mmWave Radar and Image Fusion for Depth Completion: a Two-Stage Fusion Network

Tieshuai Song
School of Electronic and Information Engineering
Beihang University
Beijing, China
4399song@buaa.edu.cn

Bin Yang
School of Electronic and Information Engineering
Beihang University
Beijing, China
young_being@buaa.edu.cn

Jun Wang
Hangzhou Innovation Institute of Beihang University
Zhejiang, China
wangj203@buaa.edu.cn

Guidong He
School of Electronic and Information Engineering
Beihang University
Beijing, China
GuidongHe@buaa.edu.cn

Zhao Dong
School of Electronic and Information Engineering
Beihang University
Beijing, China
Dongzhao@buaa.edu.cn

Fengjun Zhong
School of Electronic and Information Engineering
Beihang University
Beijing, China
SY2202115@buaa.edu.cn

***Abstract*—Pixel-wise depth completion using multi-sensor fusion is crucial in areas such as autonomous driving. While LiDAR and image fusion methods exhibit reliability, it can face challenges in adverse weather conditions, such as rain and fog. In contrast, mmWave radar, emerged in recent years, has stronger anti-interference capability. However, radar point typically features high sparsity. And mmWave radar has lower resolution in the height dimension, leading to increased errors when projected onto the image plane. To solve the problem, this paper proposes a two-stage fusion convolutional neural network. In the first stage, image features are utilized to filter the noisy radar point cloud and learn the mapping of radar points to image regions. In the second stage, we perform multi-scale fusion of the image with the coarse depth map generated in the first stage to predict the missing depth values. Experiment results indicate that our improved strategy reduces the error of depth value estimation. Our network shows a 4.5% improvement in RMSE(root-mean-square error) compared to the previous method.**

***Index Terms*—fusion, depth completion, convolutional neural network**

I. INTRODUCTION

Pixel-wise depth completion has received widespread attention due to its crucial role in fields such as autonomous driving, 3D reconstruction, and navigation [1]. Its purpose is to estimate the depth value of a real spatial target at the corresponding position of image. Through precise depth estimation, the intelligent platform can accurately perceive the surrounding environment in real time, including the locations of obstacles, pedestrians and other targets [2]. Accurate depth estimation plays an irreplaceable role in improving the safety and reliability of automatic driving systems.

Currently, automated driving vehicles primarily use three kinds of sensors to perceive the environment, including cameras, mmWave radar, and LiDAR [3]. Cameras can capture rich color and texture information of the external environment; however, the image itself does not contain distance information [4]. LiDAR can capture accurate distance information; however, it is larger in size, consumes more energy, and more susceptible to adverse weather conditions. In contrast, mmWave radar is several orders of magnitude cheaper, and has better anti-interference capability [5]. In the past few years, mmWave radar has already played an important role in vehicle collision detection [6], parking assistance and other aspects [7]. However, compared with LiDAR, mmWave radar point cloud is more sparse, and struggle to detect some soft or absorbent surfaces [8]. This limitation leads to insufficient information in certain scenarios. To address these issues, many deep-learning-based fusion methods, with the development of artificial intelligence, have been proposed in recent years [9].

In terms of depth completion, some progress has been made in the fusion of lidar point clouds and images using neural networks [10], [11], [12], [13], [14]. Unlike LiDAR, most of the current mmWave radars in vehicular scenarios lack height dimension antenna arrays or have low height dimension resolution [15], and their fusion with the camera images faces two challenges. First, strong reflected signals at different heights may enter the same distance gate of the mmWave radar. As shown in Fig. 1(a), there are three targets A, B, and C, at different heights with a radar distance of R, and their corresponding true depth values should be x_1, x_2, and x_3. And for mmWave radar with low vertical resolution, when receiving reflections from three targets at the same slant range, regardless of their orientation relative to the radar's axis or main beam, they are perceived as being at the same height due to depth and height coupling. Second, when projecting the point cloud of the mmWave radar to the forward-looking image plane, image occlusion and point cloud overlap can significantly affect the accuracy of depth estimation. As shown in Fig. 1(b)(c)(d), in the front-view camera perspective, car 2 is blocked by car 3. Millimeter wave radar has certain diffraction capabilities. And car 2 can be detected behind car 3 by bypassing it. When projecting the radar point cloud to the front view, the radar points corresponding to car 2 and car 3 will fall within the image pixel area corresponding to car 3. thus incorrectly affecting the estimation of the true depth value of the target. Therefore, when fusing mmWave radar point cloud information, it is necessary to filter the erroneous radar point cloud.

Our model adopts a two-stage architecture. In the first stage, we filter and enhance the point cloud, then associate it with the corresponding regions in the image. Subsequently, in the second stage, we complete the depth in the missing regions. To address the aforementioned high blurriness issue, we select points with depth differences within a certain range as the associated regions for supervising the network training, aiming to predict the associated regions for each radar point.

