DeepPoint: A Deep Learning Model for 3D Reconstruction in Point Clouds via mmWave Radar

Yue Sun¹, Honggang Zhang², Zhuoming Huang³, and Benyuan Liu⁴

Abstract—Recent research has shown that mmWave radar sensing is effective for object detection in low visibility environments, which makes it an ideal technique in autonomous navigation systems such as autonomous vehicles. However, due to the characteristics of radar signals such as sparsity, low resolution, specularity, and high noise, it is still quite challenging to reconstruct 3D object shapes via mmWave radar sensing. Built on our recent proposed 3DRIMR (3D Reconstruction and Imaging via mmWave Radar), we introduce in this paper DeepPoint, a deep learning model that generates 3D objects in point cloud format that significantly outperforms the original 3DRIMR design. The model adopts a conditional Generative Adversarial Network (GAN) based deep neural network architecture. It takes as input the 2D depth images of an object generated by 3DRIMR’s Stage 1, and outputs smooth and dense 3D point clouds of the object. The model consists of a novel generator network that utilizes a sequence of DeepPoint blocks or layers to extract essential features of the union of multiple rough and sparse input point clouds of an object when observed from various viewpoints, given that those input point clouds may contain many incorrect points due to the imperfect generation process of 3DRIMR’s Stage 1. The design of DeepPoint adopts a deep structure to capture the global features of input point clouds, and it relies on an optimally chosen number of DeepPoint blocks and skip connections to achieve performance improvement over the original 3DRIMR design. Our experiments have demonstrated that this model significantly outperforms the original 3DRIMR and other standard techniques in reconstructing 3D objects.

1. INTRODUCTION

The advantage of Millimeter Wave (mmWave) radar in object sensing in low visibility environment has been actively studied recently in applying it in autonomous vehicles [1] and search/rescue in high risk areas [2]. However further application of mmWave radar in object imaging and reconstruction is quite difficult due to the characteristics of mmWave radar signals such as low resolution, sparsity, and high noise due to multi-path and specularity. Recent work [1]–[3] attempt to design deep learning systems to generate 2D depth images based on mmWave radar signals. 3DRIMR [4] further introduces a design that generates 3D object shapes based on mmWave radar, but the end results are still not satisfactory.

In this paper, we introduce DeepPoint, a deep learning model that generates 3D objects in dense and smooth point clouds based on the union of multiple rough and sparse input point clouds, which are directly converted from the 2D depth images generated by the Stage 1 of 3DRIMR which takes raw radar data as input. The training of DeepPoint follows conditional GAN architecture, and it significantly outperforms the original 3DRIMR design.

The 3DRIMR [4] architecture consists of two stages, and each has a generator network. Stage 1’s generator network \( G_{r2i} \) takes 3D radar intensity data as input and generates 2D depth images; Stage 2’s generator network \( G_{p2p} \) takes as input a set of multiple 2D depth images and outputs the 3D shape of the object in the format of point cloud. Each stage’s generator network is jointly trained with a separate discriminator network, using conditional GAN architecture.

Our major contributions are as follows:

1) DeepPoint, a novel generator network that can generate smooth and dense point cloud representation of a 3D object based on the union of multiple rough and sparse point clouds directly converted from the 2D depth images derived from raw mmWave radar sensor data. The generator network utilizes a sequence of DeepPoint blocks or layers to extract essential features of those input point clouds of an object when observed from various viewpoints, even though those input point clouds may contain many incorrect points due to the imperfect generation process of 3DRIMR’s Stage 1.

2) Novel designs such as a conditional GAN architecture design for the training of DeepPoint, an optimally chosen number of layers and skip connection. Those designs have resulted in the performance improvement over the original 3DRIMR.

3) An improved 3DRIMR system implementation that can conduct fast 3D object reconstruction by using a com-
modiﬁ cy mmWave radar sensor, instead of a slow full-scale SAR scan. The whole system takes advantage of convolutional operation and point cloud based neural network for efﬁ cient 3D shape generation with detailed geometry.

In the rest of the paper, we brieﬂ y discuss related work and preliminaries in Sections II and III. Then we discuss the design of DeepPoint model in Section IV. Experiment results are given in Section V. Finally the paper concludes in Section VI.

