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

MD MURSALIN1 and SYED MOHAMMED SHAMSUL ISLAM1
1Centre for AI & ML, Discipline of Computing and Security, School of Science, Edith Cowan University, Joondalup, WA 6027, Australia
Corresponding author: Md Mursalin (e-mail: m.mursalin@ecu.edu.au).

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 made publicly available to create the 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 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 use 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 [15]–[18], landmark-based [19], Faster R-CNN based [20], pixel-wise [21], and geometric-based [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 open the door for 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. The point cloud representation is becoming more popular as it reserves the 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 the points. Therefore, most work using 3D images generally converts point cloud data to Euclidean structured format before sending it to the 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. introduced two novel deep learning architectures named as PointNet [25] and PointNet++ [26] that can identify features directly on 3D point clouds. These two networks provide the solution to the problem mentioned before and open the door to solve many other re-
search questions in classification and semantic segmentation, including Engelmann [27], PointSIFT [28], 3DContextNet [29], ShellNet [30], LSA-Net [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], DPAM [43]. For more details the reader should see the comprehensive survey on point cloud data presented by Guo et al. [44].

Here, 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 2 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 the PointNet++ [26] architecture. To handle the pose variation 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 the 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 used for quantitative evaluation of the EarNet.
4) Comprehensive experiments are conducted demonstrating the 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 the existing approaches that 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 et al. [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 the 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 142 test images only.

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 et al. [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 the 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).

B. DEEP LEARNING-BASED APPROACHES

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

Recently, Zhu et al. [50] proposed an ear segmentation approach using PointNet++. The authors trained their network using transfer learning on the 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. The existing 3D ear detection methods are summarized in Table 1. Notice that the authors usually reported various performance metrics and depended 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 by using two publicly available statistical models to train the proposed EarNet. The ground-truth of real data sets is generated by utilizing the 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] are used. The aim of using two models is to increase the variations in the training data. Both of the models were created using a dimensionality reduction technique named Principal Component Analysis (PCA). By varying the shape parameters, different face instances are generated. It is straightforward to label the ear points of these generated data because a known one-to-one correspondence exists among 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 a full-face image. Therefore, we transform the full-face image of the statistical models to left-side profile face data. The following steps are conducted to create the left-side profile face data. First, the nose tip is detected using a coarse to fine approach proposed by Mian et al. [55], where each 3D face data is sliced horizontally at multiple steps. The location on the slice with the largest altitude triangle is regarded as a possible nose tip and is given a confidence value equivalent to the altitude. This process is iterated for all the slices to get one candidate point per slice corresponding to the nose ridge. Some points that do not correspond to the nose ridge are considered outliers. The outliers are eliminated by using Random Sample Consensus (RANSAC) [56]. The point with the maximum confidence value is selected as a nose tip. Second, the detected nose tip is chosen for the current viewpoint. The full-face image is rotated to a different azimuth (−45°, −60°, −90°) and elevation (±30°) angles. This rotation facilitates pose variations in the training data. Third, the hidden point for each rotation angle is deleted by using a hidden point removal algorithm [57]. Finally, the preprocessed data is downsampled. The purpose of this downsampling is to reduce the computation for the EarNet. Three downsampling techniques (random sampling without replacement, uniform box grid, and non-uniform box grid) are applied. The non-uniform box grid method is selected as this shows better sampling quality to retain the overall geometric shape of the 3D face data. The number of points is 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 downampling, we use a threshold-based technique proposed by Gautam Kumar [58] to label the ear points. The threshold value is selected empirically. We observe that the distance between ear and nose is around 26 mm. The ear points are present within 20 mm width and 35 mm height. The training data preparation is summarized in Figure 1.

#### 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                  |
| Pfug 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. GROUND-TRUTH LABELING

Data labeling is a crucial task in deep learning. This work proposes a novel approach that involves labeling the 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 is divided into two stages. First, the ear region is detected on the 2D profile face images using the Mask R-CNN [52]. The Mask R-CNN is an extended version of the Faster R-CNN [59], which add a segment to predict an object mask within the detection bounding box detected by Faster R-CNN. The purpose of using Mask R-CNN is 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. Second, the detected mask is projected to the co-registered 3D data for labeling. We label ‘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 implementation by Waleed et al. [60] is used in this study. To train the Mask R-CNN, we randomly take a few sample images from each data set mentioned in Section IV-A. The total number of images for training and testing is 200 and 40, respectively. The VGG Image Annotator (VIA) [61] is used for labeling the 2D color images. We train 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 manually if needed.