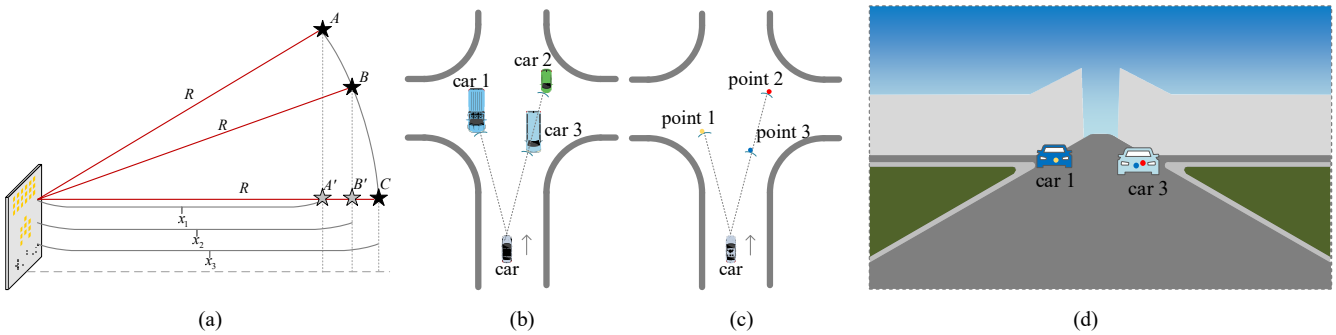

Fig. 1. Challenges in depth completion for millimeter-wave radar and image fusion, with height-distance blurring demonstrated in (a), scenarios with occlusion problems in top view shown in (b)(c), and a point cloud of an occluded target falling in the wrong region during projection shown in (d).

Regarding occlusion, the radar enhancement network acts as a filter, associating erroneous point clouds with extremely small regions or directly filtering them out. The contributions of this work are as follows:

- Proposed a Radar Enhancement Network to address the association problem between radar point clouds and image regions by fusing image information into the radar encoding process.
- Proposed a new Multi-Dimensional fusion module for the decoder part of the Radar Vision Fusion Network and used a parameterized output function to further enhance the network's performance.
- Conducted extensive experiments on the nuScenes[16] dataset, with depth completion exceeding the current best network results on this dataset.

II. Related Work

There have been some deep learning based mmWave radar point cloud and image fusion methods that have achieved better results in depth completion. Some direct fusion methods have simple network models; however, their fusion effects are relatively limited. Ma et al. [17] use an Encoder-Decoder network built on the ResNet-50 [18] backbone to directly connect images and sparse depth maps together to complete depth completion from the sparse depth maps and corresponding RGB images; Qu et al. [19] replaced the last convolutional layer with a least squares fitting module. In this model, features from the penultimate layer act as bases, with weights determined by least squares on valid pixel depths.

While other methods use two-stage networks to predict from coarse to fine. Hambarde et al. [20] proposed S2DNet, which consists of two networks: the S2DCNet and the S2DFNet, which perform the first coarser prediction and the second refinement, respectively; Long et al. [21] proposed the RCPDA model, which associates radar reflection points with nearby image pixels of the same depth to generate an enhanced and dense radar image, called "MER". Lo et al. [22] designed a radar-camera fusion depth completion model: RCDPT, based on the ViT [23] backbone, which is an improvement of the DPT [24] network in monocular depth estimation. Li et al. [25] employed a four-stage model to refine depth predictions, achieving a good depth completion effect; however, their network model is overly complex. Singh et al. [26] proposed a gated fusion mechanism for the fusion of mmWave radar point cloud with image features using a two-stage fusion network, RadarNet and FusionNet. RadarNet model notices the positional ambiguity brought about by point cloud projection, independently queries each radar point and corresponds it to the correct position in the image, and the FusionNet selectively combines radar and camera features to generate dense depth maps by considering the confidence of the correspondences. However, RadarNet does not consider the filtering role of image features in encoding radar information. This leads to an abundance of erroneous radar features in the fusion decoder. Additionally, the FusionNet decoder structure is not conducive to extracting edge information, and its feature fusion method is simplistic. And the output mapping function is not conducive for the model to learn the distribution of the depth values quickly, which results in a long network training cycle. Two-stage networks, starting coarse and refining, typically outperform direct depth estimation. Thus, we adopted a two-stage model design, but solved the above problems with some improvements. By introducing image features to the radar encoder, we effectively reduce the model parameters. In the fusion network part, we improve the encoder and the output function to reduce the training cycle and improve the accuracy of depth completion.

