Improving Depth Completion via Depth Feature Upsampling

Yufei Wang[1], Ge Zhang[2], Shaoqian Wang[1], Bo Li[1], Qi Liu[1], Le Hui[1], Yuchao Dai[1]*
[1]Northwestern Polytechnical University and Shaanxi Key Laboratory of Information Acquisition and Processing [2]Beijing Institute of Tracking and Telecommunication Technology

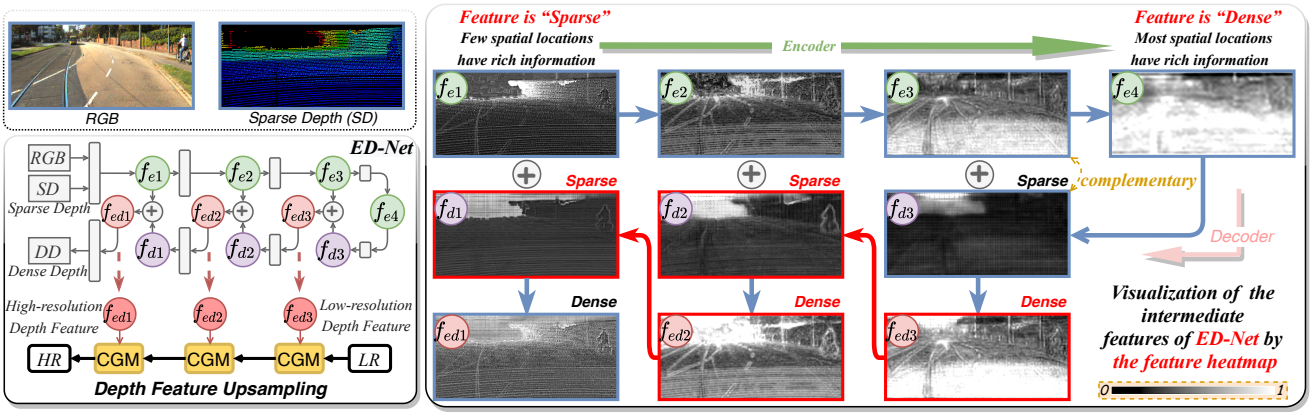

Figure 1. Typical method (S2D [22]) based on the encoder-decoder network (ED-Net) and **its intermediate features are visualized by the heatmap [46]**. The **sparse** decoder feature f_{di} derived from the **dense** fused encoder-decoder feature f_{edi} tends to complement the corresponding encoder feature f_{ei}, where the "**dense⇒sparse**" process destroys the completeness of features multiple times.

Abstract

*The encoder-decoder network (ED-Net) is a commonly employed choice for existing depth completion methods, but its working mechanism is ambiguous. In this paper, we visualize the internal feature maps to analyze how the network densifies the input sparse depth. We find that the encoder feature of ED-Net focus on the areas with input depth points around. To obtain a dense feature and thus estimate complete depth, the decoder feature tends to complement and enhance the encoder feature by skip-connection to make the fused encoder-decoder feature dense, resulting in the decoder feature also exhibits sparse. However, ED-Net obtains the **sparse** decoder feature from the **dense** fused feature at the previous stage, where the "**dense⇒sparse**" process destroys the completeness of features and loses information. To address this issue, we present a depth feature upsampling network (DFU) that explicitly utilizes these dense features to guide the upsampling of a low-resolution (LR) depth feature to a high-resolution (HR) one. The completeness of features is maintained throughout the upsampling process, thus avoiding information loss. Furthermore, we propose a confidence-aware guidance module (CGM), which is confidence-aware and performs guidance with adaptive receptive fields (GARF), to fully exploit the potential of these dense features as guidance. Experimental results show that our DFU, a plug-and-play module, can significantly improve the performance of existing ED-Net based methods with limited computational overheads, and **new SOTA results** are achieved. Besides, the generalization capability on sparser depth is also enhanced. Project page: https://npucvr.github.io/DFU.*

1. Introduction

Accurate and dense scene depth is crucial for various applications [13, 34, 41], such as autonomous navigation [33] and augmented reality [24]. Existing active depth sensors have been widely applied because they can obtain accurate depth information. However, the depth acquired by these sensors is generally highly sparse due to the imaging principles and power limitation. For instance, roughly 4% pixels in the depth provided by 64-line LiDAR have values [21, 22, 35]. To facilitate the use of the depth data, we need to address a *sparse-to-dense* problem in which the given sparse depth map is densified to the dense depth map, usually guided by the corresponding RGB image.

Depth completion is a challenging task as the input depth information is scarce. Existing approaches commonly

*Corresponding author: daiyuchao@nwpu.edu.cn.

consider this task as a pixel-wise regression problem and utilize an encoder-decoder network with skip-connection (**ED-Net**) to address it. For instance, the single-branch ED-Net [2, 14, 21, 22, 25] uses standard backbone networks [6, 19] as the encoder to extract features from the sparse depth and RGB images. Then, the decoder employs operations, such as the element-wise addition and transpose convolution, to fuse the encoder feature and upsampling. Recently, some multi-branch ED-Nets [4, 31, 32, 44, 47], including the dual-encoder ED-Net [4, 47] and the double encoder-decoder ED-Net[31, 32], *etc*., have been proposed. These methods employ separate branches to extract features from the sparse depth and RGB images, and internal features are fused at multiple scales. Although existing ED-Nets based methods have achieved considerable success [39, 48, 51], the ED-Net is often employed as a black box. ***how the network recovers a dense depth map from the input sparse depth*** has always been ambiguous.

To gain a deeper understanding of the ED-Net employed by depth completion, Fig. 1 visualizes internal features of the representative method S2D [22] through the feature heatmap [46]. The feature heatmap [46] is obtained by summing the value of the feature along the channel dimension and normalizing them to a range of (0, 1), which reflects the spatial locations that the network focuses on at different stages and provides valuable insights to infer the behaviors of the network. We observe that although the sparse depth map and dense RGB image are both fed to the network, the shallow feature of the encoder primarily focuses on few regions where the sparse depth map has values. Through multiple downsampling and convolution, the sparse regions of interest continue to expand and aggregate, the encoder feature gradually becomes more "dense". As shown in Fig. 1, the lowest-resolution encoder feature f_{e4} has rich information at most spatial locations. However, to obtain a dense feature and thus estimate complete depth, the decoder tends to obtain a complementary feature for the corresponding encoder feature, which contains multiple "**dense⇒sparse**" processes. For instance, the decoder feature f_{d1}, which is skip-connected to f_{e1}, tends to complement and enhance existing f_{e1} to make the fused encoder-decoder feature f_{ed1} dense. Therefore, the decoder feature f_{d1} also exhibits **sparse**. However, the **sparse** f_{d1} is derived from the **dense** f_{ed2} at the previous stage, where the "**dense⇒sparse**" process destroys the completeness of features and loses information. Experiments on multi-branch networks also validate this observation, which is provided in the supplemental material. To conclude, existing ED-Nets have not fully utilized the internal dense features, thereby restricting the performance of ED-Net based methods.