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 and masked 2D color image. This masked image is transferred to the co-registered 3D point cloud.

C. EAR DETECTION NETWORK (EARNET)

The proposed EarNet is a deep neural network that is customized to the PointNet++ [26] layers for ear detection. The PointNet++ part segmentation network is designed for 16 different classes with 50 parts, while we train 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 drop some of the MLP layers in [26]. As a result, the execution time is significantly reduced without decreasing the accuracy. In addition, a data augmentation block is added to rotate the full 3D point cloud with respect to the x and y axes. The purpose of this augmentation is to provide more understanding of a given object. This addition of augmentation improves the performance of ear detection in 3D point clouds. We also demonstrate that in the presence of pose variations on the 3D profile face data, the EarNet performs better than the PointNet++. The architecture of our ear detection model is shown in Table 2.

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      | \((512, [0, 0, 0, 0, 4], [32, 64, 128], [32, 32, 64], [64, 64, 128], [64, 64, 128])\) |
| SA2      | \((128, [0, 0, 0, 0, 4], [64, 128], [128, 32, 64], [128, 128, 256], [128, 128, 256], [128, 128, 256])\) |
| SA 3     | \((\text{none, none, none}, [256, 512, 1024])\) |
| FP1      | [256, 128]                                                                |
| FP2      | [128, 128]                                                                |
| FP3      | [128, 128]                                                                |
| DP1      | 0.5                                                                        |
| FC       | [128, 2]                                                                  |

The EarNet consists of several layers, including set abstraction, feature propagation, and segmentation layers. In the set abstraction layers, sampling and grouping are conducted using point convolutions and the furthest point sampling method. The skip link concatenation is used for feature propagation to the next layers of the network. In the segmentation block, a fully connected layer is utilized to estimate the per-point class scores for every point in the input data. We use the same notation as [26] for describing the architecture of EarNet, set abstraction is \(SA(K, r, [l_1, \ldots, l_d])\) where \(K\) 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. \(MLP([l_1, \ldots, l_d])\) is the multi-layer perceptron. Three consecutive multilayer perceptrons (MLPs) of size \((128, 128), (128, 128),\) and \((128, n)\) are 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 is executed. In the last two layers, dropout is applied.
D. EVALUATION METRICS

The performance of the proposed ear detection model is evaluated using two commonly used metrics, namely Accuracy, and intersection over union (IoU). These metrics are 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. The Accuracy is estimated using the following equation,

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}}$$  \hspace{1cm} (1)

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

$$\text{IoU} = \frac{TP}{TP + FP + FN}$$  \hspace{1cm} (2)

IV. EXPERIMENTS

A. DATA

Three publicly available 3D profile face data sets, namely UND F [62], UND G [63], and UND J2 [64] are 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 by 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 is 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 is kept separated for transfer learning of the proposed model and other purposes (see Sections IV-B and IV-C4), and the remaining 1385 scans are 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 use 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

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

Second, we apply transfer learning to the network using 150 real 3D scans. These scans are randomly selected and separated from the total 415 scans in the UND J2 data set. In addition, we separate another 50 scans randomly from the remaining scans to evaluate the transfer learning technique. During transfer learning, we apply rotation augmentation. The total number of data becomes 3000 (each image contains 20 rotations) after the augmentation. In this work, all experiments are performed in the Lambda Balde machine with GPU 8x 1080 Ti GeForce GTX 1080 Ti. The code is 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 obtain a 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 examine the case where there are no ear points present in the profile face point clouds. The results demonstrate the correctness of our network, which does not detect any ear points on those test data. A sample outcome is shown in Figure 4.

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

We also compare 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)
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.

for transfer learning of the basic PointNet++ network, and reported an accuracy of 93% on the remaining 1385 scans. On the other hand, our approach achieves 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 contains 50 parts. However, the UND J2 data set is entirely different from the data used to train the PointNet++. Our proposed EarNet also outperforms EpNet [49] which is 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        |

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

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

To validate our detection accuracy, we compare 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, and the ear shape is also 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 the data sets. We observe 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, and the remaining columns are the prediction of different models. Notice 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. Our model capture both global and local shapes and therefore, does not have any complete failure cases.

To demonstrate the robustness to data point resolution of
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 |

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).