III. Proposed Method

A. Overview architecture

Our depth completion network adopts the encoder-decoder structure that has been successful in making dense predictions in prior studies [27], [28]. The network processing is divided into two parts, the first part is Radar Enhancement Network: it is responsible for the initial filtering of the low resolution radar point cloud, correlating the point cloud to the image region, and extending the information that the radar point cloud can provide. The second part is the Radar Vision Fusion Network: it is responsible for fusing the image and the coarse depth map output from the Radar Enhancement Network. From previous studies, most of the mmWave radar and image fusion methods end up with the fusion of images with coarse depth maps in image-like form. For example, [29] are expanding the radar projection points in the height dimension, and [21] learns the association of radar points with image regions through an association network, although in RCPDA the network uses information from future frames, which is undesirable in real-time processing. We used single-frame radar point clouds paired with corresponding images for depth completion. The inputs of our network are the original optical image and the mmWave radar point cloud, a coarse

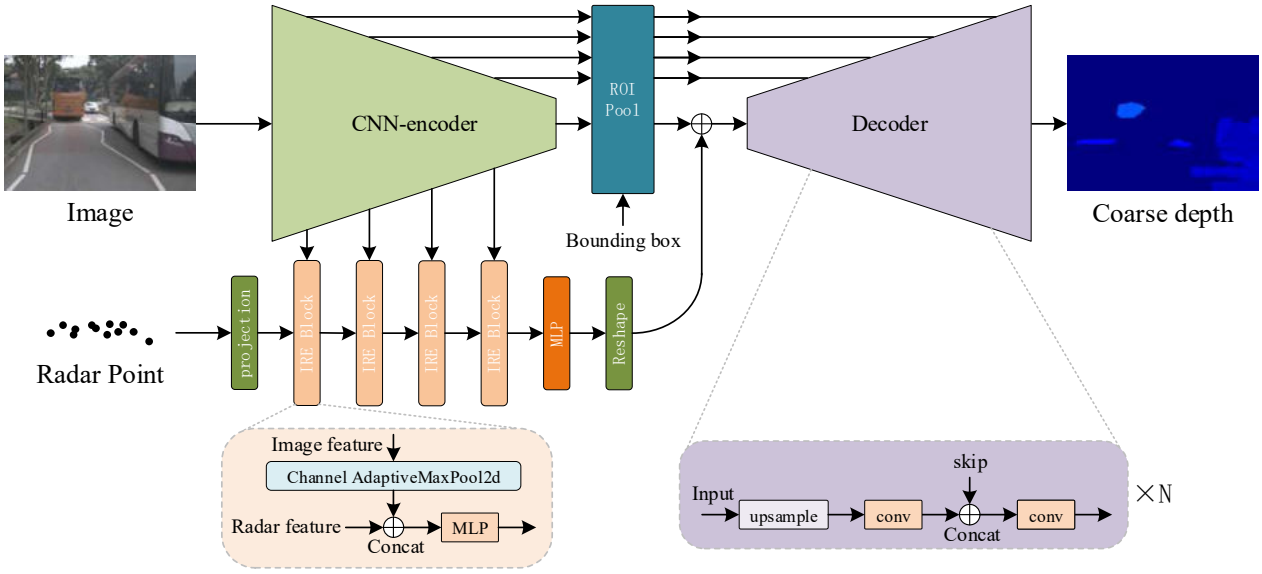

Fig. 2. Radar Enhancement Network

depth map is generated by the Radar Enhancement Network. The coarse depth map and the image are passed through a Radar Vision Fusion Network to accomplish depth completion.

B. Radar Enhancement Network

The Radar Enhancement Network architecture is shown in Fig. 2. Assume that the inputs to the Radar Enhancement Network are an RGB image $I \in \mathrm{R}^{H\times W\times 3}$ and a collection of radar point cloud $P \in \mathrm{R}^{K\times 3}$, where H and W are the height and width of the image, and K is the number of point clouds scanned by the radar in a single frame. For each radar point, we focus only on the image features in the region $h\times w$ near its projection location, called the radar association candidate region. The network eventually associates each radar point to a portion of pixels in its candidate region. We note that the lidar point clouds in the nuScenes dataset are mostly concentrated in the middle and lower regions of the image when projected onto the image plane, and the top of the image lacks valid depth value information. In order to minimize the impact of the missing depth information at the top of the image on the network, we selected the radar point cloud candidate correlation region to be $h\times w$ size from the bottom of the image, and did a certain range of cropping of the top region of the image. After the features are extracted from the image by the CNN encoder, the candidate regions for each layer of features are extracted based on the bounding box of each radar point and processed in the decoder. [26] directly projects the radar point cloud with multiple linear layers during the encoding process and outputs the feature maps with the same size of the image features, however, this approach leads to a large number of parameters. In this regard, we propose Image-Radar Enhancement Block (IRE Block) for high performance extraction of radar point cloud features, which is formulated as:

$$\mathrm{IRE}(t_I, t_R) = \mathrm{mlp}(\mathrm{cat}(\mathrm{GAP}(t_I),\ t_R)) \tag{1}$$

where t_I and t_R are the image and radar features at different scales during the encoding process, GAP(Global Average Pooling) converts a channel of feature maps into a single feature element by calculating the average value of each channel; cat is concatenation operation, where we concatenate along the channel dimension; mlp(Multilayer Perceptron) is composed of multiple linear layers. The IRE module introduces the image information into the radar encoder right during the encoding process, initially selecting more effective radar features. And it is verified by our experiments that the effect of RadarNet can be exceeded by connecting a single channel-sized radar feature map to the image features only in the last layer of the radar encoder, which drastically reduces the number of parameters of the network.