II. RELATED WORK

Frequency Modulated Continuous Wave (FMCW) Millimeter Wave (mmWave) radar sensing has been an active research area in recent years, especially in applications such as person/gesture identiﬁ cation [5], [6], car detection/imaging [1], and environment sensing [2], [3]. Usually Synthetic Aperture Radar (SAR) is used in data collection for high resolution, e.g., [7]–[10].

This paper is built on our recent work [4] on applying mmWave radar for 3D object reconstruction, in which we proposed 3DRIMR system. The deep neural network model proposed in this paper completely replaces the model in the Stage 2 of 3DRIMR, and this new model signiﬁ cantly outperforms the original 3DRIMR. There have been a few recent work on mmWave radar based imaging, mapping, and 3D object reconstruction [14]–[17], most of which use voxels to represent 3D objects. Our work is inspired by their promising research results, and due to the low cost and small form factor of commodity mmWave radar sensors, we plan to develop a simple and fast 3D reconstruction system to be attached in our UAV SLAM system [13] for search and rescue in dangerous environment.

Besides radar signals, vision community has also been working on learning-based 3D object shape reconstruction [14]–[17], most of which use voxels to represent 3D objects. Our proposed neural network model uses point cloud as a format for 3D objects to capture detailed geometric information with efﬁ cient memory and computation performance.

PointNet structure is utilized in PCN [11], which uses point cloud to reconstruct 3D object shapes, and this structure inspires us to design our model. The novelty of our model is that it has a deeper structure than PCN and skip connections are used for better capture of objects’ edges and shapes. In addition, our work adopts a conditional GAN architecture to jointly train a generator and a discriminator for better performance.

III. PRELIMINARIES

A. FMCW Millimeter Wave Radar Sensing and Imaging

Similar to [4], we use Frequency Modulated Continuous Wave (FMCW) mmWave radar sensor [18] signals to reconstruct 3D object shapes. Three Fast Fourier Transforms (FFTs) are conducted on received waveforms to generate 3D heatmaps or intensity maps of the space that represent the energy or radar intensity per voxel, written as \( z(\phi, \theta, \rho) \). Note that \( \phi \), \( \theta \), and \( \rho \) represent azimuth angle, elevation angle, and range respectively. Same as in [4], we use IWR6843ISK [19] operating at 60 GHz frequency, and for high resolution radar signals, we adopt the Synthetic Aperture Radar (SAR) operation. Unlike data from LiDAR and camera sensor, mmWave radar sensors can only give us sparse, low resolution, and highly noisy data. Partically, incorrect ghost points in radar signals can be generated due to multi-path effect. Reference [1]–[3] give more detailed discussion on FMCW mmWave radar sensing.

B. Representation of 3D Objects

In this work, we adopt point cloud format to represent 3D objects. Even though point cloud format is a standard representation of 3D objects and it is used in learning-based 3D reconstruction, e.g., [12], [20], [21], but CNN convolutional operation cannot be directly applied to a point cloud set as it is essentially an unordered point set. Furthermore, the point cloud of an object that is directly generated by raw radar signals is not a good choice to reconstruct the object due to the radar signal’s low resolution, being sparse, and with incorrect ghost points due to multi-path effect. Besides point clouds, voxel representations can also be used in 3D reconstruction [22]–[25], and the advantage of such representation is that 3D CNN convolutional operations can be applied to it. In addition, mesh representations of 3D objects are also used in existing work [26], [27]. However these two representation formats are limited by memory and computation cost.

C. Review of 3DRIMR Architecture

This paper introduces DeepPoint as the generator and discriminator networks of the Stage 2 of 3DRIMR to generate smooth and dense point clouds. For completeness, we now brieﬂ y review 3DRIMR architecture.

3DRIMR consists of two back-to-back generator networks \( G_{d2i} \) and \( G_{p2p} \). In Stage 1, \( G_{d2i} \) receives a 3D radar energy intensity map of an object and outputs a 2D depth image of the object. We let a mmWave radar sensor scans an object from multiple viewpoints to get multiple 3D energy maps. Then \( G_{d2i} \) generates multiple 2D depth images of the object. The Stage 2 of 3DRIMR pre-processes these images to get multiple coarse point clouds of the object, which are used as input to \( G_{p2p} \) to generate a single point cloud of the object. A conditional GAN architecture is designed for 3DRIMR’s training. That is, two discriminator networks \( D_{d2i} \) and \( D_{p2p} \) that are jointly trained together with their corresponding generator networks.