To address the issue, we present a depth feature upsampling network (DFU) that explicitly utilizes these dense features to guide the upsampling of a low-resolution (LR) depth feature to a high-resolution (HR) one. The completeness of features is maintained throughout the upsampling process, thus avoiding information loss. The multi-scale dense features that cover comprehensive scene depth information can progressively resolve ambiguity in super-resolving the feature of uneven depth distribution areas. Meanwhile, the upsampling process can also effectively integrate these dense features into the HR depth feature to predict the dense depth. To further improve the effectiveness of the DFU, we propose a confidence-aware guidance module (CGM) to fully exploit the potential of these dense features as guidance. Specifically, we first filter out unreliable values of the depth feature by predicting its confidence map, and then perform the guidance operation with adaptive receptive fields. In addition, the proposed network can be extended to multi-layer to achieve better results.

Our main contributions are summarized as:

- We analyze the working mechanism of popular ED-Nets by visualizing the internal feature maps and reveal that multiple "**dense⇒sparse**" processes exist in the decoder network, which destroy the completeness of features and lose information.
- We propose a depth feature upsampling network with confidence-aware guidance module, which is a plug-and-play module, to significantly improve the performance of existing ED-Net based methods and enhance the generalization ability on sparser depth with limited computational overheads.
- Extensive experiments on popular datasets prove the effectiveness of our method on existing popular ED-Nets, including single-branch, multi-branch, and SPN-based networks, and new SOTA results are achieved.

2. Related Work

Single-branch ED-Nets. The single-branch ED-Net contains one encoder and one decoder. The pioneer approaches [21, 26] concatenates the sparse depth and RGB image directly, and feeds them into a standard encoder-decoder network to predict the dense depth. Unlike using convolution in the last layer, [28] proposes to use the least squares to fit the relationship between the extracted features and depth values. Considering the gap between the RGB and depth information, subsequent methods [10, 20, 42] generally use two separate convolutional layers to extract features from the sparse depth and RGB image, respectively. Then, the multi-modal features are concatenated and fed into the network. This type of approach is simple and straightforward, but its performance is generally limited.

Multi-branch ED-Nets. Recently, some multi-branch ED-Nets [31, 32, 40, 44, 47] have been proposed for better extraction and fusion of the RGB and depth information. For example, the dual-encoder ED-Net [4, 47]

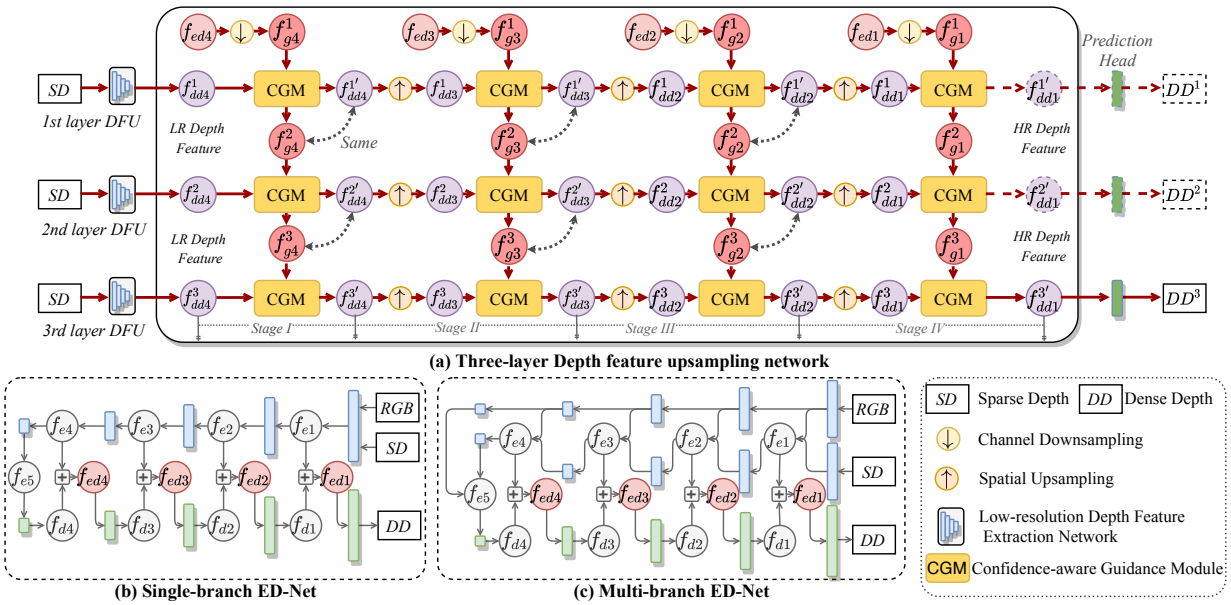

Figure 2. The proposed depth feature upsampling network (DFU). It explicitly employs the internal dense features of ED-Nets that are not fully utilized by existing methods to guide the upsampling of a low-resolution (LR) depth feature to a high-resolution (HR) one.

commonly employs two separate encoders to extract features from RGB images and sparse depth, respectively. Then, the extracted RGB and depth features are fused at single-scale or multi-scale by the channel-wise concatenation, element-wise summation, or other sophisticated fusion modules [32, 37, 44, 47]. GuideNet [32] proposes a double encoder-decoder network, which extracts the RGB and depth feature by a whole encoder-decoder network. The decoder feature of the RGB branch and the encoder feature of the depth branch are fused by the guided convolutional network at multiple scales. The multiple encoder-decoder network is proposed by RigNet [44], which uses repetitive hourglass networks to extract discriminative RGB features, and the fusion of RGB and depth information is also performed in a repetitive manner.