As shown in Fig. 2, in the Radar Enhancement Network, the last layer of the encoder outputs image feature maps and radar feature maps spliced by channel, according to the bounding box to find each radar point corresponding to the region of interest (ROI) feature $r \in \mathrm{R}^{h\times w\times C}$. For K radar points, the decoder inputs the feature map $F \in \mathrm{R}^{K\times h\times w\times C}$, and finally predicts a correlation map $F' \in \mathrm{R}^{K\times h\times w\times 1}$ for each radar point cloud, with the element value representing the probability that the location in the image belongs to the current radar point, so that for each pixel of the image the corresponding radar depth value can be selected according to the probability. The ground truth crop $Gt_{crop,n} \in \mathrm{R}^{h\times w\times 1}$ corresponding to the candidate correlation region for each radar point p_n is divided into correlation and non-correlation points based on the difference Δd_n with the radar point cloud depth value, and the pixel is considered as belonging to the radar point p_n when the difference in depth is less than a threshold value λ, i.e., the correlation map can be expressed as:

$$\Re(i,j) = \begin{cases} 1, & \text{if } \Delta d_n(i,j) < \lambda \\ 0, & \text{else} \end{cases} \tag{2}$$

For multiple radar points, a regional correlation map $\Re_n \in \square^{h\times w\times 1}$ is used to supervise the output values of the

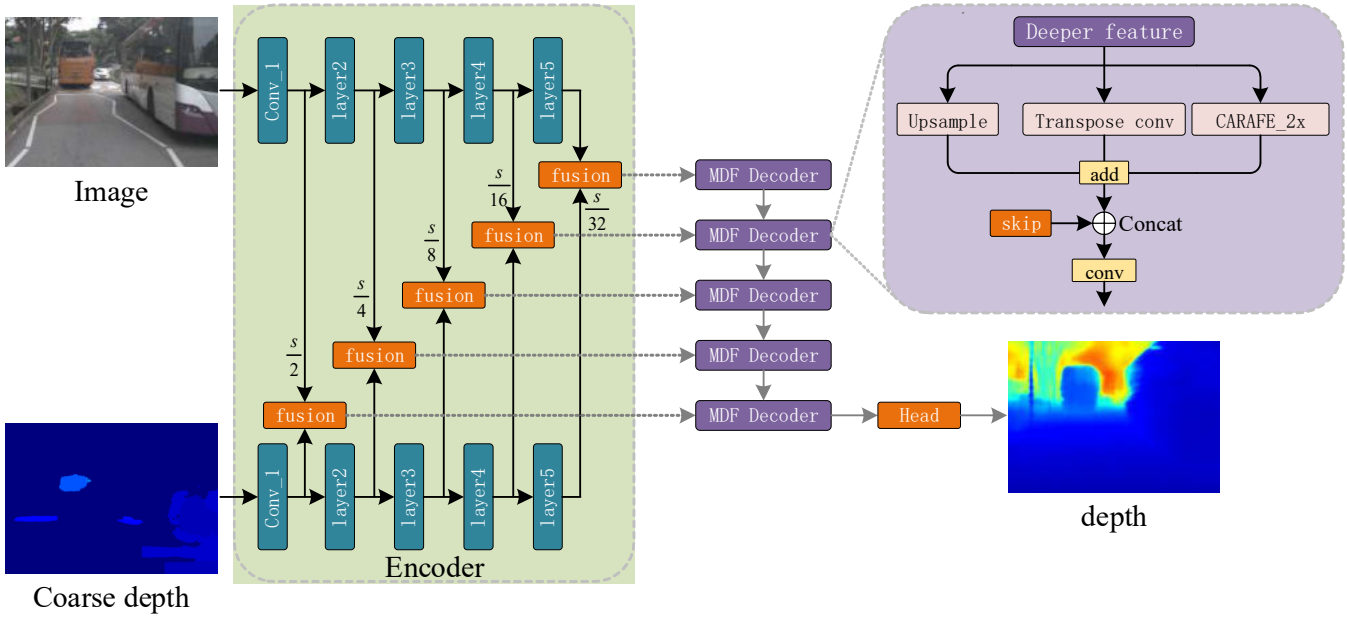

Fig. 3. Radar Vision Fusion Network

Radar Enhancement Network, and the parameters of the Radar Enhancement Network are optimized using cross-entropy loss. Finally, the coarse depth map is obtained by assigning the depth value of the pixel to the depth value of the corresponding radar point based on the predicted correlation probability.

