Abstract—LiDAR depth-only completion is a challenging task to estimate a dense depth map only from sparse measurement points obtained by LiDAR. Even though the depth-only completion methods have been widely developed, there is still a significant performance gap with the RGB-guided methods that utilize extra color images. We find that existing depth-only methods can obtain satisfactory results in the areas where the measurement points are almost accurate and evenly distributed (denoted as normal areas), while the performance is limited in the areas where the foreground and background points are overlapped due to occlusion (denoted as overlap areas) and the areas where there are no available measurement points around (denoted as blank areas), since the methods have no reliable input information in these areas. Building upon these observations, we propose an effective Coupled U-Net (CU-Net) architecture for depth-only completion. Instead of directly using a large network for regression, we employ the local U-Net to estimate accurate values in the normal areas and provide the global U-Net with reliable initial values in the overlap and blank areas. The depth maps predicted by the two coupled U-Nets complement each other and can be fused by learned confidence maps to obtain the final completion results. In addition, we propose a confidence-based outlier removal module, which identifies the regions with outliers and removes outliers using simple judgment conditions. The proposed method boosts the final dense depth maps with fewer parameters and achieves state-of-the-art results on the KITTI benchmark. Moreover, it owns a powerful generalization ability under various depth densities, varying lighting, and weather conditions. The code is released at https://github.com/YufeiWang777/CU-Net.

Index Terms—Computer vision for transportation, range sensing, depth completion

I. INTRODUCTION

Acquiring scene geometry using active depth sensors, e.g., LiDAR, has multiple applications in robotics [1] and SLAM [2]. However, due to the limited number of scan beams and angular resolution, existing LiDARs can only provide sparse measurements. Therefore, depth completion methods have been developed to predict dense depth maps from sparse depth measurements. According to whether the color images are utilized for guidance, existing depth completion methods are divided into the RGB-guided methods and the depth-only methods. Despite the RGB-guided methods outperforming the depth-only methods by a wide margin in the evaluation metrics, their performance heavily depends on the quality and style of color images, the unpredictable imaging conditions may cause adverse effects on such methods, especially when encountering rain, snow, and night [3]. In this paper, we focus on the depth-only methods.

Existing depth-only methods usually employ off-the-shelf network structures to directly regress the sparse depth maps to the dense depth maps, such as sparsity invariant CNN [4],...
encoder-decoder [5], [6], and hourglass [7], [8]. Although the performance of the depth-only methods has been dramatically improved, there is still a large gap with the RGB-guided methods. We observe that there are three different distributions of measurement points in the areas to be filled: 1) the measurement points are almost accurate and the distribution is even, 2) the measurement points are overlapped by foreground and background points due to occlusion, and some of the points whose depth values abruptly change should be removed as outlier points [9], [10], 3) there are no available points around. As shown in Fig. 1, according to the distributions, we divide the areas that need to be filled into normal areas, overlap areas, and blank areas. Through experiments, we observe that existing depth-only methods such as S2D (depth-only) [6] and Ip_basic [11] can obtain promising results in normal areas, the performance gap between them and state-of-the-art RGB-guided methods such as PENet [12] and S2D (RGB-guided) [6] primarily derives from the overlap and blank areas. We consider that this is because the depth-only methods have no reliable input information in these areas, while the RGB-guided methods can take advantage of the rich semantic information of the color images to get better results.

Therefore, predicting initial input information for the blank areas and overlap areas, and removing the outlier points from the overlap areas are the key to the depth-only methods. To address the issue, we propose the Coupled U-Net (CU-Net) method, which is enhanced by an outlier removal module. First, we adopt the first U-Net to predict an initial depth map and a corresponding confidence map from the sparse LiDAR measurements. The initial dense depth map has accurate depth values and high confidence in the normal areas, and it can also provide the second U-Net with reliable initial depth values in the overlap and blank areas. Second, we propose a confidence-based outlier removal method. Unlike removing the outliers by judging whether each measurement point meets complex conditions [10], the proposed method uses the confidence map predicted by the first U-Net to identify the regions with outliers and removes the outliers by a simple judgment condition. Then, the corrected sparse depth map and the initial dense map are fed to the second U-Net to obtain a dense depth map with improved results in the overlap and blank areas and the corresponding confidence map. Since the depth map predicted by the first U-Net shows satisfactory results in the ordered depth points and their local areas, while the depth map predicted by the second U-Net has good results in other global areas, we refer to these two U-Nets as the local U-Net and the global U-Net, respectively. The dense depth maps predicted by the local U-Net and the global U-Net are fused by the confidence maps to obtain the final completion result.

