A Novel Point Cloud Generation Method based on Monocular Images and FMCW Radar

Zecheng Li[1], Yuying Song[1], Fuyuan Ai[1], Jingxuan Wu[1], Chunyi Song[1,2,3*], Zhiwei Xu [1,2,3]
[1] Institute of Marine Electronic and Intelligent System, Ocean College, Zhejiang University
[2] Engineering Research Center of Oceanic Sensing Technology and Equipment, Ministry of Education
[3] Donghai Laboratory
{lizechng, yysong, fyai, 22034223, cysong, xuzw}@zju.edu.cn

Abstract

Millimeter-wave (mmWave) radar is widely used in autonomous driving because of its capability for harsh weather conditions. With the recent advancement in hardware, mmWave radar technology is now commonly employed in high-level perception tasks of mobile robots and autonomous driving. However, compared to LiDAR point clouds, mmWave radar point clouds have limitations. These point clouds tend to be sparse and include several "ghost" targets. Nevertheless, monocular images, which provide efficient semantic information, can improve the accuracy and density of the generated point cloud. Therefore, in this study, we propose a novel point cloud generation method based on monocular images and mmWave radar, and define a new evaluation metric. Our data collection and experimentation in real-world scenarios demonstrate that our method outperforms other techniques in producing high-quality radar point clouds. Moreover, the generated radar point cloud can be applied to object detection, localization, and mapping tasks, which will contribute to the advancement of autonomous driving perception.

1. Introduction

In recent years, 77GHz millimeter wave radar technology has rapidly emerged as a cost-effective alternative for advanced driver assistance systems [14]. This technology is used for various applications, including autonomous emergency braking and adaptive cruise control. Furthermore, millimeter-wave radar is highly robust in different lighting and weather conditions and provides accurate measurements of distance, azimuth, and instantaneous velocity [27].

Advancements in integrated circuit technology have significantly increased the integration level of millimeter-wave radar in recent years. The continuous improvements in RF technology have led to enhanced performance and reduced manufacturing costs of radar systems over time [1]. As a result, millimeter-wave radar has gained widespread applications in various domains, such as mobile robots and autonomous driving [1, 5, 7, 10, 17]. Furthermore, millimeter-wave radar plays a crucial role in intelligent perception during adverse weather conditions, including rain and fog [9].

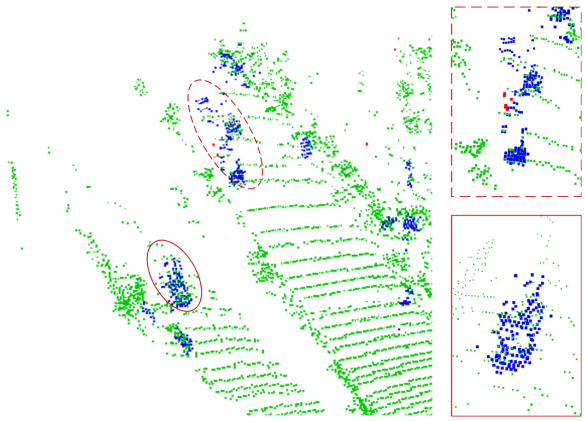


Figure 1. Comparison of LiDAR points (green), radar points generated by OS-CFAR (red), and pseudo radar points generated using our proposed method (blue). The red circles represent the vehicle points.

Millimeter-wave radar point clouds are an important data format in autonomous driving scenarios, similar to LiDAR point clouds. They are used for environmental mapping, localization and object detection [13]. However, unlike LiDAR point clouds, typical millimeter-wave radar point clouds have four limitations: Sparsity [2]; Limited vertical field of view [15]; Clutter points [4]; Viewpoint variation and temporal variation [3].

Millimeter-wave signals exhibit adaptability in various environments due to their wavelength. However, the aforementioned limitations restrict further application of

*Corresponding author

millimeter-wave radar point clouds in perception for autonomous driving environments.

The limitations of mmWave radar point clouds are primarily due to two factors [27]: the inherent physical characteristics of the hardware and the inadequacies of classical radar signal processing algorithms. The mmWave wavelength results in high surface reflection on objects, but not all reflections from objects propagate back to the mmWave receiver, unlike light signals that scatter in all directions [2, 17]. Additionally, the low-cost single-chip design of mmWave radar results in fewer antennas and lower angular resolution. In response to these issues, some teams [4, 23, 26] have attempted to optimize mmWave radar signal processing algorithms to generate higher quality point clouds. But limited by the information dimensionality of mmWave signals, these methods can only meet the needs of some simple scenarios, which is still inadequate for autonomous driving scenes. Some researchers [12, 20] have proposed sensor fusion solutions to comprehensively utilize the complementary advantages of different sensors. For instance, radar can directly obtain distance and radial velocity information of objects, and it can locate objects on a two-dimensional plane parallel to the ground. However, radar sensors cannot obtain height information. Cameras provide rich appearance features, but they are not a good source of depth estimation. The features obtained from cameras and radars are complementary. However, current camera-radar fusion methods mainly apply to downstream application tasks of point clouds, such as 3D reconstruction, object detection, etc. These fusion algorithms can only target specific perception tasks and are difficult to directly apply in different tasks. Generating mmWave radar point clouds of higher quality using camera-radar fusion methods to meet the requirements of various perception tasks is still a relatively unexplored area.