Let \( m_r \) denote a 3D radar intensity map of an object captured from a viewpoint, and let \( g_{zd} \) be a ground truth 2D depth image of the same object captured from the same viewpoint. \( G_{d2i} \) generates \( \hat{g}_{zd} \) that predicts or estimates \( g_{zd} \) given \( m_r \). If there are \( k \) different viewpoints \( v_1, \ldots, v_k \), generator \( G_{d2i} \) predicts their corresponding 2D depth images \( \{g_{zd,i} | i = 1, \ldots, k\} \). Each \( g_{zd,i} \) can be directly converted to a coarse and sparse 3D point cloud. Then we can have \( k \) coarse point clouds \( \{P_{zd,i} | i = 1, \ldots, k\} \) of the object. The Stage 2 of 3DRIMR unions the \( k \) coarse point clouds to form an initial estimated coarse point cloud of the object, denoted as
Block, and their designs on distance (EMD) in the loss function for the training of 3-dimensional matrix \( \hat{F} \).

Then, it applies a point-wise maxpooling on \( F_p \) and extracts a global feature vector \( g_{P_r} \). To produce a complete point cloud for an object, we need both local and global features, therefore, we concatenate the global feature \( g_{P_r} \) with each of the point features \( f_i \) and form another matrix \( F'_{P_r} \).

**Remarks.** Note that there are two major differences between the proposed generator network and the generator in Stage 2 of 3DMIMR [4]. First, the proposed generator is “deeper” than the generator of 3DMIMR as ours has more layers of DeepPoint blocks. Second, 3DMIMR uses fully connected layers and apply reshape operation to derive output point cloud from a high-dimensional global feature vector, which can only get an rough overall shape without many fine, local, and detailed characteristics. However, our new generator network design generates an output point cloud from both local and global features and hence can generate fine, local and detailed characteristics of an object.

C. Discriminator \( D_{p2p} \)

We design an improved discriminator with two-stream inputs, as shown in Fig. 1. The inputs pass through a simple shared MLP and get expanded into higher dimensional matrices. A point-wise max pooling and an average pooling are used to extract two global feature vectors. Then the two global feature vectors are concatenated to form final global features, which are further concatenated and fed into two fully connected layers to derive a score. The score is used to indicate whether the input is real or fake, i.e., generated point cloud.

D. Loss Function

When training the generator network, we concurrently train \( D_{p2p} \) to minimize \( L_{D_{p2p}} \), and train \( G_{p2p} \) to minimize \( L_{G_{p2p}} \). \( L_{D_{p2p}} \) is calculated as the mean MSE (Mean Square Error) of \( D_{p2p} \)’s prediction error. The loss function of generator \( L_{G_{p2p}} \) is a weighted sum, consisting of \( L_{GAN}(G_{p2p}) \), Chamfer loss \( L_{cf} \) between predicted point clouds and the ground truth, and EMD [11] loss \( L_{emd}(G_{p2p}) \).

Note that Chamfer distance [11] calculates the average closest distance between input and output points. The symmetric version of it is defined as:

\[
d_{cf}(S_1, S_2) = \frac{1}{|S_1|} \sum_{x \in S_1} \min_{y \in S_2} \|x - y\|_2 + \frac{1}{|S_2|} \sum_{y \in S_2} \min_{x \in S_1} \|y - x\|_2
\]

(1)

Then, Chamfer loss is defined as:

\[
L_{cf}(G_{p2p}) = d_{cf}(\hat{P}_r, P_{true})
\]

(2)

In addition, Earth Mover’s Distance (EMD) [11] can find a bijection \( \phi : S_1 \rightarrow S_2 \), which can minimize the average distance between pairs of corresponding points. The equation of EMD is:

\[
d_{emd}(S_1, S_2) = \min_{\phi : S_1 \rightarrow S_2} \frac{1}{|S_1|} \sum_{x \in S_1} \|x - \phi(x)\|_2
\]