C. Radar Vision Fusion Network

In the second stage, we use the coarse depth map generated by the Radar Enhancement Network to fuse with the optical image, and the encoder stage is processed using gated fusion approach [28], which fuses from five scales. The overall architecture of the fusion network is shown in Fig. 3. Assuming that the intermediate features of the encoder part of the coarse depth map are $skip_{R,i}$, where $i \in 1 \sim 5$, respectively, represent feature maps with resolution $s/2 \sim s/32$, which are fused with the image features $skip_{I,i}$, the fusion result at each layer is:

$$fusion_i = skip_{I,i} + (\text{conv}(skip_{R,i})) \otimes (\text{project}(skip_{R,i})) \quad (3)$$

Where $\otimes$ is the Hadamard product and the project function can be a convolution.

In the decoder stage, direct using bilinear interpolation causes distortion of feature details and fails to restore information about the edges, whereas using transposed convolution better preserves the details of the feature maps. In addition the CARAFE module can obtain a large sensory field during feature reorganization and use the input feature map to predict the up-sampling kernel. Therefore, we propose the Multi-dimensional fusion (MDF) module. The MDF module processes the input features through the three sub-modules of up-sampling convolution, transposition convolution, and CARAFE block, respectively, and then superimposes the outputs, which are spliced with the fusion results at different scales into the next layer of the decoder. The MDF module can be represented as:

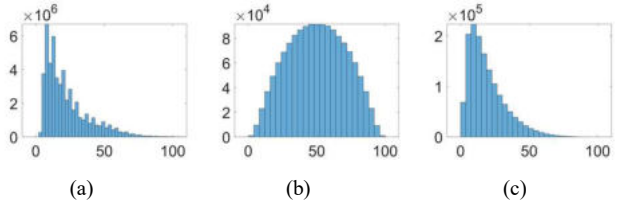

Fig. 4. Statistical distribution of the depth values

$$\begin{aligned} \text{MDF}_i &= \text{conv}(\text{cat}(x_1 + x_2 + x_3, skip_i)) \\ x_1 &= \text{Upsample}(x) \\ x_2 &= \text{TransposeConv}(x) \\ x_3 &= \text{CARAFE}(x) \end{aligned} \quad (4)$$

In the output layer section, in contrast to the direct output of previous models or the use of a fixed activation function mapping, we chose to use a projection function with learnable parameters:

$$depth = \text{Head}(x) = \frac{d_{\min}}{1/(1+e^{-k(x+b)}) + d_{\min}/d_{\max}} \quad (5)$$

where $d_{\min}$ is the minimum distance predicted, $d_{\max}$ is the maximum distance predicted, and k, b is a learnable model parameter. We use this projection function to replace original output function, and after 90 epochs of training, we find that the RMSE error is reduced compared to Singh's model, and the final k and b are stabilized around 0.9 and -1.8. The output function at this point is able to map to valid depth values in both the positive and negative regions of the x-axis, whereas Singh's model only maps to valid depth values when x values are negative.

We counted the distribution of depth values of the ground truth, as shown in Fig. 4(a), and the depth values are approximated to the Rayleigh distribution. The initialized

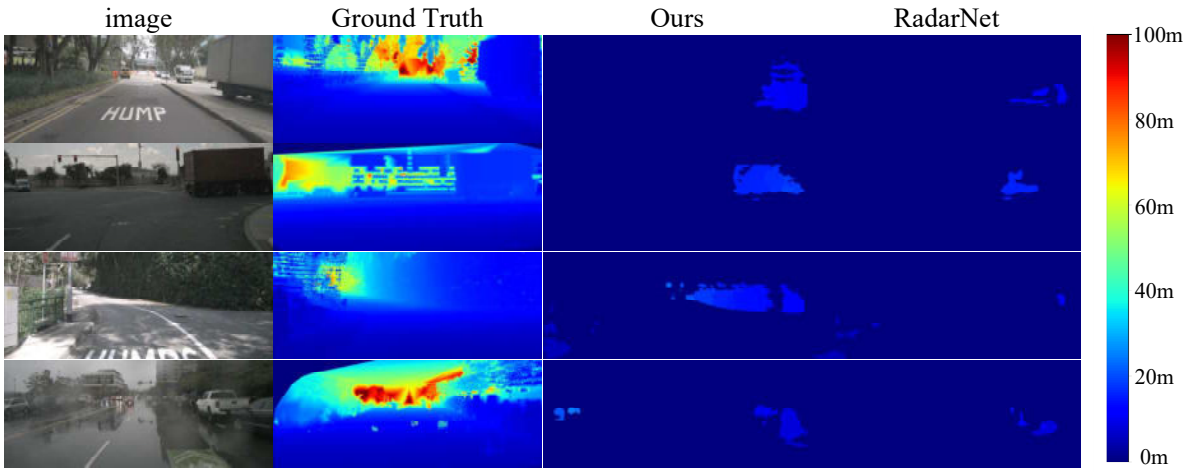

Fig. 5. Comparison of results on nuScenes. The first column is the input image, the second is Ground Truth, the third is the effect of our radar-enhanced network, and the fourth is the effect of Singh's model. Our model achieves an RMSE of 1.41, compared to RadarNet's 1.53.

model output approximates a Gaussian distribution. After mapping the output function of Singh[26] and ours, the resulting distribution of the output depth values is shown in Fig. 4(b)(c). It can be seen, with our proposed output mapping function, the distribution of depth values is closer to the statistical distribution of the ground truth. This makes our model more suitable for network training.