The aim of our work is to generate high-quality millimeter-wave radar point clouds for autonomous driving scenarios. To this end, we propose a novel method that combines camera and millimeter-wave Radar BEV images to produce superior millimeter-wave radar point clouds. Our approach consists of three modules: Hessian matrix based peak enhancement (HPE), Affinity matrix based BEV feature diffusion (AFD), and graph based point correction (GPC). In addition, we introduce a new evaluation metric specifically designed for evaluating the quality of millimeter-wave radar point clouds in autonomous driving perception tasks. To validate the proposed method, we construct a real-world dataset, including urban streets, elevated highways, and tunnels. Experiments demonstrate that our approach outperforms existing methods, resulting in fewer clutter points and denser real point clouds. Furthermore, the radar point clouds generated by the proposed method can be applied to object detection, localization, and mapping tasks, which will contribute to the advancement of autonomous driving perception.

In summary, this article has made contributions to the following aspects:

- To the best of our knowledge, our research is the first to concentrate on enhancing the quality of point clouds through the camera-radar fusion method. This is a fundamental issue that has not been extensively explored when using millimeter-wave radar for autonomous driving. We conducted experiments to comprehensively analyze this problem.
- In this paper, we introduce a novel method that fuses camera and millimeter-wave data to generate high-quality millimeter-wave radar point clouds for autonomous driving's environmental perception tasks. Additionally, we define a new metric to evaluate the quality of millimeter-wave radar point clouds for autonomous driving applications.
- We developed a prototype platform and demonstrated the effectiveness of our approach through real-world scenario-based experiments. Furthermore, we collected a raw dataset that includes diverse scenarios, such as urban streets, highways, and tunnels, to aid in research on environment perception for autonomous driving.

2. Related Work

2.1. Radar Point Cloud Generation

The generation of effective and accurate millimeter-wave (mmWave) radar point clouds typically involves two steps: raw point cloud generation and point cloud postprocessing. Some researchers [8,25] have focused on improving the signal processing algorithm to increase the angular resolution of the radar point cloud. Others [4, 23, 26] have explored postprocessing techniques. For instance, this work [25] applied the concept of synthetic aperture radar imaging to improve the side-view point cloud quality by increasing the radar point cloud angular resolution in a two-chip cascade automotive radar. Additionally, some researchers [4] have used modified PointNet network architecture to classify real targets and ghost targets in mmWave radar point clouds. However, it should be noted that postprocessing improvements are limited as much information is lost during raw radar point cloud generation. To enhance the quality of mmWave radar point clouds, we propose a novel approach that combines camera and mmWave RF images to generate superior point clouds, as discussed in the previous section.

2.2. Pseudo Point Cloud Generation

Several recent approaches [19, 29, 31] have greatly improved the performance of the monocular 3D perception

task. A common feature among them is that they first estimate the depth map from the input RGB image and then transform it into pseudo-LiDAR (point cloud) using the camera calibration information. The generated pseudo-LiDAR signals are then directly processed using off-the-shelf LiDAR-based methods [11, 24]. For example, AM3D [19] has proposed a multi-modal feature fusion module that embeds the complementary RGB cues into the generated pseudo-LiDAR representation. Additionally, this work [19] has proposed a depth prior-based background point segmentation module to mitigate the issues caused by the inaccuracy of point cloud annotation. Another method proposed by [31] uses a 2D-3D bounding box consistency loss that alleviates the local misalignment issue. However, these methods rely heavily on the accuracy of the depth map. Although pseudo-LiDAR based methods have achieved impressive accuracy in 3D perception tasks, the underlying mechanism is still not well understood in the research community.

3. The proposed method

Our proposed method involves several steps to generate a high-quality pseudo radar point cloud for autonomous driving scenarios, as shown in Fig. 2. Firstly, we convert the range-azimuth heatmap (RAM) into radar BEV images, and enhance them using the Hessian matrix to improve the target reflection intensity. Additionally, we generate BEV representations from the monocular images using the LSS model [22]. Subsequently, we feed the radar BEV images into the radar encoder to obtain the radar encoded features and radar BEV affinity matrix. The latter can complete the sparse monocular BEV features. We then combine the monocular BEV features with the radar BEV features and pass them to the BEV decoder to generate the pseudo radar point cloud. Finally, we optimize the pseudo radar point cloud using KNN graph with radar ground-truth points to achieve better accuracy and density.