We conduct comprehensive experiments to verify the effectiveness and generalization of our method on the KITTI dataset [13], [14] and DDAD dataset [15]. Our contributions can be summarized as follows:

- To address the issue, we propose a two-stage network with learned intermediate confidence maps, where the first network provides initial depth values of the overlap and blank areas for the second network. Furthermore, we propose a confidence-based outlier removal method to enhance the proposed method, which employs a learned confidence map to identify the areas with outliers and remove them.

- Experimental results on the popular KITTI benchmark and the DDAD dataset show that our method achieves state-of-the-art performance among all published papers that employ only depth data during training and inference. Meanwhile, it shows powerful generalization capabilities under different depth densities, changing lighting, and weather conditions.

## II. RELATED WORK

This section introduces representative works of the RGB-guided methods and the depth-only methods.

**RGB-guided methods.** The input of the RGB-guided methods includes the sparse depth maps and their corresponding color images. How to fuse the information of these two different modalities is an open problem, a straightforward approach called “early-fusion” is to concatenate the depth maps and the color images to form a 4D tensor. Ma et al. [6] propose a “later-fusion” method that extracts the feature of the color images and the sparse depth maps separately, and feeds the fusion features into an Encoder-decoder network. Gansbeke et al. [8] propose a method based on color images guidance and uncertainty, which employs two branches and achieves better results. Qiu et al. [16] consider that the color images and depth maps are not strongly correlated, and propose a method that consists of the surface normal guided branch and the RGB guided branch. The proposed method first predicts the surface normal from the color images, and the results of two branches are fused through confidence images to obtain the final dense depth maps. Hu et al. [12] add the 3D information of the sparse depth maps to the convolution operation. Considering that the existing fusion methods between the color images and depth maps are too simple, Tang et al. [17] propose a guided convolutional network for feature fusion. To address the dense depth maps predicted by end-to-end networks that are blurred at the boundaries of objects, a series of affinity-based spatial propagation methods [18], [19], [20], [21], [22] have been proposed.

**Depth-only methods.** The depth-only methods adopt only the given sparse depth maps to predict the corresponding dense depth maps. Many classical methods can only efficiently fill relatively dense depth maps, such as bilateral filtering. To fill highly sparse depth maps, Ku et al. [11] propose a method that employs traditional image processing techniques, such as morphological transformations and image smoothing. The method has a fast processing speed and its performance exceeds that of contemporary learning methods. Zhao et al. [10] also propose a non-learning method based on surface geometry, which is enhanced by an outlier removal algorithm. In the deep learning era, considering the sparsity of the input depth maps, Uhrig et
al. [4] propose a sparsity invariant CNN. Cheng et al. [23] further point out that standard CNN can also handle sparse depth maps as long as the network is deep enough. Then, Ma et al. [5], [6] propose an Encoder-decoder model that uses an encoder network to extract features and a decoder network to recover the resolution. Lv et al. [24] propose a method based on multi-task learning that employs image reconstruction as an auxiliary task. The method is not a pure depth-only completion method, which needs color images in training.

### III. PERFORMANCE ANALYSIS OF THE DEPTH-ONLY METHODS FOR DEPTH COMPLETION

Compared with the RGB-guided methods, the depth-only methods have obvious limitations in evaluation metrics. In this section, we investigate the causes of the limitations by comparing the results of the two methods in different areas. We divide the areas to be filled into three categories according to the distributions of measurement points. To reduce the influence of other factors, we choose a typical depth-only method and RGB-guided method for comparison. These two methods both adopt a single U-Net, and the only difference is that the RGB-guided method also feeds the corresponding color images into the network.