And the loss function set to:

$$Loss = w_0 \cdot \mathrm{SmoothL1}(\Delta d) + w_1 \cdot \mathrm{GradLoss}(I, \Delta d) \quad (6)$$

where $\Delta d = d_{pred} - d_{gt}$, The SmoothL1 loss function can be expressed as:

$$\mathrm{SmoothL1}(x) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} \beta x^2, & |x| < 1 \\ |x| - \beta, & x < -1 \text{ or } x > 1 \end{cases} \quad (7)$$

The GradLoss function weights Δd by acquiring the image lateral and vertical gradients, so that the output depth map gradient variation can be close to the optical image to avoid excessive edge errors.

IV. Experiment Results

A. Dataset

To evaluate the performance of our two-stage fusion network, we chose to conduct our experiments on the nuScenes dataset. NuScenes dataset contains data from LiDAR, mmWave radar, surround-view camera, and inertial measurement unit (IMU), which consists of about 40,000 labeled aligned samples. The approximate ratio of the training set, validation set, and test set is 7:1.5:1.5. In our experiments, we first project both the forward-looking radar point cloud and the lidar point cloud to the forward-looking camera plane based on the coordinate transformation matrix of each sample. In order to fit the ground truth at a particular sample moment, we choose to use the lidar point cloud data from 40 frames before and after that frame for projection accumulation. Data from the top 260 heights of the image were excluded from the experiments due to the absence of lidar point cloud information in these regions.

B. Implementation details

We implemented our network using Pytorch and trained it on an NVIDIA RTX A5000. We pay attention to the details of the experiments in [26]. For the radar-enhanced network part, 6 radar points were randomly selected for training for each sample, and the size of the candidate association region for each radar point was 512 × 288. Starting learning rate was 1e-3, and after a duration of 30 epochs, the training was completed after 90 epochs using a 1e-4 continuum. Batch size was set to 12. We use the Adam optimizer with momentum and weight decay using 0.9 and 0.999 correspondingly. We further use data enhancement on the images to improve the robustness, including: image contrast, saturation, brightness range adjustment, random horizontal and vertical flipping, adding Gaussian distributed random noise, etc. The encoder part of the fusion network is initialized using [26] pre-training parameters. Batch size is set to 2, and the learning rate is sequentially set to 1e-3, 1e-4, and 1e-5 lasting for 40, 20, and 20 epochs, respectively. Data augmentation is done in the same way as Radar Enhancement Network.

C. Results

For the depth completion task of radar image fusion, we have selected several commonly used metrics:

Root Mean Square Error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\| D_{pred}^{i} - D_{gt}^{i} \right\|^2} \quad (8)$$

Absolute Relative Error(AbsRel):

$$AbsRel = \frac{1}{N}\sum_{i=1}^{N}\left| \frac{D_{pred}^{i} - D_{gt}^{i}}{D_{gt}^{i}} \right| \quad (9)$$

The threshold accuracy δ_n :

$$\delta_n = \frac{1}{N}\sum_{i}\left(\max\left(\frac{D_{gt}^{i}}{D_{pred}^{i}}, \frac{D_{pred}^{i}}{D_{gt}^{i}} \right) < (1.25)^n \right) \quad (10)$$

We compare the results of our model with those of previous models. In the first stage of the Radar Enhancement Network, [26] generates a rough manipulated depth map