3.1. BEV enhancement using Hessian matrix

The strength of the range-azimuth heatmap for the radar data indicates the probability of the presence of objects within the BEV space, where the energy of the target area is higher compared to empty areas. The radar data is transformed from polar to Cartesian coordinates through bilinear interpolation. Likewise, the LiDAR data is also converted from a point cloud to a Cartesian coordinate system as the ground-truth label matrix. Moreover, to align the LiDAR matrix with the reflection principle of radar, we apply RANSAC [6] to filter out ground points in the LiDAR data.

The radar BEV images obtained after conversion are often fuzzy. The responses of targets such as cars, bicycles, and pedestrians are ambiguous. To address this issue, we propose an algorithm for enhancing radar BEV images using the Hessian matrix, as depicted in Algorithm 1. The Hessian matrix is a matrix of second-order partial derivatives that can identify the local curvature of an image. Our algorithm starts by smoothing the grayscale radar BEV image $I_r(x, y)$ using a Gaussian filter $G(x, y)$, which can reduce the noise in the image. Then, we slide a square kernel of size $n \times n$ over the smoothed image to detect peaks. If a peak is detected, we calculate the Hessian matrix $H(x, y)$ at that pixel and use it to enhance the peak. We also smooth the peak area using a Gaussian filter and compute the energy of the peak E_p. Finally, we scale the smoothed peak by the energy value and return the enhanced radar BEV image $I_e(x, y)$.

Algorithm 1 Peak Enhancement Algorithm

Require: $I_r(x, y)$
Ensure: $I_e(x, y)$
$I_g(x, y) = I_r(x, y) * G(x, y)$
for each pixel (x, y) in $I_g(x, y)$ **do**
  if $I_g(x, y)$ is a peak **then**
    $I_e(x, y) = I_g(x, y) * H(x, y)$
    $I_{s,p}(x, y) = I_e(x, y) * G(x, y)$
    $E_p = \sum_{i,j} I_{s,p}(i, j)^2$
    $I_e(x, y) = I_{s,p}(x, y) * E_p$
  else
    $I_e(x, y) = I_g(x, y)$
  end if
end for
Return $I_e(x, y)$

3.2. BEV features diffusion based on affinity matrix

Similar to the LSS Model proposed by [22], we assume that $X \in R^{3\times H\times W}$ represents an image with extrinsic $\mathbf{E}$ and intrinsic $\mathbf{I}$ parameters, and that p denotes a pixel in the image with coordinates (h, w). To each pixel, we associate $|D|$ points $\{(h, w, d) \in R^3 | d \in D\}$, where D is a set of discrete depths defined by $\{d_0 + \triangle, ..., d_0 + |D|\triangle\}$. Note that this transformation does not involve any learnable parameters. Simply put, we create a large point cloud with size $D \times H \times W$ for a given image. For each point in the point cloud, the context vector is parameterized to match a notion of attention and discrete depth inference. At pixel p, the network predicts a context $\mathbf{c} \in R^C$ and a depth distribution $\alpha \in \triangle^{|D|-1}$ for each pixel. We define the feature $\mathbf{c}_d \in R^C$ associated with point p_d as the context vector for pixel p scaled by α_d:

$$\mathbf{c}_d = \alpha_d \mathbf{c}$$

However, the sparsity of monocular BEV features poses a challenge when fusing them with radar BEV features. To address this issue, we propose a novel approach that completes the BEV features based on the radar BEV affinity matrix. Specifically, given the monocular BEV feature map