SPN-based Networks. The spatial propagation network (SPN) [2, 3, 14, 18, 25, 43] is a hot topic, which iteratively refines the predicted depth by aggregating the reference and neighbor pixels. The initial SPN [17] updates each pixel by three adjacent pixels from the previous row or column. CSPN [2] improves it by updating all pixels simultaneously by fixed-local neighbors, and CSPN++ [3] assembles the results predicted by using the neighbors in different ranges. To obtain non-local neighbors, DSPN [43] and NLSPN [25] learn the offsets to the regular grid. DySPN [14] gives variable weights to neighbors with different distances to make the kernel weights change dynamically during the update process. GraphCSPN [18] integrates 3D information into the update process. LRRU [36] proposes to employ multi-scale guidance features to guide the prediction of neighbors and weights, which makes them change adaptively throughout the update process.

3. Methodology

In this section, we present a depth feature upsampling network (DFU) that effectively utilizes internal dense features of ED-Net to guide the upsampling of a low-resolution (LR) depth feature to a high-resolution (HR) one. The DFU is introduced in three parts: (1) feature extraction, including extracting the LR depth feature and multi-scale dense features from ED-Net as guidance, (2) depth feature upsampling guided by the proposed confidence-aware module, (3) going deeper, where DFU is extended to multi-layer to obtain better results. For descriptive convenience, **the feature is marked as f_i, where $i = \{1, 2, 3, 4, 5\}$ denotes the feature of different resolutions**, namely $f_i \in \mathbb{R}^{H/n \times W/n \times D_i}, n = 2^{i-1}$.

3.1. Features Extraction

The LR depth feature $f_{dd4} \in \mathbb{R}^{H/8 \times W/8 \times D_{dd4}}$ is extracted from the sparse input depth using a convolutional network, where D_{dd4} is set to 64 in this paper. The network consists of 8 residual blocks, 2 at full-resolution, 2 at 1/2-resolution, 2 at 1/4-resolution, and 2 at 1/8-resolution. Although the input depth is sparse, the depth information continues to expand and aggregate by multiple downsampling and convolutional layers, making the LR feature f_{dd4} dense (most spatial positions of the feature contain rich information).

As shown in Fig. 2 (b) and (c), the guidance feature can be obtained from any ED-Net, including single-branch ED-Net, multi-branch ED-Net, *etc*. In this paper, we generally extract five encoder features $\{f_{e1}, f_{e2}, f_{e3}, f_{e4}, f_{e5}\}$ at different resolution. Then, the decoder features $\{f_{d4}, f_{d3}, f_{d2}, f_{d1}\}$ and fused encoder-decoder features

$\{f_{ed4}, f_{ed3}, f_{ed2}, f_{ed1}\}$ are obtained in the decoder. We select the features $\{f_{ed4}, f_{ed3}, f_{ed2}, f_{ed1}\}$ to guide the upsampling of the LR depth feature. These features are generally dense and cover comprehensive scene depth information, which can progressively resolve ambiguity in super-resolving the feature of uneven depth distribution areas. Furthermore, the upsampling process effectively integrates these dense features that are underutilized in existing ED-Nets, thus improving the accuracy of depth completion.

To reduce computational complexity, we reduce the channel number of these features $\{f_{ed4}, f_{ed3}, f_{ed2}, f_{ed1}\}$ to the same as the LR depth feature by

$$f_{gi} = \mathbf{Conv_{1\times1}}(f_{edi}), \tag{1}$$

where $\mathbf{Conv_{1\times1}}$ is a 1×1 convolutional layer, and the feature f_{gi} is employed as the guidance feature in the DFU.

3.2. Depth Feature Upsampling

As shown in Fig. 2 (a), the LR depth feature is upsampled to HR through four stages, which are guided by the features $\{f_{g4}, f_{g3}, f_{g2}, f_{g1}\}$, respectively. To fully exploit the potential of the guidance feature, we propose a **Confidence-aware Guidance Module (CGM)**, which first filters out unreliable values of the depth feature by predicting its confidence map, and performs the the guidance operation with adaptive receptive fields to handle uneven depth distribution. The first stage of the DFU consists of a CGM, while other stages adopt the "upsampling + CGM". As shown in Fig. 3, for the depth feature f_{ddi}, we denote the output feature of the CGM as f'_{ddi}. In this paper, the upsampling is implemented by using a deconvolutional layer of stride 2, and the details of the CGM are introduced as follows:

Confidence-aware. The LR depth feature is extracted from the input depth that is highly sparse and noisy. Besides, ambiguity in super-resolving the feature of uneven depth distribution areas often exists. Therefore, we predict an element-wise attention map to filter out unreliable feature values. Specifically, we concatenate the depth feature f_{ddi} and the guidance feature f_{gi}. The confidence map c_i is predicted by a 3×3 convolutional layer as:

$$c_i = \sigma(\mathbf{Conv_{3\times3}}(\mathbf{conc.}(f_{ddi}, f_{gi}))), \tag{2}$$

here σ denotes the sigmoid function to compress the confidence value to $0 - 1$.

Guidance with Adaptive Receptive Fields (GARF). In the field of depth completion, guided convolutional networks have been studied widely for multi-modal feature fusion. For example, the pioneering GuideNet [32] predicts spatially-variant depthwise convolutional kernels [30] $\mathbf{W} \in \mathbb{R}^{hw\times k^2\times c}$ from the RGB feature, and then applies them to extract the depth feature. Here, h, w, and c are the height, weight, and channel number of the feature map,

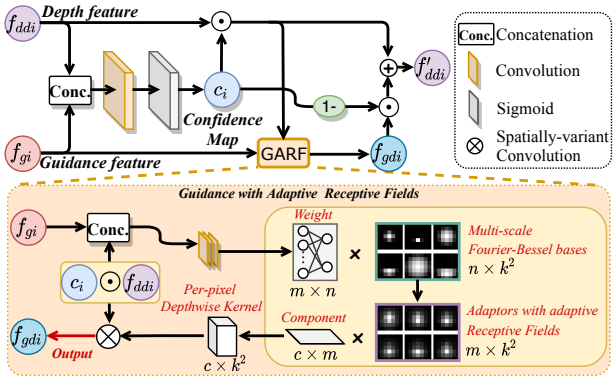

Figure 3. The Confidence-aware Guidance Module (CGM)

and k is the kernel size. To reduce computational complexity, [37] decomposes the depthwise kernel $\mathbf{W}_{\mathbf{p}_i} \in \mathbb{R}^{k^2\times c}$ at each spatial location $\mathbf{p}_i$ into content-adaptive adaptors $\mathbf{A} \in \mathbb{R}^{k^2\times m}$ multiplied by the spatially-shared component $\mathbf{D} \in \mathbb{R}^{m\times c}$. The guided convolutional network has been proven to be effective since it can transfer structural information from the guidance feature to the target feature [12]. Therefore, in our method, we adopt this strategy to fully utilize the guidance feature. However, the kernel size of existing guided convolutional networks is usually fixed, which cannot handle the areas with uneven depth distribution well. To address this issue, we propose a guidance module with adaptive receptive fields.