#### A. Division of the Filled Area

In this subsection, we use $A$ to denote the filled area and the subtraction “−” to denote the difference between two areas. Depth completion needs to estimate a dense depth map $A_{all}$ with the same resolution as the sparse depth map. We first dilate all the measurement points in the sparse depth map to obtain the areas $A_{sparse}$. Subsequently, we perform the same operation to dilate the outlier points collected by the algorithm [10] to obtain the areas $A_{outlier}$. As follows, we divide the areas to be filled into three categories according to the distributions of measurement points.

- **Normal areas** $A_{normal}$: the measurement points in these areas are accurate and evenly distributed.
  \[ A_{normal} = A_{sparse} - A_{outlier} \]  

- **Overlap areas** $A_{overlap}$: the measurement points in these areas are a mixture of accurate points and outlier points.
  \[ A_{overlap} = A_{outlier} \]  

- **Blank areas** $A_{blank}$: there are no available measurement points in these areas.
  \[ A_{blank} = A_{all} - A_{normal} - A_{overlap} \]

#### B. RGB-guided vs Depth-only

In this subsection, we compare the performance of the typical depth-only method and RGB-guided method in various areas, which are proposed in [6] and denoted as S2D (depth-only) and S2D (RGB-guided), respectively. For quantitative evaluation, we employ root mean squared error of the inverse depth (iRMSE[1/km]), mean absolute error of the inverse depth (iMAE[1/km]), root mean squared error (RMSE[mm]), mean absolute error (MAE[mm]), relative absolute error (Rel), and $\delta$ inlier ratios (maximal mean relative error of $\delta_i$ for $i = 1, 2, 3$). Please refer to [21] for their specific definitions.

As shown in Tab. I, there is a significant performance gap between S2D (RGB-guided) and S2D (depth-only). We find that the difficulty of filling varies in different areas, and the primary performance gap between the depth-only methods and the RGB-guided methods is reflected in the overlap areas and blank areas. Specifically, S2D (depth-only) and S2D (RGB-guided) both achieve satisfactory and comparable performance in the normal areas, where the gap in RMSE is 21.3. However, in the overlap areas and blank areas, S2D (RGB-guided) obtains better results, where the gaps in RMSE are 626.0 and 132.0, respectively. We consider that in the normal areas, whether color images are used or not, the areas to be filled can be well estimated because the measurement points around are accurate and evenly distributed. However, since the blank areas have no valid measurement points and the overlap areas have only overlapped measurement points, it is challenging for the depth-only methods to achieve satisfactory results in these areas due to the lack of reliable input information. On the contrary, RGB-guided methods can exploit the semantic information of color images to improve their performance. In summary, the reason for the limited performance of the depth-only methods is that they have no reliable source of information in the overlap areas and blank areas.

### IV. METHODS

To alleviate the limitations of the depth-only methods, we propose the Coupled U-Net model that consists of a local U-
Global U-Net

GU-Conf

Sparse Depth

Res1

Res2

Res3

Res4

Res5

U-Net also estimates a confidence map \( \hat{C} \). Different from existing depth-only methods, the local U-Net further enhances the results in the overlap areas and blank map, and then employs another U-Net, the global U-Net, to obtain satisfactory results in the normal region. However, there is no effective outlier removal algorithm to remove incorrect points in the sparse depth map before feeding it into the global U-Net. CU-Net fuses the dense depth maps predicted by these two U-Nets through learned confidence images to obtain the final results.

A. Coupled U-Net Model

Existing depth-only methods treat depth completion as an image-to-image task. Given a sparse depth map \( D_{\text{sparse}} \), they usually use an end-to-end network such as U-Net to learn the mapping from the sparse depth map \( D_{\text{dense}} \) to the dense depth map \( D_{\text{dense}} \). As shown in Eq. (4), U-Net consists of an encoder network and a decoder network. The encoder network \( f_{\text{encoder}} \) extracts the feature representation \( X \) from the sparse depth map \( D_{\text{sparse}} \). Then, the decoder network \( f_{\text{decoder}} \) uses operations such as deconvolution or upsampling to predict a dense depth map \( \hat{D}_{\text{dense}} \) with the same resolution as \( D_{\text{dense}} \).