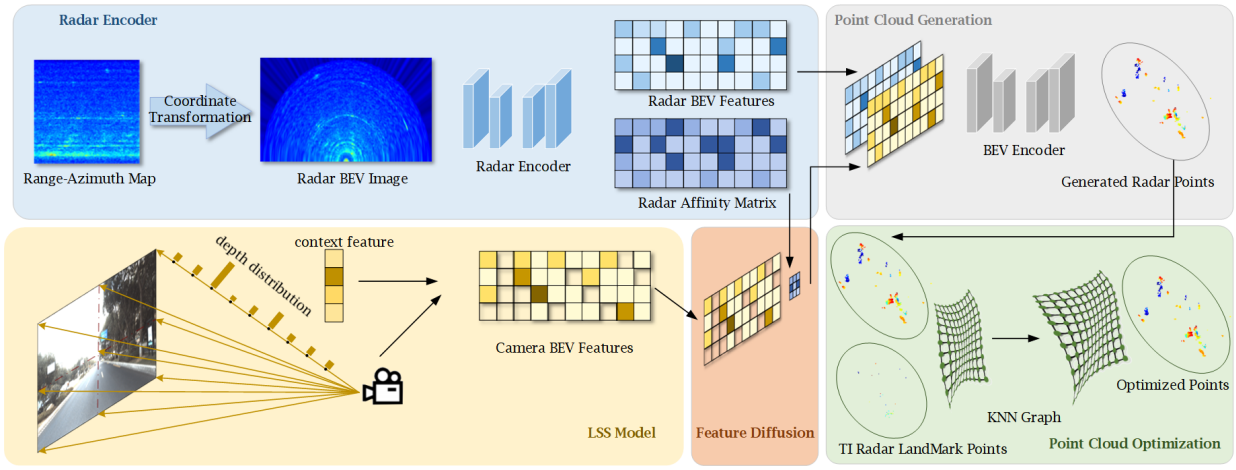
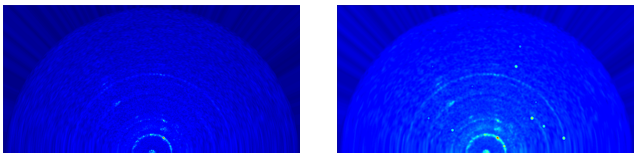
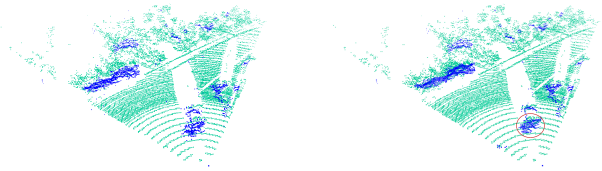


Figure 2. The whole pipeline of the proposed method.



(a) (b)

Figure 3. (a) Radar BEV image w/o enhancement. (b) Radar BEV image through hessian matrix, which significantly improves the clarity of the radar signal reflections.



(a) (b)

Figure 4. Comparision of mmWave radar point clouds generated with and without the GPC module. By using the GPC module, the target points (shown in the red circle) can be better optimized. (a) Radar point cloud w/o GPC module. (b) Radar point cloud with GPC module.

$F \in R^{m\times n\times c}$ and the corresponding radar affinity matrix $A \in R^{m\times n\times 8}$, we complete the monocular BEV features using N iterations. Our approach not only enhances feature details but also improves per-pixel features.

Without loss of generality, we can express the convolutional transformation functional for each time step t with a kernel size of k as follows:

$$\mathbf{F}_{i,j,t+1} = \kappa_{i,j}(0,0)\mathbf{F}_{i,j,0} + \sum_{\substack{a,b=-(k-1)/2 \\ a,b\neq 0}}^{(k-1)/2} \kappa_{i,j}(a,b)\odot\mathbf{F}_{i-a,j-b,t}$$

where,

$$\kappa_{i,j} = \frac{\hat{\kappa}_{i,j}(a,b)}{\sum_{a,b\neq 0}|\hat{\kappa}_{i,j}(a,b)|}$$

$$\kappa_{0,0} = 1 - \sum_{a,b\ neq 0} \kappa_{i,j}(a,b)$$

The transformation kernel $\hat{\kappa}_{i,j} \in \mathbf{R}^{k\times k\times c}$ is computed by an affinity network, which is spatially dependent on the input image. The kernel size k is typically set as an odd number to ensure the symmetry of the computational context around pixel (i,j). The element-wise product of $\odot$ is applied here. To stabilize the model when the condition $\sum_{a,b\neq 0}|\kappa_{i,j}(a,b)| \leq 1$ is met, we normalize the kernel weights to the range of $(-1,1)$, as proposed in [16]. Finally, we perform N iterations until a stable state is reached.

3.3. Point cloud optimization through KNN graph

Our method significantly improves the quality of the pseudo radar point cloud by more precisely rendering object contours. However, Fig. 4 shows that some objects may be missed or malposed due to noise and unclear features in the radar bev images and occlusions in the monocular images.

This is because we are only using a portion of the information from the radar, resulting in the loss of some targets. CFAR utilizes all the information from the radar data, and generates point clouds that, while sparse, still retain some points. Although this problem can be alleviated with increasing the dimensionality of the radar BEV data, but this would place high demands on the GPUs.

We explore a hybrid approach to correct the bias by leveraging radar ground-truth (GT) points. While radar GT points are too sparse to capture object shapes and cannot be used alone for detection, they can be used as "landmark" points. We present a graph-based point correction (GPC)

algorithm, which effectively integrates pseudo radar point clouds that render object shapes with sparse, accurate radar ground-truth points.

We utilize two point clouds, namely Radar (R) and Pseudo-Radar (PR), as inputs to our method. To begin, we analyze the local shapes in the PR point cloud by establishing a directed K-nearest-neighbor (KNN) graph that connects each 3D point to its KNNs with appropriate weights. We assume that ground-truth radar points are available for the first n points, while the remaining m points have no ground-truth data. The pseudo-radar points are represented by $Z \in R^{n+m}$, while the radar ground-truth points is represented by $G \in R^n$.