TABLE 6: The mean detection speed comparison between different models.

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

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).

3) Detection Speed

The proposed EarNet achieves faster detection speed compared to PointNet++. The average inference time per 3D real scan (non-synthetic) is 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.

4) Other Experiments

We perform experiments to evaluate the effects of training data size (synthetic) on the network performance. First, we train our network with 35,000 synthetic data. Then we retrain the network multiple times, dropping 5,000 data each time.
Our experiments demonstrate 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 around 90%. The real data has more variability, which our trained model is 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 performance. Therefore, we also conduct experiments to see the effects of adding real data from the UND J2 data set using transfer learning. A total of 415 scans from each subject are separated from the UND J2 data set. We perform three experiments selected from the 415 scans: first 50 subjects, randomly selected 50 subjects, and 50 subjects that seem hard to detect visually. The hard 50 data are selected manually and are visually challenging data in terms of occlusion. We perform the experiment three times for the randomly selected data and report the average result. As illustrated in Figure 12, the 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 are observed by conducting transfer learning using more than 50 real data (see Figure 13). Therefore, for transfer learning we use 150 real data randomly selected from the pool of 415 scans (kept separated from the testing set).

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

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

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).

FIGURE 9: Sample ear detection results with respect to different point cloud resolutions. (Best seen in color).
we will also investigate different deep learning-based 2D detection model into an ear recognition pipeline. In addition, a possible direction for future research is to incorporate the proposed ear segmentation networks for the ground-truth labeling pipeline. A large set of synthetic profile face data is 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 the 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.

### 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 is 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 the 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 will also investigate different deep learning-based 2D segmentation networks for the ground-truth labeling pipeline.

### REFERENCES

[1] H. Nejati, L. Zhang, T. Sim, E. Martinez-Marroquin, and G. Dong, “Wonder ears: Identification of Identical Twins from Ear Images,” in Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Nov 2012, pp. 1201–1204.

[2] F. N. Sibai, A. Nuaimi, A. Maamari, and R. Kowair, “Ear Recognition with Feed-forward Artificial Neural Networks,” Neural Computing and Applications, 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,” in Biometric technology for human identification VII, vol. 7667. International Society for Optics and Photonics, 2010, p. 76670T.

[4] H. Chen and B. Bhanu, “Contour Matching for 3D Ear Recognition,” in 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION’05)-Volume 1, vol. 1. IEEE, 2005, pp. 123–128.

[5] A. Iannarelli, “Forensic identification series: Ear identification,” Paramont, California, vol. 5, 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 Proceedings of 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), Dec 2016, pp. 421–426.

[7] S. Eddine and P.-Y. Richard, “Human Ear Surface Reconstruction Through Morphable Model Deformation,” in International Conference on Digital Image Computing: Techniques and Applications, 2018.

[8] C. Li, Z. Fu, F. Zhang, and S. Wang, “A Novel 3D Ear Reconstruction Method using a Single Image,” in Proceedings of the 10th World Congress on Intelligent Control and Automation. IEEE, 2012, pp. 4891–4896.

[9] M. Burge and W. Burger, “Ear biometrics,” in Biometrics. Springer, 1996, pp. 273–285.

[10] Ž. Emeršič, B. Meden, P. Peer, and V. Štruc, “Evaluation and Analysis of Ear Recognition Models: Performance, Complexity and Resource Requirements,” Neural computing and applications, pp. 1–16, 2018.

[11] Ž. Emeršič, V. Struc, and P. Peer, “Ear Recognition: More Than a Survey,” Neurocomputing, vol. 255, pp. 26–39, 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 Computing Surveys (CSUR), vol. 44, no. 3, p. 14, 2012.

[13] P. Srivastava, D. Agrawal, and A. Bansal, “Ear Detection and Recognition Techniques: A Comparative Review,” in Advances in Data and Information Sciences. 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, 2018.

[15] P. L. Galdámez, W. Raveane, and A. G. Arrieta, “A Brief Review of the Ear Recognition Process using Deep Neural Networks,” Journal of Applied Logic, vol. 24, pp. 62–70, 2017.

[16] M. Moniruzzaman and S. Islam, “Automatic Ear Detection using Deep Learning,” in In the proceedings of the International Conference on Machine Learning and Data Engineering, 2017, pp. 108–114.