\[
\hat{D}_{\text{dense}} = f_{\text{U-Net}} (D_{\text{sparse}}) \implies \begin{cases} X = f_{\text{encoder}} (D_{\text{sparse}}) \\ \hat{D}_{\text{dense}} = f_{\text{decoder}} (X) \end{cases} \tag{4}
\]

As discussed in Sec. III, a single U-Net-based method can obtain satisfactory results in the normal region. However, there are undesirably large errors in the overlap and blank areas due to the lack of reliable input information. To solve this issue, we first use the local U-Net to predict an initial depth map, and then employ another U-Net, the global U-Net, to further enhance the results in the overlap areas and blank areas. Different from existing depth-only methods, the local U-Net also estimates a confidence map \( \hat{C}_{LU} \) that reflects the reliability of the initial depth map. The local U-Net is described in Eq. (5).

\[
(\hat{D}_{LU}, \hat{C}_{LU}) = f_{\text{LocalUNet}} (D_{\text{sparse}}) \tag{5}
\]

Although the dense depth map predicted by the local U-Net has unsatisfactory results in the overlap areas and blank areas, it is adequate to provide better initial values than the sparse depth maps. Consequently, we employ it to guide the global U-Net to generate better results in these areas. We feed the dense depth map \( \hat{D}_{LU} \) and the sparse depth map \( D_{\text{sparse}} \) into the global U-Net. The global U-Net also predicts a dense depth map \( D_{GU} \) and the corresponding confidence map \( \hat{C}_{GU} \). The global U-Net is described in Eq. (6).

\[
(D_{GU}, \hat{C}_{GU}) = f_{\text{GlobalUNet}} \left( \hat{D}_{LU}, D_{\text{sparse}} \right) \tag{6}
\]

B. Outlier Removal

In this section, we propose a confidence-based outlier removal method to remove incorrect points of the sparse depth maps. The algorithm [10] removes outlier points by judging whether each point in the sparse depth map meets the set of complex conditions. For example, the outlier points may change the relative position with some surrounding correct points, have relatively larger values, or create denser areas. However, due to the irregular distributions of points in the sparse depth map, it will also remove some correct points that meet the conditions, resulting in the loss of information. Unlike the algorithm [10] which works on the whole sparse depth map, we propose a confidence-based outlier removal method that focuses on the overlap areas of the sparse depth maps, where the outlier points are mainly clustered. Since the confidence maps predicted by the local U-Net have high values in the normal areas and low values in the overlap and blank areas, and there are no points in the blank areas, we can employ them to identify the overlap areas by setting a confidence threshold. Since the outlier points usually arise from the further background, the proposed method can achieve effective outlier removal by removing the points whose depth values are larger than the surrounding points, thus retaining more useful information.

The proposed outlier removal algorithm is shown in Alg. 1. \( N_d \) denotes the nonzero points set, which consists of the
Algorithm 1 The proposed outlier removal algorithm

Initialization: $D_r \leftarrow \emptyset$

for $d \in D_{\text{sparse}}$ and $d \neq 0$ do

$(u, v) \leftarrow \text{Index}(d)$

if $\hat{C}_{LU}(u, v) > \epsilon$ or $|d - 1/N_d| \sum_{d_i \in N_d} d_i < \epsilon$ then

$D_r = D_r \cup \{d\}$

return $D_r$

neighborhood points of a measurement point $d$, and $|N_d|$ is the cardinal number. A measurement point $d$ is deemed valid, only if two conditions are met. First, the corresponding confidence $\hat{C}_{LU}(u, v)$ exceeds a threshold $\epsilon$. Second, the difference between the measurement point and the average of its nonzero neighborhood is less than the threshold $\epsilon$. The valid measurement $d$ is added to $D_r$, which will be concatenated with $D_{LU}$.