Recently, some novel dynamic convolution networks [27, 38] have shown that the convolutional kernel can be equivalently represented as a combination of pre-fixed bases, such as Fourier-Bessel(FB) bases [1]. Inspired by these works [27, 38], we further decompose the adaptors over multi-scale FB bases to selectively choose a receptive field at each spatial position. Specifically, we pre-generate n FB bases of different sizes, such as 3×3, 5×5, and 7×7, and then unify them to the same size by zero padding. Unlike [37] directly predicts the adaptors, we predict the weight of the multi-scale FB bases. As shown in Fig. 3, the adaptors with adaptive receptive fields can be obtained by multiplying the predicted weight with the FB bases. Then, we preform the guidance operation similar as [37] and obtain the refined depth feature f_{gdi} by:

$$f_{gdi} = GARF(c_i \odot f_{ddi}, f_{gi}) \tag{3}$$

Feature Updating. The confidence map c_i assigns low confidence to the unreliable depth features, which are exactly the areas we want the guidance module to prioritize. Therefore, we obtain the output of CGM f'_{ddi} by:

$$f'_{ddi} = c_i \odot f_{ddi} + (1 - c_i) \odot f_{gdi} \tag{4}$$

3.3. Going Deeper

To further enhance the performance of the proposed method, we extend the depth feature upsampling network

Table 1. The performance comparison between the baseline models and their improved model by using the multi-layer DFU.

			KITTI Validate Dataset [5]				NYUv2 Depth Dataset [23]				
	Params. [M]	FLOPs [G]	RMSE ↓ [mm]	MAE ↓ [mm]	iRMSE ↓ [1/km]	iMAE ↓ [1/km]	RMSE ↓ [mm]	REL ↓ (×1000)	$\delta_{1.25}$ ↑ [%]	$\delta_{1.25^2}$ ↑ [%]	$\delta_{1.25^3}$ ↑ [%]
Our single-branch ED-Net	11.95	291.24	772.87	210.26	2.28	0.95	111.94	16.62	99.32	99.88	**99.98**
Improving by **One-layer** DFU	+ 1.67	+ 43.79	750.58	205.32	2.14	0.93	100.67	13.92	99.46	**99.91**	**99.98**
Improving by **Two-layer** DFU	+ 2.41	+ 65.65	**745.27**	202.34	2.18	0.93	98.99	13.63	99.48	**99.91**	**99.98**
Improving by **Three-layer** DFU	+ 3.14	+ 87.52	746.98	**201.08**	**2.09**	**0.89**	**97.78**	**13.59**	**99.49**	**99.91**	**99.98**
Our Dual-branch ED-Net	22.28	498.84	753.81	204.39	2.12	0.92	101.12	14.40	99.46	99.91	**99.98**
Improving by **One-layer** DFU	+ 1.65	+ 38.34	741.39	199.65	2.00	0.88	98.63	14.21	99.49	**99.92**	**99.98**
Improving by **Two-layer** DFU	+ 2.38	+ 57.24	738.20	198.85	2.03	**0.87**	97.62	13.70	**99.50**	**99.92**	**99.98**
Improving by **Three-layer** DFU	+ 3.10	+ 76.15	**736.60**	**198.61**	**2.01**	**0.87**	**96.27**	**12.92**	**99.50**	**99.92**	**99.98**

to multi-layer. In this subsection, the feature is marked as f_i^j, where the superscript j denotes the feature in the j-th layer DFU. As shown in Fig. 2, we deepen the depth of each stage of the network. Specifically, we employ the feature f_{ddi}^j of the j-th layer DFU as the guidance feature f_{gi}^{j+1} of the corresponding module of the j+1-th layer DFU.

4. Experiments

4.1. Implementation Details

We employ PyTorch to implement our model, and conduct experiments with GeForce RTX 3090 GPUs. **Our training process is divided into two stages**, which both employ the $L1 + L2$ loss to train 35 epochs. The initial learning rate is 10^{-3}, and it is reduced by 50% every 5 epochs. We employ the AdamW optimizer with a batch size of 8, and set $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay is 10^{-6}. In the first stage, we first train the ED-Net to obtain the internal dense features that cover comprehensive scene depth information. Then, we train the depth feature upsampling network (DFU), in which these dense features are explicitly employed as the guidance features. For the multi-layer DFU, we only supervise the output of the last-layer DFU.

4.2. Datasets and Metrics

KITTI Dataset [5] consists of sparse depth maps projected from raw LiDAR scans and corresponding RGB images, which is a popular real-world autonomous driving dataset. The dataset contains 86k frames for training, 1k selected frames for validation, and 1k frames without ground truth that need to be tested on the online benchmark.

NYUv2 Dataset [23] is a popular indoor RGBD dataset. Following existing works [16, 36, 45, 51], our model is trained with 50K images sampled from the training set, and tested on 654 officially labeled images. The images of size 640×480 are downsampled to half and then center-cropped to 304×228 for both training and inference.

Evaluation Metrics. Following exiting methods [25, 36, 52], we employ Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), inverse RMSE (iRMSE), inverse MAE (iMAE), mean absolute relative error (REL), and percentage of pixels satisfying δ_τ for quantitative evaluation.

4.3. Improving Single-branch ED-Net

Following most methods [10, 20, 22, 42], the single-branch ED-Net used in this paper first extracts the depth and RGB features by two independent convolutional layers, and the features are concatenated and fed into an encoder-decoder network. The encoder consists of five stages, and each stage contains 2 residual blocks. The encoder feature is extracted at full-resolution in the first stage and is progressively halved in the next four stages. The feature channels of the five stages are 64, 128, 256, 256 and 256. We use the deconvolutional layer with stride 2 to upsample features, and fuse the encoder and decoder feature by the concatenate. As shown in Fig. 2 (b), the fused encoder-decoder features $f_{ed4}, f_{ed3}, f_{ed2}, f_{ed1}$ are employed as the guidance features in the proposed DFU.