[17] S. Wang, Y. Du, and Z. Huang, “Ear Detection using Fully Convolutional Networks,” in Proceedings of the 2nd International Conference on Robotics, Control and Automation. ACM, 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 and Computation: Practice and Experience, vol. 32, no. 1, p. e5197, 2020.

[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 et al., “Automatic Ear Detection and Feature Extraction using Geometric Morphometrics and Convolutional Neural Networks,” IET Biometrics, vol. 6, no. 3, pp. 211–223, 2016.

[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, 2017.

[21] Ž. Emeršič, L. L. Gabriel, V. Struc, and P. Peer, “Pixel-wise Ear Detection with Convolutional Encoder-Decoder Networks,” arXiv preprint arXiv:1702.00307, 2017.

[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, 2019.

[23] J. Lei, X. You, and M. Abdel-Mottaleb, “Automatic Ear Landmark Localization, Segmentation, and Pose Classification in Range Images,” IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 46, no. 2, pp. 165–176, 2015.

[24] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, “Multi-view 3D Object Detection Network for Autonomous Driving,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1907–1915.

[25] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 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 Advances in neural information processing systems, 2017, pp. 5099–5108.
A. Pflug, A. Winterstein, and C. Busch, “Ear Detection in 3D Point Cloud Semantic Segmentation,” arXiv preprint arXiv:1807.00652, 2018.

W. Zeng and T. Gevers, “3DContextNet: K-d Tree Guided Hierarchical Learning of Point Clouds Using Local and Global Contextual Cues,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2018, pp. 0–0.

Z. Zhang, B.-S. Hua, and S.-K. Yeung, “ShellNet: Efficient Point Cloud Convolutional Neural Networks using Concentric Shells Statistics,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1607–1616.

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,” arXiv preprint arXiv:1905.05442, 2019.

Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen, “PointCNN: Convolution on X-Transformed Points,” Advances in neural information processing systems, vol. 31, pp. 820–830, 2018.

S. Wang, S. Suo, W.-C. Ma, A. Pokrovsky, and R. Urtasun, “Deep Parametric Continuous Convolutional Neural Networks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2589–2597.

A. Boulch, “ConvPoint: Continuous Convolutions for Point Cloud Processing,” Computers & Graphics, vol. 88, pp. 24–34, 2020.

H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, “KPConv: Flexible and Deformable Convolution for Point Clouds,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6411–6420.

J. Mao, X. Wang, and H. Li, “Interpolated Convolutional Networks for 3D Point Cloud Understanding,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1578–1587.

Q. Huang, W. Wang, and U. Neumann, “Recurrent Slice Networks for 3D Segmentation of Point Clouds,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2626–2635.

F. Engemann, T. Kontogiannis, A. Hermans, and B. Leibe, “Exploring Spatial Context for 3D Semantic Segmentation of Point Clouds,” in Proceedings of the IEEE international conference on computer vision workshops, 2017, pp. 716–724.

X. Ye, J. Li, H. Huang, L. Du, and X. Zhang, “3D Recurrent Neural Networks with Context Fusion for Point Cloud Semantic Segmentation,” in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 403–417.

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 Transactions On Graphics (tog), vol. 38, no. 5, pp. 1–12, 2019.

L. Landrieu and M. Simonovsky, “Large-scale Point Cloud Semantic Segmentation with Superpoint Graphs,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4558–4567.

L. Wang, Y. Huang, Y. Hou, S. Zhang, and J. Shan, “Graph Attention Convolution for Point Cloud Semantic Segmentation,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10 296–10 305.

J. Liu, B. Ni, C. Li, J. Yang, and Q. Tian, “Dynamic Points Agglomeration for Hierarchical Point Sets Learning,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7546–7555.

Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, “Deep Learning for 3D Point Clouds: A Survey,” IEEE transactions on pattern analysis and machine intelligence, 2020.

H. Chen and B. Bhanu, “Shape Model-Based 3D Ear Detection from Side Face Range Images,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)-Workshops. IEEE, 2005, pp. 122–122.

J. Zhou, S. Cadavid, and M. Abdel-Mottaleb, “Histograms of Categorized Shapes for 3D Ear Detection,” in 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, 2010, pp. 1–6.

S. Prakash and P. Gupta, “An Efficient Technique for Ear Detection in 3D: Invariant to Rotation and Scale,” in 2012 5th IAPR International Conference on Biometrics (ICB). IEEE, 2012, pp. 97–102.