C. Depth Fusion by Confidence

The dense depth maps predicted by the local U-Net $\hat{D}_{LU}$ have accurate values in the normal areas and coarse values in the overlap areas and blank areas, while the dense depth maps predicted by the global U-Net $\hat{D}_{GU}$ have the opposite. It indicates that these two dense depth maps have unique and complementary attributes. To take advantage of their respective strengths, we fuse them through the learned confidence maps $\hat{C}_{LU}$ and $\hat{C}_{GU}$. The fused depth $\tilde{D}$ is obtained by Eq. (7), $(u, v)$ denotes the coordinate of the pixel.

$$\tilde{D}(u, v) = \frac{\hat{C}_{LU}(u, v) \cdot \hat{D}_{LU}(u, v) + \hat{C}_{GU}(u, v) \cdot \hat{D}_{GU}(u, v)}{\hat{C}_{LU}(u, v) + \hat{C}_{GU}(u, v)}.$$ (7)

D. Loss Function

We use an L2 loss that is defined in Eq. (8).

$$L(\tilde{D}) = \| (\tilde{D} - D_{gt}) \odot 1_{\{D_{gt} > 0\}} \|^2.$$ (8)

where $\tilde{D}$ is the predicted dense map, $D_{gt}$ is the ground truth for supervision. Since the ground truth depth maps are semi-dense, which are created by aggregating 11 consecutive LiDAR scans into one, we only supervise the valuable parts, $1$ indicates whether there is a value in the ground truth, $\odot$ denotes the element-wise multiplication. In training, we supervise the dense depth map predicted by the local U-Net $\hat{D}_{LU}$, the dense depth map predicted by the global U-Net $\hat{D}_{GU}$, and the fused dense depth map $\tilde{D}$ simultaneously. The two confidence maps $\hat{C}_{LU}$ and $\hat{C}_{GU}$ are learned without direct supervision. The final loss is defined in Eq. (9), where $\lambda_{LU}$, $\lambda_{GU}$, and $\lambda_{Fused}$ are three hyper-parameters.

$$L = \lambda_{LU}L(\hat{D}_{LU}) + \lambda_{GU}L(\hat{D}_{GU}) + \lambda_{Fused}L(\hat{D}).$$ (9)

V. EXPERIMENTS

We perform extensive experiments to verify our method. We first introduce the training setting and the datasets in Sec. V-A. Then, we compare our method with the state-of-the-art depth-only methods on the KITTI dataset in Sec. V-B. In addition, we conduct comprehensive ablation experiments to verify the effectiveness of each component of the proposed method in Sec. V-C. Last but not least, in Sec. V-D, we perform experiments in various settings including different densities of the input depth, changing lighting, and weather conditions to prove the generalizability of our method.

A. Experiment Setups

We use PyTorch to implement our method, which is trained with a single Tesla V100 GPU. During training, we set the batch size to 6 and use a step learning rate decay policy with the Adam optimizer. The initial learning rate is set to $10^{-3}$, and we reduce the learning rate by 50% every 5 epochs. We set $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay is $10^{-6}$, and the total number of epochs is 15. In outlier removal, we select a $7 \times 7$ window as the neighborhood, $\epsilon$ and $\epsilon$ are set to 0.7 and 1, respectively. The joint training hyper-parameters $\lambda_{LU}$, $\lambda_{GU}$, and $\lambda_{Fused}$ in Eq. (9) are set to 0.3, 0.3, and 0.5, respectively. We use the same ResNet block as in [12].

KITTI Dataset. The KITTI dataset [13], [14] is a large autonomous driving real-world dataset, which consists of sparse depth maps obtained by raw LiDAR scans and corresponding color images. The dataset contains 86k training data, 1k selected validation data, and 1k test data without ground truth, which needs to be evaluated on the KITTI online benchmark server. Since the top of the sparse depth map does not have valid LiDAR points, we crop the input images to $256 \times 1216$ for both training and inference as [17].

DDAD Dataset. The Dense Depth for Autonomous Driving (DDAD) dataset [15] is a new autonomous driving benchmark from the Toyota Research Institute for long-range (up to 250m). The dataset contains monocular color images and accurate ground-truth depth generated from a high-density, long-range Luminar-H2 LiDAR sensor. Unlike the KITTI dataset captured under good lighting and weather conditions, the DDAD dataset contains scenes with different lighting conditions (such as morning, sunset, and overcast) and weather conditions (for example rain). We conduct experiments under various conditions in the DDAD dataset to verify the generalizability of our model. In experiments, we crop the input depth maps to $512 \times 1216$.