As shown in Table 1, our single-branch ED-Net contains 11.95 M parameters and requires 291.24 G FLOPs, and the RMSE and MAE are 772.87 mm and 210.26 mm. Note that our single-branch ED-Net outperforms S2D [22] built on the Resnet-34 with fewer parameters since we use the effective stochastic depth strategy [8] in the residual block as [14]. We observe that the single-branch ED-Net is significantly improved by using one-layer DFU. Specifically, the improvement in terms of various metrics is 22.29 mm in RMSE, 4.94 mm in MAE, 0.14 in iRMSE, and 0.02 in iMAE. Meanwhile, the performance in terms of RMSE and MAE is further improved when we deepen the DFU to two layers. However, this improvement is relatively small compared to the previous one. When we extend the DFU to three layers, the MAE is further reduced, while the results in the term of RMSE appear saturated. In addition, the experimental results conducted on the NYUv2 dataset consistently demonstrate that our DFU can effectively improve the performance of the single-branch ED-Net.

Table 2. The performance comparison between LRRUs (SPN-based method) and their improved model by using the multi-layer DFU.

	LRRU-Mini				LRRU-Tiny				LRRU-Small				LRRU-Base			
KITTI [5] **Validate Dataset**	RMSE↓ [mm]	MAE↓ [mm]	iRMSE↓ [1/km]	iMAE↓ [1/km]	RMSE [mm]	MAE [mm]	iRMSE [1/km]	iMAE [1/km]	RMSE [mm]	MAE [mm]	iRMSE [1/km]	iMAE [1/km]	RMSE [mm]	MAE [mm]	iRMSE [1/km]	iMAE [1/km]
Baseline	806.27	210.16	2.28	0.89	763.80	198.89	2.12	0.85	745.31	195.69	2.00	0.83	729.49	188.78	1.92	0.80
+ One-layer DFU	787.28	206.56	2.27	0.88	747.39	195.68	2.04	0.83	732.42	192.27	1.96	0.81	718.27	188.05	1.89	0.80
+ Two-layer DFU	775.93	202.71	2.20	0.86	744.92	193.10	2.00	0.82	730.89	190.56	1.94	0.81	716.02	186.69	1.89	0.80
+ Three-layer DFU	**767.67**	**199.48**	**2.16**	**0.85**	**742.83**	**192.31**	**1.99**	**0.81**	**729.24**	**189.41**	**1.93**	**0.80**	**713.32**	**185.55**	**1.87**	**0.79**
NYUv2 [23] **Depth Dataset**	RMSE↓ [mm]	REL↓ (x1000)	$\delta_{1.25}$↑	$\delta_{1.25^2}$↑	RMSE [mm]	REL (x1000)	$\delta_{1.25}$	$\delta_{1.25^2}$	RMSE [mm]	REL (x1000)	$\delta_{1.25}$	$\delta_{1.25^2}$	RMSE [mm]	REL (x1000)	$\delta_{1.25}$	$\delta_{1.25^2}$
Baseline	100.86	13.34	99.44	**99.91**	95.36	12.25	99.51	99.92	93.36	11.85	99.53	99.92	91.27	11.21	99.56	99.92
+ One-layer DFU	99.22	12.77	99.46	**99.91**	92.58	11.76	99.54	**99.93**	91.57	11.48	**99.56**	99.93	90.79	**11.19**	**99.57**	**99.93**
+ Two-layer DFU	98.94	12.73	**99.47**	**99.91**	92.31	11.63	**99.55**	**99.93**	91.15	**11.43**	**99.56**	**99.98**	**90.61**	11.22	**99.57**	**99.93**
+ Three-layer DFU	**98.29**	**12.69**	**99.47**	**99.91**	**92.29**	**11.59**	**99.55**	99.92	**91.08**	**11.43**	**99.56**	**99.98**	90.77	11.27	**99.57**	**99.93**

4.4. Improving Multi-branch ED-Net

We experimentally found that the performance gap between dual-branch networks, double encoder-decoder networks, and multiple encoder-decoder networks is not significant. Therefore, we select the most cost-effective dual-encoder network as the baseline model to verify the improvement effect of our proposed method on the multi-branch ED-Net. As shown in Fig. 2 (c), the difference between the dual-encoder ED-Net and single-branch ED-Net is that the dual-encoder ED-Net utilizes two encoders to extract features from the sparse depth and corresponding RGB image, respectively. Specifically, the RGB encoder works independently, and the extracted RGB features at multiple scales are gradually injected into the depth encoder to effectively integrate information from different modalities.

As shown in Table 1, our dual-encoder ED-Net has nearly double the number of parameters and computational complexity in comparison to the single-branch ED-Net as it employs two encoders and the multi-scale fusion strategy. However, the dual-encoder ED-Net also outperforms the single-branch ED-Net significantly. When we apply one-layer DFU to the dual-encoder ED-Net, the errors including RMSE, MAE, iRMSE, and iMAE are reduced by 12.42 mm, 4.74 mm, 0.12 and 0.04, respectively. The performance of the method is continuously improved as the DFU deepens, while the magnitude of the performance improvement gradually decreases. We observe that the performance of the method closes saturation when using the three-layer DFU. Additionally, similar conclusions can be drawn from the results on the NYUv2 dataset, which adequately demonstrate that our proposed method can effectively improve the performance of the dual-branch ED-Net.

4.5. Improving SPN-based Model

The SPN is a popular module to refine the predicted depth map. To validate the effectiveness of our method for SPN-based methods, we apply the proposed DFU to LRRU [36], which introduces an effective SPN model and achieves top-ranking performance on the KITTI benchmark. As described in [36], LRRU has already utilized internal dense features of the guided-feature extraction network to guide the SPN model. However, these features are employed individually, and the features at different scales can not be aggregated to improve the robustness and effectiveness of the SPN model. To address this issue, we add the DFU between the guided-feature extraction network and the recurrent update process of the LRRU to integrate the information from the multi-scale guidance features (more details in the supplementary materials).

As shown in Table 2, we conduct comprehensive experiments on four variants of LRRU with varying network scales. On the KITTI dataset, the performance of all four variants of LRRU is significantly improved by using the DFU, especially for the mini model. Specifically, by using one-layer DFU, the RMSE of LRRU-Mini is reduced by 18.99 mm, and the RMSE of LRRU-Base is reduced by 11.22 mm. Furthermore, the performance of LRRUs continues to enhance at a gradually decreasing rate when utilizing the multi-layer DFU. The experimental results on the NYUv2 dataset consistently demonstrate the effectiveness of our method. Noted that, due to the limited number of samples of the NYUv2 dataset, when we continue to add our method to the LRRU-Base model with a large network size, the network appears to be overfitting. Therefore, on the NYUv2 dataset, our method only has a limited performance improvement effect on the LRRU-Base model.

