S. Islam, "Automatic ear detection using deep learning," in *Proc. Int. Conf. Mach. Learn. Data Eng.*, 2017, pp. 108–114.
[17] S. Wang, Y. Du, and Z. Huang, "Ear detection using fully convolutional networks," in *Proc. 2nd Int. Conf. Robot., Control Autom. (ICRCA)*, 2017, pp. 50–55.
[18] I. I. Ganapathi, S. Prakash, I. R. Dave, and S. Bakshi, "Unconstrained ear detection using ensemble-based convolutional neural network model," *Concurrency Comput., Pract. Exper.*, vol. 32, no. 1, Jan. 2020, Art. no. e5197.
[19] C. Cintas, M. Quinto-Sánchez, V. Acuña, C. Paschetta, S. de Azevedo, C. C. S. de Cerqueira, V. Ramallo, C. Gallo, G. Poletti, M. C. Bortolini, S. Canizales-Quinteros, F. Rothhammer, G. Bedoya, A. Ruiz-Linares, R. Gonzalez-José, and C. Delrieux, "Automatic ear detection and feature extraction using geometric morphometrics and convolutional neural networks," *IET Biometrics*, vol. 6, no. 3, pp. 211–223, May 2017.
[20] Y. Zhang and Z. Mu, "Ear detection under uncontrolled conditions with multiple scale faster region-based convolutional neural networks," *Symmetry*, vol. 9, no. 4, p. 53, Apr. 2017.
[21] Ž. Emeršič, L. Lan Gabriel, V. Štruc, and P. Peer, "Pixel-wise ear detection with convolutional encoder-decoder networks," 2017, *arXiv:1702.00307*.
[22] A. Tomczyk and P. S. Szczepaniak, "Ear detection using convolutional neural network on graphs with filter rotation," *Sensors*, vol. 19, no. 24, p. 5510, Dec. 2019.
[23] J. Lei, X. You, and M. Abdel-Mottaleb, "Automatic ear landmark localization, segmentation, and pose classification in range images," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 46, no. 2, pp. 165–176, Feb. 2016.
[24] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1907–1915.
[25] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.
[26] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108.
[27] F. Engelmann, T. Kontogianni, J. Schult, and B. Leibe, "Know what your neighbors do: 3D semantic segmentation of point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 395–409.
[28] M. Jiang, Y. Wu, T. Zhao, Z. Zhao, and C. Lu, "PointSIFT: A SIFT-like network module for 3D point cloud semantic segmentation," 2018, *arXiv:1807.00652*.
[29] W. Zeng and T. Gevers, "3DContextNet: K-D tree guided hierarchical learning of point clouds using local and global contextual cues," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 314–330.
[30] Z. Zhang, B.-S. Hua, and S.-K. Yeung, "ShellNet: Efficient point cloud convolutional neural networks using concentric shells statistics," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1607–1616.
[31] L.-Z. Chen, X.-Y. Li, D.-P. Fan, K. Wang, S.-P. Lu, and M.-M. Cheng, "LSANet: Feature learning on point sets by local spatial aware layer," 2019, *arXiv:1905.05442*.
[32] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, "PointCNN: Convolution on X-transformed points," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 820–830.
[33] S. Wang, S. Suo, W.-C. Ma, A. Pokrovsky, and R. Urtasun, "Deep parametric continuous convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2589–2597.
[34] A. Boulch, "ConvPoint: Continuous convolutions for point cloud processing," *Comput. Graph.*, vol. 88, pp. 24–34, May 2020.