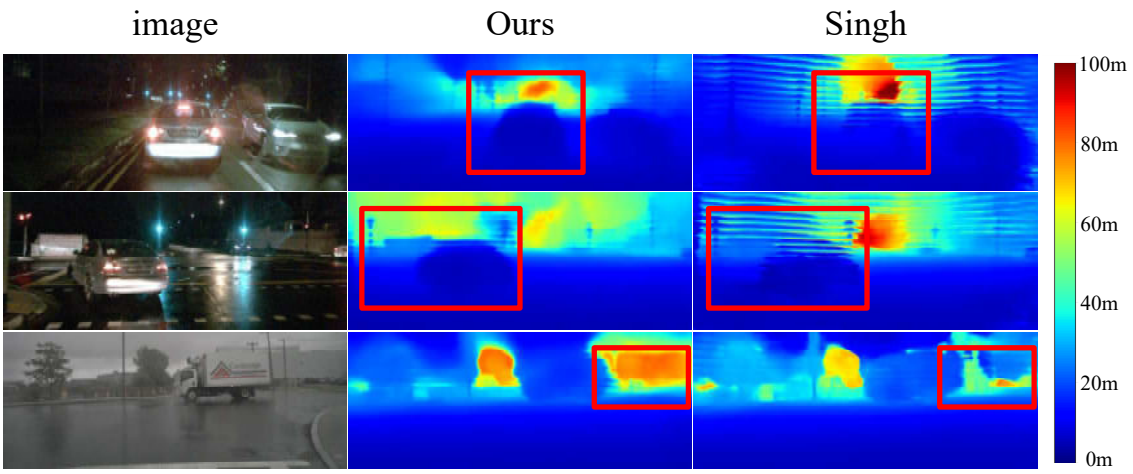

Fig. 6. Comparison of depth completion. First column is the image, the second column is the depth map of ours, Singh's is the third column . In the first two examples, singh's method produces many horizontal stripes, while our model is smoother in this regard. In the third example, singh's model has a large detection error in the background region, and our model is more accurate.

TABLE I. QUANTITATIVE COMPARISON OF MODELS

Method	RMSE↓	AbsRel↓	δ_1 ↑
RCPDA[21]	7.692	-	-
RCDPT[22]	5.165	0.095	0.901
S2D[17]	5.628	0.115	0.876
Lin[30]	5.180	0.100	0.901
Lee[31]	5.209	0.104	0.895
Singh[26]	4.899	0.102	0.897
Ours	**4.679**	**0.082**	**0.922**

TABLE II. COMPARISON OF DIFFERENT COMBINATIONS

Combination Mode	RMSE↓
Only Upsample	4.812
Only Transpose Conv	4.721
Only CARAFE	4.940
Upsample + Transpose Conv	4.729
Upsample + CARAFE	4.795
Transpose Conv + CARAFE	4.711
Upsample + Transpose Conv + CARAFE	**4.679**

with an RMSE of 1.53, while our method outputs a coarse depth map with an approximate RMSE of 1.41, and the parameters of our radar enhancement model are only 1.8M, which is a reduction of about 78% in the size of the model parameters as compared to the size of the model parameters in [26], which is 8.2M. model parameters by about 78%. As shown in Fig. 5, from left to right, the optical image, Ground Truth, the rough depth image generated by our Radar Enhancement Network, and the coarse depth map generated by RadarNet are shown. It can be clearly seen that our generated coarse depth map outperforms RadarNet in terms of accuracy and information content. On the nuScenes dataset, the ratio of the average number of effective pixels per frame to the image size of our Radar Enhancement Network is about 0.0263, while the ratio is only 0.0179 in the RadarNet model, and our model is able to use fewer parameters to provide more and more accurate information about depth values, which will help in the second stage of depth completion.

In the second stage of the Radar Vision Fusion Network, we use the RMSE, AbsRel, and δ_n metrics commonly used in depth estimation as well as depth completion to measure the error of the model, and we compare our model with the other models in Table I. We have directly selected the results from their papers for this section because we were not able to reproduce the effects of many models in their papers. The results of the network performance comparison are shown in TABLE I. Our model outperforms all the remaining baseline models in the above three metrics, and we selected some of the depth maps to compare with [26], and the comparison results are shown in Fig. 6. It can be seen that [26] produces many horizontal stripes in the black scene, which may be caused by its long training time, while our model performs well in the overall effect, and in the third row, [26] presents a large error in the depth estimation value at the background, while our model performs well.

D. Ablation Study

To further validate the effectiveness of our method, we conducted multiple sets of experiments. We analyzed the role of the radar-enhanced network in the overall depth completion task, and for our network, the RMSE for depth completion using coarse depth maps was 4.679, while for depth completion without coarse depth maps was 5.011. Further we design experiments to verify the effectiveness of the improved part of our Radar Enhancement Network. We use the coarse depth maps generated by our Radar Enhancement Network to be trained using Singh's model in the second stage, and get the final training result with RMSE of 4.838, an improvement over Singh's model with RMSE of 4.899.

In the second stage, we enhance network performance using three submodule operators inside the decoder MDF module. We experimented with different combinations in the MDF module: "Upsample" denotes bilinear interpolation, "Transpose Conv" denotes transposed convolution, and "CARAFE" uses a CARAFE module. Experimental results are shown in TABLE II. It can be seen that the network works best when the three operators in the MDF module are employed simultaneously. We replace the final output mapping function with our proposed function using [26] and achieve an RMSE of 5.01 after 60 training epochs. While FusionNet[26], with 200 epochs, has an RMSE of 5.2 only.

V. Conclusion

In this paper, we proposed a two-stage fusion network for radar vision fusion. Specifically, we introduced the IRE module to integrate image features into radar encoding, enhancing the network's ability to extract valid depth information from radar point cloud. In the Radar Vision Fusion Network, we innovatively employed the MDF module and refined the depth completion output mapping function, leading to improved network training efficiency and error reduction. Our improved method made the network easier to train and effectively reduced the RMSE. Experiment results demonstrated the effectiveness of our approach. Our future work includes extending the current method to additional high-quality indoor and outdoor datasets and analyzing its performance under various mmWave radar signal-to-noise ratios.