Let N_i denote the set of neighbors for the i^{th} point. Additionally, we define a weight matrix, $W \in R^{(n+m)\times(n+m)}$, where W represents the edge-weight between points i and j. Drawing inspiration from prior work in manifold learning [30], we select the weights as coefficients that reconstruct the points from their respective neighbors in N_i. We can obtain these weights by solving a constrained quadratic optimization problem as follows:

$$W = argmin_W ||Z - WZ||_2^2$$

$$s.t. W\mathbf{1} = \mathbf{1} \text{ and } W_{ij} = 0 \text{ if } j \notin N_i$$

Here, $\mathbf{1} \in R^{n+m}$ denotes the all-ones vector. For the condition that the points are in general positions and $k > 3$, there exist infinitely many solutions that satisfy $Z = WZ$. To ensure robustness, we select the solution with the minimum L_2 norm.

Let us define the corrected points as $Z' \in R^{n+m}$, where $Z' = [Z'_R; Z'_{PR}]$, $Z'_R \in R^n$, and $Z'_{PR} \in R^m$. For the n points with ground-truth radar measurements, we update the points Z'_R to be equal to G. Next, we solve for Z'_{PR} by using G and the weighted KNN graph encoded in W. Specifically, we update the remaining points Z'_{PR} so that the point i can be accurately reconstructed as a weighted sum of its KNN neighbors' points using the learned weights W. We can obtain the final Z' directly by solving another quadratic optimization problem as follows:

$$Z' = argmin_{Z'} ||Z'W - Z'||^2, s.t. Z'_{1:n} = G$$

4. Experiment and Evaluation

In this section, we introduce our vehicle acquisition platform and dataset, which comprises 70 sequences collected in three different scenes: urban streets, elevated highways, and tunnels. Furthermore, we propose a new evaluation metric specifically designed for mmWave radar point clouds and perform a quantitative evaluation of various radar point cloud generation methods. The results demonstrate that our method generates higher quality point clouds. Finally, we conduct several ablation studies to evaluate the contribution of each component in our approach.

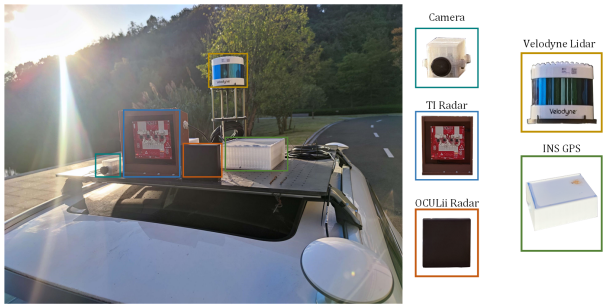


Figure 5. The acquisition platform and equipment

.

4.1. Dataset

The acquisition equipment details are presented in Table 1, while Fig. 5 illustrates the acquisition platform. Our dataset comprises manually selected data from 70 different scenarios including urban streets, elevated highways, and tunnels, in accordance with the nuScenes settings. Each scene is a 20-second sequence, and its keyframes are sampled as training data. All image data is saved at a resolution of 512x1024. Ultimately, we sample 37 scenes (1774 frames) for the training set, and 33 scenes (1744 frames) for the test set.

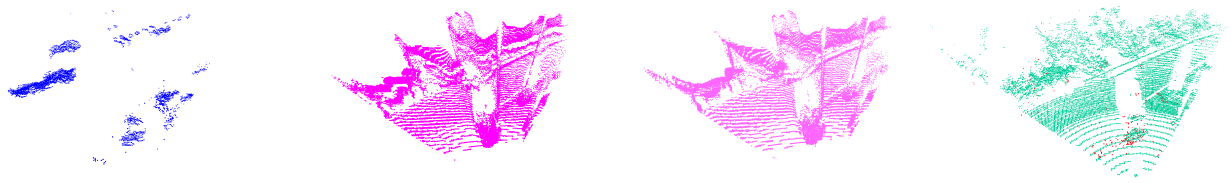
Sensors	Details
Camera	(Leopard LI-IMX490-GMSL2-90H) RGB, 20Hz capture frequency, 1876×1896 resolution, auto exposure, JPEG compressed
LiDAR	(Velodyne VLS-128) Spinning, 128 beams, 10Hz capture frequency, 360° horizontal FOV, $-25°$ to $15°$ vertical FOV, up to 2.4M points per second
Radar	(TI) $\leq 250m$ range, 77GHz, FMCW, 13Hz capture frequency, $\pm 0.1km/h$ vel. accuracy
GPS/IMU	(NovAtel PwrPak7D) GPS, IMU, AHRS, 0.2 heading, 0.1 roll/pitch, 20mm RTK positioning, 20Hz update rate

Table 1. The detailed parameters of the acquisition equipment.