[35] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, "KPConv: Flexible and deformable convolution for point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6411–6420.
[36] J. Mao, X. Wang, and H. Li, "Interpolated convolutional networks for 3D point cloud understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1578–1587.
[37] Q. Huang, W. Wang, and U. Neumann, "Recurrent slice networks for 3D segmentation of point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2626–2635.
[38] F. Engelmann, T. Kontogianni, A. Hermans, and B. Leibe, "Exploring spatial context for 3D semantic segmentation of point clouds," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 716–724.
[39] X. Ye, J. Li, H. Huang, L. Du, and X. Zhang, "3D recurrent neural networks with context fusion for point cloud semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 403–417.
[40] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, 2019.
[41] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4558–4567.
[42] L. Wang, Y. Huang, Y. Hou, S. Zhang, and J. Shan, "Graph attention convolution for point cloud semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10296–10305.
[43] J. Liu, B. Ni, C. Li, J. Yang, and Q. Tian, "Dynamic points agglomeration for hierarchical point sets learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7546–7555.
[44] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3D point clouds: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4338–4364, Dec. 2021.
[45] H. Chen and B. Bhanu, "Shape model-based 3D ear detection from side face range images," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Sep. 2005, p. 122.
[46] J. Zhou, S. Cadavid, and M. Abdel-Mottaleb, "Histograms of categorized shapes for 3D ear detection," in *Proc. 4th IEEE Int. Conf. Biometrics, Theory, Appl. Syst. (BTAS)*, Sep. 2010, pp. 1–6.
[47] S. Prakash and P. Gupta, "An efficient technique for ear detection in 3D: Invariant to rotation and scale," in *Proc. 5th IAPR Int. Conf. Biometrics (ICB)*, Mar. 2012, pp. 97–102.
[48] A. Pflug, A. Winterstein, and C. Busch, "Ear detection in 3D profile images based on surface curvature," in *Proc. 8th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, Jul. 2012, pp. 1–6.
[49] M. Mursalin and S. M. S. Islam, "EpNet: A deep neural network for ear detection in 3D point clouds," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.* Cham, Switzerland: Springer, 2020, pp. 15–26.
[50] Q. Zhu and Z. Mu, "PointNet++ and three layers of features fusion for occlusion three-dimensional ear recognition based on one sample per person," *Symmetry*, vol. 12, no. 1, p. 78, Jan. 2020.
[51] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.
[52] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," 2017, *arXiv:1703.06870*.
[53] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *Proc. 6th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, Sep. 2009, pp. 296–301.
[54] H. Dai, N. Pears, W. Smith, and C. Duncan, "Statistical modeling of craniofacial shape and texture," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 547–571, Nov. 2019.
[55] A. S. Mian, M. Bennamoun, and R. Owens, "An efficient multimodal 2D-3D hybrid approach to automatic face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 11, pp. 1927–1943, Nov. 2007.
[56] M. S. Bartlett, "Independent component representations for face recognition," in *Face Image Analysis by Unsupervised Learning*. Boston, MA, USA: Springer, 2001, pp. 39–67.
[57] S. Katz, A. Tal, and R. Basri, "Direct visibility of point sets," *ACM Trans. Graph.*, vol. 26, no. 3, p. 24, 2007.
[58] G. Kumar. (2021). *3D Ear Recognition*. [Online]. Available: https://github.com/gautamkumarjaiswal/3DEarRecognition
[59] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[60] W. Abdulla, "Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow," *GitHub Repository*, 2017. [Online]. Available: https://github.com/matterport/Mask_RCNN
[61] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2276–2279.
[62] P. Yan and K. Bowyer, "Empirical evaluation of advanced ear biometrics," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Sep. 2005, p. 41.
[63] P. Yan and K. W. Bowyer, "An automatic 3D ear recognition system," in *Proc. 3rd Int. Symp. 3D Data Process., Visualizat., Transmiss. (3DPVT)*, Jun. 2006, pp. 326–333.
[64] P. Yan and K. W. Bowyer, "Biometric recognition using 3D ear shape," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 8, pp. 1297–1308, Aug. 2007.
[65] S. M. S. Islam, R. Davies, M. Bennamoun, and A. S. Mian, "Efficient detection and recognition of 3D ears," *Int. J. Comput. Vis.*, vol. 95, no. 1, pp. 52–73, Oct. 2011.
[66] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

MD. MURSALIN received the undergraduate degree in computer science and information technology from the Islamic University of Technology (IUT), Bangladesh, and the master's degree in computer science and engineering from the University of Jinan, China, through the prestigious Chinese Government Scholarship. He is currently a Ph.D. Fellow with Edith Cowan University (ECU), Australia, through the ECU Higher Degree by Research Scholarship (HDRS). He worked as an Assistant Professor in computer science with the Pabna University of Science and Technology, Bangladesh, and published several research articles in reputed journals and conferences. His research interests include biomedical signal processing, brain–computer interface, human–computer interaction, and machine learning. He has professional affiliations with internationally recognized societies, including IEEE and ACS. Furthermore, he has served as a Reviewer in various reputed journals and conferences, including IEEE Access.

SYED MOHAMMED SHAMSUL ISLAM (Senior Member, IEEE) received the B.Sc. degree in EEE and the M.Sc. degree in computer engineering (CE), in 2000 and 2005, respectively, and the Ph.D. degree (Hons.) in CE from The University of Western Australia (UWA), in 2011. He worked in different teaching and research positions at UWA and Curtin University in the past, and is currently working as a Senior Lecturer in computing with Edith Cowan University, Australia. He has supervised the completion of five higher degrees by research students and published over 60 scholarly articles. He has attracted tens of public media releases, awards, and grants (of over a million Australian dollars). His research interests include (but not limited to) artificial intelligence, medical imaging, biometrics, machine learning, networking, and computer vision. He is an Organizing/Technical Committee Member of over 35 conferences and a Senior Member of the Australian Computer Society. He is a regular reviewer of over ten Q1 journals and an Associate Editor of IEEE Access.