4.6. Comparison with state-of-the-arts

To evaluate our methods against state-of-the-art (SOTA) methods, we compare their experimental results on the KITTI benchmark (outdoor) and NYUv2 (indoor). As shown in Table 3, we report the results of our baseline methods and the improved methods by using three-layer DFU. Meanwhile, we select some representative methods from the last three years for comparison. Since most SOTA methods employ the multi-branch ED-Net structure, the performance of our single-branch ED-Net is slightly worse

Table 3. Quantitative evaluation on the KITTI online leaderboard and NYUv2. The results of improved methods by using the three-layer DFU are highlighted with a gray background and the best and second-best results are highlighted in **red** and **blue** colors, respectively.

	KITTI Online Benchmark [5]				**NYUv2 Depth Dataset** [23]					Publication
	RMSE[mm]↓	MAE[mm]↓	iRMSE[1/km]↓	iMAE[1/km]↓	RMSE[m]↓	REL↓	$\delta_{1.25}$↑	$\delta_{1.25^2}$↑	$\delta_{1.25^2}$↑	
TWISE [11]	840.20	195.58	2.08	0.82	0.097	0.013	99.6	99.9	100.0	CVPR 2021
FCFRNet [15]	735.81	217.15	2.20	0.98	0.106	0.015	99.5	99.9	100.0	AAAI 2021
PENet [7]	730.08	210.55	2.17	0.94	-	-	-	-	-	ICRA 2021
ACMNet [49]	744.91	206.09	2.08	0.90	0.105	0.015	99.5	99.9	100.0	TIP 2021
PointFusion [9]	741.90	201.10	1.97	0.85	**0.090**	0.014	99.6	99.9	100.0	ICCV 2021
GFormer [29]	721.48	207.76	2.14	0.97	-	-	-	-	-	CVPR 2022
DySPN [14]	709.12	192.71	1.88	0.82	**0.090**	**0.012**	99.6	99.9	100.0	AAAI 2022
GraphCSPN [18]	738.41	199.31	1.96	0.84	**0.090**	**0.012**	99.6	99.9	100.0	ECCV 2022
RigNet [44]	712.66	203.25	2.08	0.90	**0.090**	0.013	99.6	99.9	100.0	ECCV 2022
CompletionFormer [48]	708.30	203.45	2.01	0.88	0.091	**0.012**	99.6	99.9	100.0	CVPR 2023
BEV@DC [51]	697.44	**189.44**	**1.83**	**0.82**	**0.089**	**0.012**	99.6	99.9	100.0	CVPR 2023
PointDC [45]	736.07	201.87	1.97	0.87	**0.089**	**0.012**	99.6	99.9	100.0	ICCV 2023
Our Single-branch ED-Net	745.16	209.86	2.22	0.95	0.112	0.016	99.3	99.9	100.0	-
Improved Single-branch	719.65	201.92	2.06	0.91	0.098	0.014	99.5	99.9	100.0	-
Our Dual-branch ED-Net	720.96	203.73	2.07	0.92	0.101	0.014	99.5	99.9	100.0	-
Improved Dual-branch	706.23	199.14	1.99	0.89	0.096	0.013	99.5	99.9	100.0	-
LRRU-Base (SPN-based) [36]	**696.51**	189.96	**1.87**	**0.81**	0.091	**0.011**	99.6	99.9	100.0	ICCV 2023
Improved LRRU-Base	**686.46**	**187.95**	**1.83**	**0.81**	0.091	**0.011**	99.6	99.9	100.0	-

in comparison. However, the performance of our single-branch ED-Net is significantly improved by using one-layer DFU. Specifically, on the KITTI benchmark, the errors including RMSE, MAE, iRMSE, and iMAE are reduced by 25.51 mm, 7.94 mm, 0.16, and 0.04, respectively. Besides, five metrics evaluated on the NYUv2 are also enhanced considerably. The experimental results demonstrate that the proposed DFU can effectively improve the performance of the single-branch ED-Net. Compared to the single-branch ED-Net, our dual-branch ED-Net shows better performance, which already outperforms the latest PointDC [45], GraphCSPN [18], GFormer [29], *etc*. in the term of RMSE. However, the proposed DFU still greatly improves the performance of the dual-branch ED-Net on two benchmarks. Note that our method is effective for the LRRU-Base (SPN-based method) as well, which is the current SOTA method on the KITTI benchmark. Specifically, the improved LRRU-Base by using three-layer DFU achieves new SOTA results, which **ranks *Top-1*** on the KITTI benchmark.

The results on the NYUv2 show that the methods [9, 18, 51, 51] that fuse 3D information generally achieve better results, such as BEV@DC [51] and PointDC [51]. Our method can also greatly improve the performance of the single-branch ED-Net and dual-branch ED-Net on NYUv2, while the performance improvement on LRRU-Base is negligible due to overfitting. Fig. 4 and Fig. 5 show the qualitative results of the baseline models and improved models by using three-layer DFU. Since our DFU effectively uses internal dense features of ED-Nets that cover comprehensive scene depth information, the improved method has more accurate results in fine and small structures, such as the gap area between two adjacent objects.

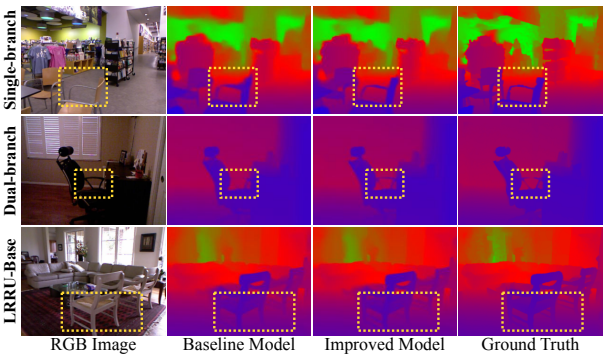

Figure 4. Qualitative results on the NYUv2 depth dataset.

4.7. Ablation Studies

In this section, we conduct experiments on the KITTI dataset to verify the effectiveness of the confidence-aware guidance module, which consists of the confidence-aware model and the guidance model with adaptive receptive fields. As shown in Table 4 (a) and (e), we choose our single-branch ED-Net as the baseline, and its results are significantly improved by using standard one-layer DFU.