B. Comparison with state-of-the-art Methods

We evaluate our method on the KITTI depth completion. We model our approach end-to-end from scratch on the train set and compare the performance with other top-ranking depth-only completion methods on the KITTI leaderboard. As a depth-only solution, our method employs only depth data during training and inference. So we first compare it with the methods with the same settings, including non-learning methods IP Basic [11] and Physical Surface [10], as well as learning methods SparseConvs [4], Nconv [25], pNCNN [26], etc. As shown in Tab. II, our method outperforms all other methods in the main RMSE metric and achieves the best performance in MAE, iRMSE, and iMAE metrics.

Furthermore, we compare our method with other methods that use extra information during training. For example, Spade-SD [27] uses the synthetic data, Glob_guide [28] uses semantic
Table II: Quantitative comparison with state-of-the-art depth-only completion algorithms on the KITTI leaderboard. Whether the methods use extra information in the training is indicated in the column “ExInfo”.

| Methods         | iRMSE | iMAE | RMSE | MAE | ExInfo |
|-----------------|-------|------|------|-----|--------|
| Spade-sD [27]   | 2.60  | 0.98 | 1035.29 | 248.32 | Synthetic |
| Glob_guide [28] | 2.80  | 1.07 | 922.93  | 249.11 | Semantic |
| IR_L2 [24]      | 4.92  | 1.35 | 901.43  | 292.36 | RGB     |
| SparseConvs [4] | 4.94  | 1.78 | 1601.33 | 481.27 | No      |
| IR_Basic [11]   | 3.78  | 1.29 | 1288.46 | 302.60 | No      |
| ADNN [29]       | 59.39 | 3.19 | 1325.37 | 439.48 | No      |
| Nconv-CNN [25]  | 4.67  | 1.52 | 1268.22 | 360.28 | No      |
| S2D (depth-only) [6] | 3.21  | 1.35 | 954.36  | 288.64 | No      |
| HMS-Net [30]    | 2.93  | 1.14 | 937.27  | 258.48 | No      |
| pNCNN [26]      | 3.37  | 1.05 | 960.05  | 251.77 | No      |
| Physical_Surface [10] | 3.76 | 1.21 | 1239.84 | 298.30 | No      |
| DTP [31]        | 2.94  | 1.07 | 937.27  | 247.81 | No      |
| Ours            | 2.69  | 1.04 | 917.76  | 244.36 | No      |

segmentation labels, and IR_L2 [24] needs RGB as auxiliary information. Compared with these methods, our method still achieves comparable results. Specifically, our method achieves better results than Glob_guide [28] across all four metrics. Meanwhile, it has smaller RMSE metrics than Spade-sD [27] and smaller MAE, iRMSE, and iMAE than IR_L2 [24].

C. Ablation Study

1) Coupled U-Net Model: We first compare our method with the typical depth-only methods and RGB-guided methods, both of which adopt a single U-Net. The results in Tab. I show that our method improves all metrics compared with the typical depth-only methods, and the improvement in the areas of overlap and blanks is more significant. Since the overlap area is mainly concentrated on the boundary of the object, which is difficult for the depth-only method, the result of the overlap area is still far from that of other areas. Meanwhile, it is worth mentioning that our method adopts a more lightweight U-Net, so the network parameters are fewer.

By varying the input of the global U-Net, our method has six different variants, denoted by $M_1$ to $M_6$. The detailed configuration and comparison results of these six variants are shown in Tab. III, where SD denotes the sparse depth map, SCD denotes the sparse depth map after removing outlier points, and LU-Depth denotes the dense depth map predicted by the local U-Net. The results show that the setup of our method achieves the minimum RMSE that is the basis for the ranking of the KITTI leaderboard. In addition, we also verify the effect of employing an extra guidance image that is applied to guide the global U-Net (denoted by LU-Guidance), as done in [8] and [16]. The results show that the additional guidance maps do not provide any performance benefits.