(3)
Fig. 1: DeepPoint architecture proposed in this paper. The system first pre-processes multiple 2D depth images of an object to get multiple coarse and sparse point clouds, which may contain many incorrect points. Those depth images are generated for the same object but viewed from four different viewpoints. Combining those coarse point clouds we can derive a single coarse point cloud, which is used as the input of the generator $G_{p2p}$ of DeepPoint. The generator of DeepPoint outputs a dense and smooth point cloud representation of the object. A conditional GAN architecture is used to train both generator $G_{p2p}$ and discriminator $D_{p2p}$ jointly.

In our case, EMD loss is calculated as:

$$L_{emd}(G_{p2p}) = d_{emd}(\hat{P}_r, P_{true})$$  \hspace{1cm} (4)

$L_{G_{p2p}}$ is given by Eqn. (5). Note that $\lambda_{dcf}$ and $\lambda_{demd}$ are hand-tuned to 100 and 1 respectively in our experiments.

$$L_{G_{p2p}} = L_{GAN}(G_{p2p}) + \lambda_{dcf}L_{cf}(G_{p2p}) + \lambda_{demd}L_{emd}(G_{p2p})$$  \hspace{1cm} (5)

V. IMPLEMENTATION AND EXPERIMENTS

We implement DeepPoint and use it as Stage 2’s generator network of 3DRIMR [4] system. The system first generates 2D depth images from 3D radar intensity maps from multiple views of an object, and then passes these output depth images to the generator network of DeepPoint to produce a 3D point cloud of the object. We now present our experiment results in this section.

A. Datasets

We conduct experiments on cars with average size of $445cm \times 175cm \times 158cm$. The input data to the proposed generator network is the output depth images produced by 3DRIMR’s Stage 1. We follow a procedure that is similar to 3DRIMR [4], to generate ground truth point clouds. Fig. 2 shows an example scene.

B. Model Training and Testing

We use the 2D depth images generated in the Stage 1 of 3DRIMR to form a dataset of coarse and sparse point clouds, which includes 1600 point clouds with 200 point clouds of each car model. We train the proposed generator network and discriminator network for 200 epochs using 1520 point clouds with batch size 4. The learning rate for the first 100 epochs is $2 \times 10^{-4}$ and linearly decreases to 0 in the rest 100 epochs. Then we test the generator network using the remaining 80 point clouds. Fig. 3 shows some example results of generated 2D depth images from the Stage 1 of 3DRIMR.

C. Evaluation Results

In this section, we compare our generator network performance with the original 3DRIMR [4] since both of them aim at reconstructing objects’ 3D point clouds from sparse radar data. We also validate our design choices for generator and discriminator networks by controlled experiments. We conduct experiments by varying the number of layers, i.e., DeepPoint blocks, in our proposed generator network model. Specifically, as shown in Table I, we tested 1, 2, 5 and 7 layers of DeepPoint blocks in our generator network.

1) Comparison with 3DRIMR: In Table I, we can see the our generator network with 5-Block and 7-Block significantly outperforms the original 3DRIMR in terms of

| Method          | CD   | EMD  | F-score |
|-----------------|------|------|---------|
|                 | avg. | std. | avg.    | std. | avg. | std. |
| 3DRIMR          | 7.89 | 4.11 | -       | -    | 8.41 | 3.22 |
| 1-Block, w/o sc| 10.10| 4.49 | 5.01    | 4.24 | 8.40 | 3.40 |
| 2-Block         | 9.75 | 4.00 | 4.56    | 3.96 | 8.47 | 3.44 |
| 3-Block         | 9.40 | 4.70 | 4.83    | 4.84 | 9.40 | 4.22 |
| 5-Block         | 7.79 | 4.37 | -       | -    | 13.10| 5.97 |
| 7-Block + 3sc   | 7.68 | 4.15 | 4.53    | 4.19 | 13.23| 6.34 |

TABLE I: Quantitative results under different setups. Note that the units of CD and EMD in this table are cm, and the magnitude of F-score is $10^{-2}$. 