A. Pfug, A. Winterstein, and C. Busch, “Ear Detection in 3D Profile Images Based on Surface Curvature,” in 2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing. IEEE, 2012, pp. 1–6.

M. Mur salin and S. M. S. Islam, “EpNet: A Deep Neural Network for Ear Detection in 3D Point Clouds,” in International Conference on Advanced Concepts for Intelligent Vision Systems. Springer, 2020, pp. 15–26.

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, 2020.

A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su et al., “ShapeNet: An Information-Rich 3D Model Repository,” arXiv preprint arXiv:1512.03012, 2015.

K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, “Mask R-CNN,” CoRR, vol. abs/1703.06870, 2017. [Online]. Available: http://arxiv.org/abs/1703.06870

P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, “A 3D Face Model for Pose and Illumination Invariant Face Recognition,” in 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. Ieee, 2009, pp. 296–301.

H. Dai, N. Pears, W. Smith, and C. Duncan, “Statistical Modeling of Craniofacial Shape and Texture,” International Journal of Computer Vision, vol. 128, no. 2, pp. 547–571, Nov 2019.

A. Mian, M. Bennamoun, and R. Owens, “An Efficient Multimodal 2D-3D Hybrid Approach to Automatic Face Recognition,” IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 11, pp. 1927–1943, 2007.

M. S. Bartlett, “Independent Component Representations for Face Recognition,” in Face Image Analysis by Unsupervised Learning. Springer, 2001, pp. 39–67.

S. Katz, A. Tal, and R. Basi, “Direct visibility of point sets,” in ACM Transactions on Graphics (TOG), vol. 26. ACM, 2007, p. 24.

G. Kumar. (2021) 3d ear recognition. [Online]. Available: https://github.com/gautamkumarjaiswal/3DEarRecognition

S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks,” in Advances in neural information processing systems, 2015, pp. 91–99.

W. Abdulla, “Mask R-CNN for Object Detection and Instance Segmentation on Keras and TensorFlow,” 2017.

A. Dutta and A. Zisserman, “The VIA Annotation Software for Images, Audio and Video,” in Proceedings of the 27th ACM International Conference on Multimedia, ser. MM ’19. New York, NY, USA: ACM, 2019.

P. Yan and K. Bowyer, “Empirical Evaluation of Advanced Ear Biometrics,” in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)-Workshops. IEEE, 2005, pp. 41–41.

P. Yan and K. W. Bowyer, “An Automatic 3D Ear Recognition System,” in Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPTV’06). IEEE, 2006, pp. 326–333.

———, “Biometric Recognition Using 3D Ear Shape,” IEEE Transactions on pattern analysis and machine intelligence, vol. 29, no. 8, pp. 1297–1308, 2007.

S. M. Islam, R. Davies, M. Bennamoun, and A. S. Mian, “Efficient Detection and Recognition of 3D Ears,” International Journal of Computer Vision, vol. 95, no. 1, pp. 52–73, 2011.

D. P. Kingma and J. Ba, “Adam: A method for Stochastic Optimization,” arXiv preprint arXiv:1412.6980, 2014.
MD MURSALIN is currently a PhD fellow at Edith Cowan University (ECU), Australia, and a recipient of the ECU Higher Degree by Research Scholarship (HDRS). He received the prestigious Chinese Government Scholarship for completing the master’s degree in computer science and engineering from the University of Jinan, China. He completed the undergraduate course in Computer Science and Information Technology from the Islamic University of Technology (IUT), Bangladesh. He worked as an Assistant Professor in Computer Science at Pabna University of Science and Technology, Bangladesh, and published several research articles in different reputed journals and conferences. His research interests span 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.

DR ISLAM completed his PhD with Distinction in Computer Engineering (CE) from the University of Western Australia (UWA) in 2011. He also received an MSc (CE) and BSc (EEE) in 2005 and 2000 respectively. He worked in different teaching and research positions at UWA and Curtin University in the past, and currently working as a Senior Lecturer in Computing at Edith Cowan University, Australia. He has supervised to completion five higher degree 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). He is an Associate Editor of IEEE Access, a regular reviewer of over 10 Q1 journals, Organising/Technical Committee member of over 35 conferences and Senior member of both IEEE and Australian Computer Society. His research interests include (but not limited to) Artificial Intelligence, medical Imaging, Biometrics, Machine Learning, Networking, and Computer Vision.