Some typical examples are shown in Fig. 3, which show that the confidence images predicted by the local U-Net have high confidence values in the normal areas, whereas they have low confidence values in the overlap areas and blank areas. The confidence images predicted by the global U-Net are the opposite. This indicates the dense depth maps predicted by the local U-Net and the dense depth maps predicted by the global U-Net are complementary. We also make a visual comparison between the dense depth maps predicted by our
Table III: Performance of different models obtained by varying the input of global U-Net on the KITTI validation dataset. The best results are in bold, and our method, the results are highlighted with a gray background, achieves the minimum RMSE.

| Models        | SD | SCD | LU-Depth | LU-Guidance | iRMSE ↓ | iMAE ↓ | RMSE ↓ | MAE ↓ | Rel ↓ | δ↑ | δ↓ | δ↑↓ |
|---------------|----|-----|----------|-------------|---------|--------|--------|-------|------|----|----|-----|
| M₁            |    | ✓   | ✓        |             | 3.4     | 1.2    | 962.9  | 260.7 | 0.0146 | 99.59% | 99.87% | 99.95% |
| M₂            |    | ✓   |          |             | 2.9     | 1.1    | 959.8  | 254.9 | 0.0138 | 99.60% | 99.87% | 99.94% |
| M₃            |    | ✓   | ✓        |             | 2.9     | 1.1    | 960.4  | 258.0 | 0.0142 | 99.59% | 99.87% | 99.94% |
| M₄            |    | ✓   |          |             | 2.7     | 1.0    | 963.6  | 242.5 | 0.0129 | 99.60% | 99.87% | 99.94% |
| M₅            | ✓  | ✓   | ✓        |             | 2.8     | 1.0    | 958.8  | 245.3 | 0.0131 | 99.59% | 99.87% | 99.94% |
| M₆            | ✓  | ✓   |          |             | 2.8     | 1.1    | 959.8  | 249.2 | 0.0133 | 99.60% | 99.87% | 99.95% |

Table IV: Performance comparison with different outlier removal algorithms on the KITTI validation dataset.

| Depths        | iRMSE ↓ | iMAE ↓ | RMSE ↓ | MAE ↓ | Rel ↓ | δ↑ | δ↓ | δ↑↓ |
|---------------|---------|--------|--------|-------|------|----|----|-----|
| Raw depth map | 3.0     | 0.8    | 1595.2 | 202.5 | 100.00% |
| Baseline method [10] | 2.1 | 0.4 | 662.9 | 81.7 | 94.00% |
| Ours          | 1.7     | 0.3    | 466.1  | 73.2  | 95.34%  |

Table V: Ablation experiments of the outlier removal method on the KITTI validation dataset.

| Algorithms        | iRMSE ↓ | iMAE ↓ | RMSE ↓ | MAE ↓ | Keep ratio ↓ |
|-------------------|---------|--------|--------|-------|--------------|
| Baseline method   [10] | 2.8 | 1.0 | 958.8 | 245.3 |              |
| Ours              | 2.8 | 1.0 | 958.8 | 245.3 |              |

Figure 4: Examples of the outlier removal. (a) the RGB images; (b) the sparse input depths; (c) and (d) are the outlier mask predicted by our method and the depth map after removing outlier points; (e) and (f) are the outlier mask predicted by Physical_Surface [10] and the depth map after removing outlier points.

Figure 5: The performance of the methods under different depth densities.

D. Generalization Capability

To prove the generalization capability of our method, we evaluate its performance under different depth densities, changing lighting, and weather conditions.

1) Generalization on more sparse depth maps: The sparse depth map of the KITTI dataset is obtained by a 64-line Velodyne LiDAR. However, in many practical applications, only 32-line or 16-line LiDAR will be employed due to cost constraints, which only provide more sparse depth maps. Therefore, it is crucial to analyze the performance of the method on sparse depth maps with different sparsity levels.