Fig. 3: Example results of generated 2D depth images from the Stage 1 of 3DRIMR. The generated images are used as inputs to DeepPoint (which is used as the Stage 2 of 3DRIMR). The 1st row shows the 3D radar intensity data from 2 snapshots only. The 2nd row shows the outputs from 3DRIMR’s Stage 1. The 3rd row shows the ground truth depth images.

both Chamfer Distance (CD) and F-score. Note that EMD results are not available for 3DRIMR. In addition, Fig. 4 demonstrates that the visual improvement of output point clouds is even more obvious. This is because 3DRIMR can only reconstruct an overall shape of the object whereas our proposed generator network can recover more fine details of the shape, e.g., correct orientation, and the shapes of wheels of the car. This significant improvement is due to the “deeper” structure of the generator network, the optimal number of DeepPoint blocks, the introduction of skip connections, and the use of a more efficient training loss metric, i.e., Earth Mover’s Distance (EMD).

2) Performances of different layers of DeepPoint Blocks:
The DeepPoint blocks in our generator network first expand each input point’s dimension and then shrink them back to 3 to get each output point’s coordinates. The more such DeepPoint blocks means the generator network is “deeper”, which seems achieve better performance. However, this is not always true. There exists an optimal number of layers. As shown in Fig. 5, the performance of our generator improves with the number of DeepPoint blocks increasing, i.e., CD decreases from 9.8 cm to 7.8 cm and EMD decreases from 4.6 cm to 4.4 cm as the number of DeepPoint Blocks increases from 1 to 5. Correspondingly, F-score is even improved around 55%. However, after the number of DeepPoint blocks reaches an upper bound, say 5 in our experiment, further increasing the number of DeepPoint blocks to 7 can no longer largely improve the performance. We can see that all these 3 evaluation metrics are very similar in these two cases.

3) Skip Connections in Generator Network: As shown in Fig. 1, we can see inside each DeepPoint block, we concatenate the raw input points’ Cartesian coordinates with feature matrix obtained by passing through the shared MLP to further form the point features. Our experiment results (e.g., Table I) clearly show that such skip connection design can improve the performance. However, blindly increasing skip connections will not always help. As shown in another experiment in which we build 3 more skip connections by concatenating the 1st and 7th point features, 2nd and 6th point features, and 3rd and 5th point features respectively. However, based on the results shown in Table I, we can see that with these additional skip connections, generator performs worse compared with the case without using them. In our future work, we will investigate an optimal placement of skip connections.

4) Variants of Discriminator:
In our discriminator network design, we use mix pooling to extract the global feature vectors. Mix pooling means that we concatenate the feature vectors from both max pooling and average pooling. We compare the results of using max pooling, average pooling and mix pooling in the discriminator in Table II. Note that the generators in these 3 experiments are the same. Table II shows that mix pooling performs best among these three pooling methods, and max pooling falls a little behind average pooling method.

| Method       | CD avg. | CD std. | EMD avg. | EMD std. | F-score avg. | F-score std. |
|--------------|---------|---------|----------|----------|--------------|--------------|
| Mix Pooling  | 9.75    | 4.00    | 4.56     | 3.96     | 8.47         | 3.44         |
| Max Pooling  | 10.52   | 4.80    | 4.79     | 4.33     | 7.29         | 2.71         |
| Average Pooling | 10.28   | 4.11    | 4.60     | 4.11     | 7.71         | 3.18         |

TABLE II: Quantitative Results of Stage 2 using different pooling methods in the discriminator. Note that the units of CD and EMD in this table are cm, and the magnitude of F-score is $10^{-2}$. 
VI. CONCLUSIONS AND FUTURE WORK

We have proposed DeepPoint, a deep learning model that generates 3D objects in smooth and dense point clouds. It utilizes a sequence of novel DeepPoint blocks to extract essential features of the union of multiple rough and sparse input point clouds of an object when observed from various viewpoints, even though the inputs contain many incorrect points. It relies on a deep structure design, an optimally chosen number of DeepPoint blocks, and skip connections to achieve good 3D reconstruction performance. For future work, we will find the optimal placement of skip connections and introduce new techniques to improve the detailed geometry of generated point clouds. We will also conduct large scale experiments to improve our design.