4.2. Evaluation Metrics

Typically, high-quality point clouds are essential for autonomous vehicles. Our goal is to make the radar point clouds closer to the LiDAR point clouds. Therefore, we use the LiDAR point clouds as the ground truth and introduce a new evaluation metric to evaluate the quality radar point clouds generated by different methods.

We define the sets of LiDAR and radar point clouds in a single frame as Ω_L and Ω_R, respectively. The distance



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

Figure 6. Comparison of point clouds generated by different methods. (a) Radar point cloud generated by our method. (b) Pseudo point cloud generated by [28]. (c) Pseudo point cloud generated by [18] (d) LiDAR points (Green) and TI radar points (Red). The radar point cloud generated by our method can more precisely render object contours.

Methods	Input Pattern	$S_{recall}\uparrow$	$S_{clutter}\downarrow$	$D_H\downarrow$	$D_C\downarrow$
OS-CFAR [32]	Radar	4.190	0.303	8.727	3.823
FusionNet [28]	Camera + Radar	0.126	0.313	7.306	0.840
Sparse to dense [18]	Camera + Radar	0.143	0.248	7.306	0.921
ours	Camera + Radar	1.388	**0.052**	**2.960**	2.365

Table 2. Comparison to the competing methods.

between two points, $\mathbf{p}$ and $\mathbf{q}$, in Euclidean space is denoted by $d(\mathbf{p},\mathbf{q}) = \sqrt{(p_x-q_x)^2+(p_y-q_y)^2+(p_z-q_z)^2}$.

The subset of clutter points in the radar point cloud Ω_R, denoted by Ω_{R1}, can be defined as follows:

$$\Omega_{R1} = \{\mathbf{q} \in \Omega_R, s.t. d(\mathbf{q},\mathbf{p}) > \delta, \forall \mathbf{p} \in \Omega_L\}$$

The parameter δ is the distance threshold that we set. As the distance of point $\mathbf{q}$ increases, the value of δ also increases. This implies that for points that are farther apart, a looser threshold is set. If no LiDAR point is within a distance of δ from this point, the corresponding radar point is classified as a clutter point. A smaller number of clutter points indicates that the method performs better in reducing radar clutter.

We define Ω_{L1}, which is a subset of Ω_L as

$$\Omega_{L1} = \{\mathbf{p} \in \Omega_L, s.t. d(\mathbf{q},\mathbf{p}) < \delta, \exists \mathbf{q} \in \Omega_R\}$$

Ω_{L1} is used to evaluate the density of the generated radar point clouds of real targets in our proposed method. In this regard, two metrics are defined to evaluate the performance of the proposed method: $S_{clutter}$ and S_{recall}. $S_{clutter}$ represents the severity of clutter points, while S_{recall} represents the similarity between the generated radar points and LiDAR points. The equations for $S_{clutter}$ and S_{recall} are as follows:

$$S_{clutter} = \frac{N(\Omega_{R1})}{N(\Omega_R)}$$

$$S_{recall} = \frac{N(\Omega_{L1})}{N(\Omega_R)}$$

Here, $N(\Omega_{L1})$ and $N(\Omega_{R1})$ represent the number of points in Ω_{L1} and Ω_{R1}, respectively. A higher S_{recall} and a lower $S_{clutter}$ indicate better performance of the method as it implies that the radar point clouds are more similar to the LiDAR point clouds. Additionally, we calculate the Hausdorff distance D_H between the generated radar point cloud and the LiDAR point cloud, which provides a global measure of similarity, and the Chamfer distance D_C, which provides a local measure of similarity.

$$D_H = \max\{\sup_{\mathbf{p}\in\Omega_L}\inf_{\mathbf{q}\in\Omega_R} d(\mathbf{q},\mathbf{p}), \sup_{\mathbf{q}\in\Omega_R}\inf_{\mathbf{p}\in\Omega_L} d(\mathbf{q},\mathbf{p})\}$$

$$D_C = \frac{1}{|\Omega_R|}\sum_{\mathbf{q}\in\Omega_R} \min_{\mathbf{p}\in\Omega_L} d(\mathbf{q},\mathbf{p}) + \frac{1}{|\Omega_L|}\sum_{\mathbf{p}\in\Omega_L} \min_{\mathbf{q}\in\Omega_R} d(\mathbf{q},\mathbf{p})$$

4.3. Training Details

All models used in our experiments were trained with a batch size of 4 and the Adam optimizer, with a learning rate of 0.001 and a momentum of 0.9, for 40 epochs. We implemented all models in PyTorch [21], and the experiments were conducted on desktop computers equipped with Nvidia GTX 1080Ti GPUs.

The encoder-decoder we use is modified from LSS [22] with the same sample rate. The size of the image is maintained at 512×1024, while the radar BEV image is maintained at 300×600. Data augmentation including random flips, scaling, and color dithering are also adopted to avoid overfitting.