We use the depth map obtained by the 64-line LiDAR as the baseline. Meanwhile, we employ the random sampling strategy to generate more sparse depth maps according to different ratios of density. In Fig. 5, we compare the performance of our method with S2D (depth-only) [6] and SparseConvs [4] under different sparsity levels on the KITTI validation dataset. To ensure the fairness of the comparison, our model is trained only on the KITTI train dataset without fine-tuning on depth maps with different sparsity levels, while the results of S2D (depth-only) are obtained by open-source code. Fig. 5 shows that the performance of the model gradually decreases with the depth density decreasing. However, our method outperforms other methods at different density ratios, which shows that our model has better generalization capability on more sparse depth maps.
2) Generalization under various conditions: As a depth-only solution, our method has stronger robustness under different conditions, as it is not affected by the quality and style of color images. To verify the generalization ability of our model in various conditions and datasets, we use the model trained on the KITTI dataset to directly test on five sequences with different lighting and weather conditions of the DDAE dataset. The experimental setup is difficult because there are many differences between training and inference data. We compare our method with the state-of-the-art RGB-guided method PENet [12]. The PENet model is provided by the open-source project, whose RMSE is 757 in the KITTI validation dataset. As shown in Fig. 6, compared with PENet, our method achieves better results in the five sequences and is relatively stable under changing conditions. It shows that our method has good generalization ability under various conditions.

VI. CONCLUSIONS

By investigating the reasons for the performance gap between the depth-only completion methods and the RGB-guided completion methods, we observed that the lack of reliable input information in the overlap areas and blank areas restricts the performance of the depth-only methods. To address this issue, we proposed the Coupled U-Net (CU-Net) model that is enhanced by a confidence-based outlier removal algorithm. Our method achieves the state-of-the-art performance on the KITTI leaderboard and has a good generalizability under different depth densities and various conditions. Meanwhile, our work advocates that a feasible way to improve the results in the overlap and blank areas, such as designing more complex algorithms or introducing data from other modalities, may be the focus of future research.
This part is the supplementary materials for “CU-Net: LiDAR Depth-only Completion with Coupled U-Net”. In Sec. A, we show qualitative experiments on the DDAD dataset. In Sec. B, we describe some typical examples of a continuous video sequence.

### A. QUALITATIVE EXPERIMENTS ON THE DDAD DATASET

The Dense Depth for Autonomous Driving (DDAD) dataset [15] is a new autonomous driving benchmark from Toyota Research Institute for long range (up to 250m). The dataset contains monocular color images and accurate ground-truth depth generated from a high-density, long-range Luminar-H2 LiDAR sensor. Unlike the KITTI dataset captured under good lighting and weather conditions, the DDAD dataset contains scenes with different lighting conditions (such as morning, sunset, and overcast) and weather conditions (for example, rain). We conduct experiments under various conditions in DDAD to verify the generalizability of our model. In experiments, we crop the input depth maps to $512 \times 1216$.

To verify the generalization of our model in various conditions and datasets, we use the model trained on the KITTI dataset to directly test on five sequences with different lighting and weather conditions of the DDAD dataset. The experimental setup is difficult because there are many differences between training and inference data. We compare our method with the state-of-the-art RGB-guided method PENet [12]. The PENet model is provided by the open-source project, whose RMSE is 757 in the KITTI validation dataset. Fig. 7 shows some visualization results of our model and the PENet on the DDAD dataset. The first two rows of samples are collected on rainy days, and the last two rows of samples are collected on overcast days. We observe that for adverse weather like rain, the dense depth map predicted by our method is more complete and clear compared to PENet. Meanwhile, for the overcast samples, the object boundaries in our dense depth maps, such as the edges of columns and street lights, are also sharper.

![Figure 7: Typical examples of our model and the PENet on the DDAD dataset.](https://youtu.be/WAQouCouw5E)

### B. EXPERIMENTAL RESULTS OF A CONTINUOUS VIDEO SEQUENCE

To show the experimental results more clearly, we provide some typical examples of a continuous video sequence. The results include the sparse depth map, the sparse depth map after removing outliers, the confidence map predicted by the local U-Net and the global U-Net, our results, and the ground truth. The results show that the confidence images predicted by the local U-Net have high confidence values in the normal areas. In contrast, they have low confidence values in the overlap areas and blank areas. The confidence images predicted by the global U-Net are the opposite. The video is released at https://youtu.be/WAQouCouw5E.