References

[1] Junjie Hu, Chenyu Bao, Mete Ozay, Chenyou Fan, Qing Gao, Honghai Liu, and Tin Lun Lam, "Deep depth completion from extremely sparse data: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8244-8264, 2022.

[2] Felipe Manfio Barbosa, Fernando Santos Osório, "Camera-radar perception for autonomous vehicles and ADAS: Concepts, datasets and metrics," 2023, *arXiv:2303.04302*.

[3] Zhangjing Wang, Yu Wu, and Qingqing Niu, "Multi-sensor fusion in automated driving: A survey," *IEEE Access*, vol. 8, pp. 2847-2868, 2020.

[4] Tom van Dijk, and Guido de Croon, "How do neural networks see depth in single images?" in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 2183-2191, 2019.

[5] Yong Niu, Yong Li, Depeng Jin, Li Su, and Athanasios V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: opportunities and challenges," *Wireless networks*, vol. 21, pp. 2657-2676, 2015.

[6] Junil Choi, Vutha Va, Nuria Gonzalez-Prelcic, Robert Daniels, Chandra R. Bhat, and Robert W. Heath, "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Communications Magazine*, vol. 54, no. 12, pp. 160-167, Dec. 2016.

[7] Zhiqing Wei, Fengkai Zhang, Shuo Chang, Yangyang Liu, Huici Wu, and Zhiyong Feng, "Mmwave radar and vision fusion for object detection in autonomous driving: A review," *Sensors*, vol. 22, no. 7, p. 2542, Mar. 2022.

[8] 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*, vol. 4, no. 4, pp. 1-23, 2020.

[9] Yaodong Cui, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao, "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 722-739, 2021.

[10] Zhizhuo Zhou, and Shubham Tulsiani, "Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12588-12597, 2023.

[11] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu , Guan Huang , and Stefano Mattoccia , "Completionformer: Depth completion with convolutions and vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18527-18536, 2023.

[12] Hongxiang Cai, Zeyuan Zhang, Zhenyu Zhou, Ziyin Li, and Wenbo Ding, "BEVFusion4D: Learning LiDAR-Camera Fusion Under Bird's-Eye-View via Cross-Modality Guidance and Temporal Aggregation," *arXiv:2303.17099*, 2023.

[13] 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), IEEE*, pp. 1-6, 2019.

[14] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1090-1099, 2022.

[15] Arthur Venon, Yohan Dupuis, Pascal Vasseur, and Pierre Merriaux, "Millimeter wave fmcw radars for perception, recognition and localization in automotive applications: A survey," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no 3, pp. 533-555, 2022.

[16] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 11621-11631, 2020.

[17] 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)*. pp. 4796-4803, 2018.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 770-778, 2016.

[19] Chao Qu, Ty Nguyen, Camillo J. Taylor, "Depth completion via deep basis fitting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 71-80, 2020.

[20] Praful Hambarde and Subrahmanyam Murala, "S2DNet: Depth Estimation From Single Image and Sparse Samples," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 806-817, 2020.

[21] Yunfei Long, Daniel Morris, Xiaoming Liu, Marcos Castro, Punarjay Chakravarty, and Praveen Narayanan, "Radar-camera pixel depth association for depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12507-12516, 2021.

[22] Chen-Chou Lo, and Patrick Vandewalle, "RCDPT: Radar-camera fusion dense prediction transformer," In *Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5, 2023.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv:2010.11929*, 2020.

[24] Ren´ e Ranftl, Alexey Bochkovskiy, and Vladlen Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12179-12188, 2021.

[25] Han Li, Yukai Ma, Yaqing Gu, Kewei Hu, Yong Liu, and Xingxing Zuo, "RadarCam-Depth: Radar-Camera Fusion for Depth Estimation with Learned Metric Scale," *arXiv:2401.04325, 2024*.

[26] Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, and Mani Srivastava, "Depth estimation from camera image and mmwave radar point cloud," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 9275-9285, 2023.

[27] Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Senior Member, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, and Yutao Yue, "Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review," *IEEE Transactions on Intelligent Vehicles*, 2023.

[28] Apoorv Singh, "Vision-radar fusion for robotics bev detections: A survey," in *Proceedings of the 2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1-7, Jun. 2023.

[29] Muhammad Ishfaq Hussain, Muhammad Aasim Rafique, and Moongu Jeon. "Rvmde: Radar validated monocular depth estimation for robotics," *arXiv:2109.05265*, 2021.

[30] Juan-Ting Lin, Dengxin Dai, and Luc Van Gool, "Depth estimation from monocular images and sparse radar data," in *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10233-10240, 2020.

[31] Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips, "Semantic-guided radar-vision fusion for depth estimation and object detection," in *Proceedings of the 32th British Machine Vision Conference (BMVC)*, 2021.