Confidence-aware. In this paper, the LR depth feature is extracted from a highly sparse and noisy depth map, and ambiguity in super-resolving the feature of uneven depth distribution areas often exists. Therefore, some values of the depth feature are unreliable. To filter out the unreliable depth feature values before the guidance module, we propose to predict an element-wise confidence map, which assigns these unreliable values a lower weight. Besides, we also employ the confidence map in the feature updating, which makes the guidance module focus on the unreliable values of the depth features. As shown in Table 4 (d) and (e), removing the confidence map increases the RMSE and

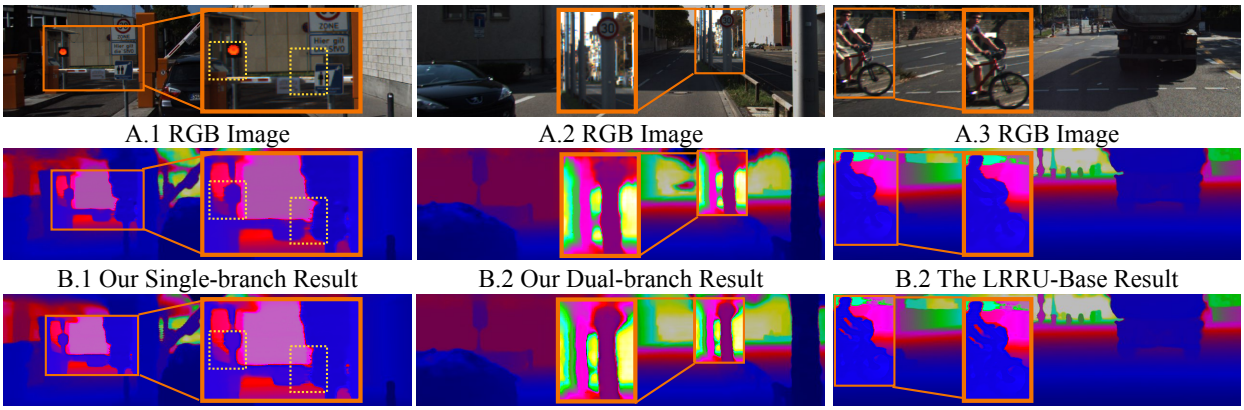

Figure 5. Qualitative results on the KITTI online leaderboard, including the baseline and improved models by using three-layer DFU.

Table 4. Ablation studies on the KITTI validation dataset.

	Confidence aware	GARF	RMSE [mm]	MAE [mm]
(a) Our Single-branch ED-Net	-	-	772.87	210.26
(b) Improving by one-layer DFU	-	-	761.24	209.04
(c) Improving by one-layer DFU	✓	-	758.67	207.51
(d) Improving by one-layer DFU	-	✓	756.31	206.82
(e) Improving by one-layer DFU	✓	✓	751.76	203.09

MAE of the improved method by 4.55 mm and 3.73 mm.

Guidance with Adaptive Receptive Fields. Traditional guided convolutional networks [32, 37, 44] generally predict spatially-variant convolutional kernels from the guidance feature, and then employ the predicted kernels to extract the target feature. These networks have been proven effective since they can transfer the structural information from the guidance feature to the target feature. However, the predicted per-pixel kernel has a fixed size, which cannot be suitable for the areas with different depth distributions. To address this issue, we propose the guidance module with adaptive receptive fields (GARF). The results in Table 4 (c) and (e) show that the performance of the improved method without GARF is significantly decreased, which demonstrates the effectiveness of GARF.

4.8. Experiments on fewer points

It is crucial to analyze the generalization ability for sparser depth maps, which are usually provided in many practical applications. We train our baseline methods and their improved methods by using three-layer DFU on the standard KITTI (64-line LiDAR depth) and test the performance (MAE[mm]) on the depth map with fewer lines. The sparser depth map is obtained by the method provided by [50]. As shown in Fig. 6, the performance improvement by using the proposed DFU is more significant on sparser depth maps. The experimental results demonstrate that our approach not only improves the performance of the method, but also enhances the generalization ability for sparser depth maps.

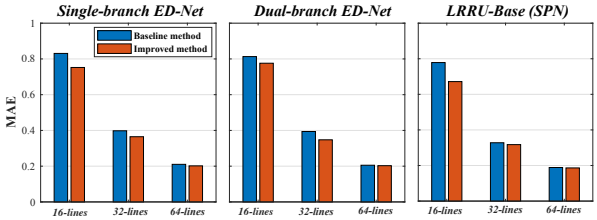

Figure 6. The performance (MAE [mm]) on fewer points.

4.9. Computational Cost

We report the parameters and FLOPs of the method in Table 1. Since we pre-reduce the channel number of the feature, our method improves the performance of existing ED-Nets with limited computational overheads. Specifically, using one-layer DFU approximately requires extra 1.7 M parameters and 40 G FLOPs, which are small compared to those of the ED-Net itself.

5. Conclusion

In this paper, we have proposed the depth feature upsampling network (DFU), a plug-and-play module to improve existing ED-Net based methods. By the proposed confidence-aware guidance module, DFU effectively utilizes internal dense features of ED-Net to guide the depth feature upsampling, the completeness of these dense features is maintained in DFU. Furthermore, DFU can be extended to multi-layer to achieve better results. Experimental results show that our method can significantly improve existing ED-Nets, including single-branch ED-Nets, multi-branch ED-Nets, and SPN-based methods, with limited computational overheads. Meanwhile, the generalization ability for sparser depth is also enhanced.

Acknowledgments

This work was supported by the National Science Fund of China (Grant Nos.62001394, 62271410, 62306238) and the Fundamental Research Funds for the Central Universities.

References

[1] Milton Abramowitz, Irene A Stegun, and Robert H Romer. Handbook of mathematical functions with formulas, graphs, and mathematical tables. courier corporation, 1988. 4

[2] Xinjing Cheng, Peng Wang, and Ruigang Yang. Learning depth with convolutional spatial propagation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2019. 2, 3

[3] Xinjing Cheng, Peng Wang, Chenye Guan, and Ruigang Yang. Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 3

[4] Chen Fu, Chiyu Dong, Christoph Mertz, and John M Dolan. Depth completion via inductive fusion of planar lidar and monocular camera. In *International Conference on Intelligent Robots and Systems (IROS)*, 2020. 2

[5] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 5, 6, 7

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[7] Mu Hu, Shuling Wang, Bin Li, Shiyu Ning, Li Fan, and Xiaojin Gong. Penet: Towards precise and efficient image guided depth completion. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2021. 7

[8] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 5

[9] Lam Huynh, Phong Nguyen, Jiří Matas, Esa Rahtu, and Janne Heikkilä. Boosting monocular depth estimation with lightweight 3d point fusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 7