4.4. Quantitative Evaluation

We compared our method with three others using the evaluation metrics we established. Our comparison in-

Methods	Input Pattern	$S_{recall}\uparrow$	$S_{clutter}\downarrow$	$D_H\downarrow$	$D_C\downarrow$	Points Num
OS-CFAR [32]	Radar	4.190	0.303	8.727	3.823	211
Baseline(LSS) [22]	Camera	1.456	0.193	4.903	5.982	575
+ Radar data	Camera + Radar	1.220	0.218	4.293	4.606	1255
+ HPE	Camera + Radar	1.422	0.114	8.069	3.177	1608
+ AFD	Camera + Radar	1.377	0.174	7.031	2.493	2062
+ GPC	Camera + Radar	1.388	0.052	2.960	2.365	2000

Table 3. Ablation Study.

cludes classic radar detectors, OS-CFAR [32], as well as two depth completion methods that generate pseudo point clouds through back projection.

We conduct experiments with distance thresholds

$$\delta = \begin{cases} 0.5m & \text{distance} \in [0m, 40m] \\ 1m & \text{distance} \in [40m, 60m] \\ 1.5m & \text{distance} \in [60m, 75m] \end{cases}$$

The evaluation result is shown in Table 2. It can be observed that the $S_{clutter}$ of our method is lower than that of other methods. Higher values of S_{recall} correspond to generated points located on dense LiDAR points areas, indicating a higher probability of target occurrence. Our method has a higher S_{recall} value when compared to that of depth completion methods, suggesting that our method focused on target points rather than the background. We can also see that on the Fig. 6. The Hausdorff distance D_H and Chamfer distance D_C provide global and local measures of points cloud similarity. The experiments demonstrate that the radar point cloud generated by our method has the best global similarity with the LiDAR point cloud. However, the local similarity of our method is lower compared to that of depth completion methods, which is largely because the point cloud generated by our method don't contain ground points.

4.5. Ablation Study

To evaluate the contribution of different components of our method, we perform ablation studies on our dataset.

1) Baseline: The LSS model [22] is taken as the baseline to convert camera features to bev-view features, and to generate the coarse radar point cloud using BEV encoder.

2) Radar BEV images: LSS model generates many clutter points in the radar point cloud, resuting in lower global and local similarity with LiDAR point cloud. Thus, we transform the radar data from polar to Cartesian coordinate to align with camara features, and take it as the additional data. It can be seen that the radar data improves the performance of the model. With the radar data, the D_C decreases approximately 23% on average.

3) HBE module: The HBE module enhances the clarity of the BEV image through the Hessian matrix, making the reflection areas clear and improving the quality of the point clouds, as shown in Fig. 3. By adding HBE module, the number of generated radar points increase by 353 points. And the $S_{clutter}$ decreases by 10.4%.

4) AFD module: The AFD module performs the BEV features diffusion of monocular images based on the radar BEV affinity matrix, resulting in an equivalent number of generated radar points as the 4D-radar, although adding some clutter points.

5) GPC module: The radar point clouds generated by the BEV encoder are processed using the GPC postprocessing module to generate the final radar point clouds. The GPC module significantly improves the quality of point clouds. After using the GPC module, the percentage of clutter points in the radar point clouds decreases by approximately 12.4% on average. Moreover, the Hausdorff and Chamfer distances decrease by 57.9% and 5.1%, respectively. Fig. 4 shows a comparison of point clouds generated with and without the GPC module.

5. Conclusion

In this article, we present a novel method that fuses camera and mmWave radar to generate high-quality point clouds for autonomous vehicle perception tasks. We validate the proposed method by constructing a real-world dataset that includes urban streets, elevated highways, and tunnels. Additionally, we introduce a new evaluation metric specifically designed to evaluate the quality of millimeter-wave radar point clouds in autonomous driving perception tasks. Experimental results demonstrate that our approach outperforms existing methods, which contains fewer clutter points and denser real points. The radar point clouds generated by the proposed method can be applied to object detection, localization, and mapping tasks, which will contribute to the advancement of autonomous driving perception.