[10] Saif Imran, Yunfei Long, Xiaoming Liu, and Daniel Morris. Depth coefficients for depth completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 5

[11] Saif Imran, Xiaoming Liu, and Daniel Morris. Depth completion with twin surface extrapolation at occlusion boundaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[12] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable kernel networks for joint image filtering. *International Journal of Computer Vision (IJCV)*, 2021. 4

[13] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[14] Yuankai Lin, Tao Cheng, Qi Zhong, Wending Zhou, and Hua Yang. Dynamic spatial propagation network for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 2, 3, 5, 7

[15] Lina Liu, Xibin Song, Xiaoyang Lyu, Junwei Diao, Mengmeng Wang, Yong Liu, and Liangjun Zhang. Fcfr-net: Feature fusion based coarse-to-fine residual learning for depth completion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 7

[16] Lina Liu, Xibin Song, Jiadai Sun, Xiaoyang Lyu, Lin Li, Yong Liu, and Liangjun Zhang. Mff-net: Towards efficient monocular depth completion with multi-modal feature fusion. *IEEE Robotics and Automation Letters (RAL)*, 2023. 5

[17] Sifei Liu, Shalini De Mello, Jinwei Gu, Guangyu Zhong, Ming-Hsuan Yang, and Jan Kautz. Learning affinity via spatial propagation networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[18] Xin Liu, Xiaofei Shao, Bo Wang, Yali Li, and Shengjin Wang. Graphcspn: Geometry-aware depth completion via dynamic gcns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 3, 7

[19] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[20] Yangqi Long, Huimin Yu, and Biyang Liu. Depth completion towards different sensor configurations via relative depth map estimation and scale recovery. *Journal of Visual Communication and Image Representation (JVCIR)*, 2021. 2, 5

[21] Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018. 1, 2

[22] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2019. 1, 2, 5

[23] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012. 5, 6, 7

[24] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohli, Jamie Shotton, Steve Hodges, and Andrew W. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011. 1

[25] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 5

[26] Jiaxiong Qiu, Zhaopeng Cui, Yinda Zhang, Xingdi Zhang, Shuaicheng Liu, Bing Zeng, and Marc Pollefeys. Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image.

In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[27] Qiang Qiu, Xiuyuan Cheng, Guillermo Sapiro, et al. Dcfnet: Deep neural network with decomposed convolutional filters. In *International Conference on Machine Learning (ICML)*, 2018. 4

[28] Chao Qu, Ty Nguyen, and Camillo Taylor. Depth completion via deep basis fitting. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2020. 2

[29] Kyeongha Rho, Jinsung Ha, and Youngjung Kim. Guideformer: Transformers for image guided depth completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7

[30] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4

[31] René Schuster, Oliver Wasenmuller, Christian Unger, and Didier Stricker. Ssgp: Sparse spatial guided propagation for robust and generic interpolation. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, 2021. 2

[32] Jie Tang, Fei-Peng Tian, Wei Feng, Jian Li, and Ping Tan. Learning guided convolutional network for depth completion. *IEEE Transactions on Image Processing (TIP)*, 2020. 2, 3, 4, 8

[33] Jiadai Wang, Jiajia Liu, and Nei Kato. Networking and communications in autonomous driving a survey. *IEEE Communications Surveys and Tutorials (CST)*, 2018. 1

[34] Shaoqian Wang, Bo Li, and Yuchao Dai. Efficient multi-view stereo by iterative dynamic cost volume. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[35] Yufei Wang, Yuchao Dai, Qi Liu, Peng Yang, Jiadai Sun, and Bo Li. Cu-net: Lidar depth-only completion with coupled u-net. *IEEE Robotics and Automation Letters (RAL)*, 2022. 1

[36] Yufei Wang, Bo Li, Ge Zhang, Qi Liu, Tao Gao, and Yuchao Dai. Lrru: Long-short range recurrent updating networks for depth completion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 3, 5, 6, 7

[37] Yufei Wang, Yuxin Mao, Qi Liu, and Yuchao Dai. Decomposed guided dynamic filters for efficient rgb-guided depth completion. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2023. 3, 4, 8

[38] Ze Wang, Zichen Miao, Jun Hu, and Qiang Qiu. Adaptive convolutions with per-pixel dynamic filter atom. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 4

[39] Alex Wong and Stefano Soatto. Unsupervised depth completion with calibrated backprojection layers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 2

[40] Alex Wong, Xiaohan Fei, Stephanie Tsuei, and Stefano Soatto. Unsupervised depth completion from visual inertial odometry. *IEEE Robotics and Automation Letters (RAL)*, 2020. 2

[41] Mochu Xiang, Yuchao Dai, Feiyu Zhang, Jiawei Shi, Xinyu Tian, and Zhensong Zhang. Towards a unified network for robust monocular depth estimation: Network architecture, training strategy and dataset. *International Journal of Computer Vision (IJCV)*, 2023. 1

[42] Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, and Hongsheng Li. Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 5

[43] Zheyuan Xu, Hongche Yin, and Jian Yao. Deformable spatial propagation networks for depth completion. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2020. 3

[44] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Baobei Xu, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 7, 8

[45] Zhu Yu, Zehua Sheng, Zili Zhou, Lun Luo, Si-Yuan Cao, Hong Gu, Huaqi Zhang, and Hui-Liang Shen. Aggregating feature point cloud for depth completion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 5, 7

[46] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 1, 2

[47] Yongchi Zhang, Ping Wei, Huan Li, and Nanning Zheng. Multiscale adaptation fusion networks for depth completion. In *International Joint Conference on Neural Networks (IJCNN)*, 2020. 2, 3

[48] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 7

[49] Shanshan Zhao, Mingming Gong, Huan Fu, and Dacheng Tao. Adaptive context-aware multi-modal network for depth completion. *IEEE Transactions on Image Processing (TIP)*, 2021. 7

[50] Yiming Zhao, Lin Bai, Ziming Zhang, and Xinming Huang. A surface geometry model for lidar depth completion. *IEEE Robotics and Automation Letters (RAL)*, 2021. 8

[51] Wending Zhou, Xu Yan, Yinghong Liao, Yuankai Lin, Jin Huang, Gangming Zhao, Shuguang Cui, and Zhen Li. Bev@ dc: Bird's-eye view assisted training for depth completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 5, 7

[52] Yufan Zhu, Weisheng Dong, Leida Li, Jinjian Wu, Xin Li, and Guangming Shi. Robust depth completion with uncertainty-driven loss functions. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022. 5