References

[1] Yasin Almalioglu, Mehmet Turan, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. Milli-rio: Ego-motion estimation with low-cost millimetre-wave radar. *IEEE Sensors Journal*, 21(3):3314–3323, 2020. 1

[2] Kshitiz Bansal, Keshav Rungta, Siyuan Zhu, and Dinesh Bharadia. Pointillism: Accurate 3d bounding box estimation with multi-radars. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 340–353, 2020. 1, 2

[3] Daniel Brodeski, Igal Bilik, and Raja Giryes. Deep radar detector. In *2019 IEEE Radar Conference (RadarConf)*, pages 1–6. IEEE, 2019. 1

[4] Mahdi Chamseddine, Jason Rambach, Didier Stricker, and Oliver Wasenmuller. Ghost target detection in 3d radar data using point cloud based deep neural network. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 10398–10403. IEEE, 2021. 1, 2

[5] Yuwei Cheng, Hu Xu, and Yimin Liu. Robust small object detection on the water surface through fusion of camera and millimeter wave radar. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15263–15272, 2021. 1

[6] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. In *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*, pages 236–243. Springer, 2003. 3

[7] Christopher Doer and Gert F Trommer. An ekf based approach to radar inertial odometry. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 152–159. IEEE, 2020. 1

[8] Maximilian Gall, Markus Gardill, Thomas Horn, and Jonas Fuchs. Spectrum-based single-snapshot super-resolution direction-of-arrival estimation using deep learning. In *2020 German Microwave Conference (GeMiC)*, pages 184–187. IEEE, 2020. 2

[9] Ziyang Hong, Yvan Petillot, and Sen Wang. Radarslam: Radar based large-scale slam in all weathers. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5164–5170. IEEE, 2020. 1

[10] Andrew Kramer, Carl Stahoviak, Angel Santamaria-Navarro, Ali-Akbar Agha-Mohammadi, and Christoffer Heckman. Radar-inertial ego-velocity estimation for visually degraded environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5739–5746. IEEE, 2020. 1

[11] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018. 3

[12] Vladimir Lekic and Zdenka Babic. Automotive radar and camera fusion using generative adversarial networks. *Computer Vision and Image Understanding*, 184:1–8, 2019. 2

[13] Gang Li, Yoke Leen Sit, Sarath Manchala, Tobias Kettner, Alicja Ossowska, Kevin Krupinski, Christian Sturm, and Urs Lubbert. Novel 4d 79 ghz radar concept for object detection and active safety applications. In *2019 12th German Microwave Conference (GeMiC)*, pages 87–90. IEEE, 2019. 1

[14] Teck-Yian Lim, Amin Ansari, Bence Major, Daniel Fontijne, Michael Hamilton, Radhika Gowaikar, and Sundar Subramanian. Radar and camera early fusion for vehicle detection in advanced driver assistance systems. In *Machine learning for autonomous driving workshop at the 33rd conference on neural information processing systems*, volume 2, page 7, 2019. 1

[15] Juan-Ting Lin, Dengxin Dai, and Luc Van Gool. Depth estimation from monocular images and sparse radar data. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10233–10240. IEEE. 1

[16] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. *Advances in Neural Information Processing Systems*, 30, 2017. 4

[17] Chris Xiaoxuan Lu, Muhamad Risqi U Saputra, Peijun Zhao, Yasin Almalioglu, Pedro PB De Gusmao, Changhao Chen, Ke Sun, Niki Trigoni, and Andrew Markham. milliego: single-chip mmwave radar aided egomotion estimation via deep sensor fusion. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 109–122, 2020. 1, 2

[18] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4796–4803. IEEE, 2018. 6

[19] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6851–6860, 2019. 2, 3

[20] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2021. 2

[21] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[22] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 3, 6, 7

[23] Robert Prophet, Javier Martinez, Juan-Carlos Fuentes Michel, Randolf Ebelt, Ingo Weber, and Martin Vossiek. Instantaneous ghost detection identification in automotive scenarios. In *2019 IEEE Radar Conference (RadarConf)*, pages 1–6. IEEE, 2019. 2

[24] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018. 3

[25] Kun Qian, Zhaoyuan He, and Xinyu Zhang. 3d point cloud generation with millimeter-wave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–23, 2020. 2

[26] In-hwan Ryu, Insu Won, and Jangwoo Kwon. Detecting ghost targets using multilayer perceptron in multiple-target tracking. *Symmetry*, 10(1):16, 2018. 2

[27] Shunqiao Sun, Athina P Petropulu, and H Vincent Poor. Mimo radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges. *IEEE Signal Processing Magazine*, 37(4):98–117, 2020. 1, 2

[28] Wouter Van Gansbeke, Davy Neven, Bert De Brabandere, and Luc Van Gool. Sparse and noisy lidar completion with rgb guidance and uncertainty. In *2019 16th international conference on machine vision applications (MVA)*, pages 1–6. IEEE, 2019. 6

[29] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 2

[30] Kilian Weinberger, Benjamin Packer, and Lawrence Saul. Nonlinear dimensionality reduction by semidefinite programming and kernel matrix factorization. In *International Workshop on Artificial Intelligence and Statistics*, pages 381–388. PMLR, 2005. 5

[31] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 3

[32] Taohua Zhou, Mengmeng Yang, Kun Jiang, Henry Wong, and Diange Yang. Mmw radar-based technologies in autonomous driving: A review. *Sensors*, 20(24):7283, 2020. 6, 7