Improving Depth Completion via Depth Feature Upsampling – Supplementary Materials –

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

Abstract

In the main paper, we have proposed the depth feature upsampling network (DFU), a plug-and-play module to improve existing methods based on the encoder-decoder network (ED-Net). In this supplementary material, we first introduce our motivation in more detail. Then, we provide addtional experiments on multi-branch networks, details of improving LRRU, etc. *Following the main paper, the feature is marked as f_i, where f_i denotes the feature of different resolutions, namely $f_i \in \mathbb{R}^{H/n \times W/n \times D_i}, n = 2^{i-1}$.*

1. More details on motivation

The Encoder-Decoder network with skip-connection (ED-Net) is a popular framework for depth completion, but its working is ambiguous as pointed out by R2. In this paper, we visualize the inside activation maps to help us understand the learned features and how the network processes input data. As shown in Fig. 1 (a), we found that the encoder features f_{ei} of ED-Net focus on the areas with input depth points around. Therefore, to obtain a dense feature and thus estimate complete depth, the decoder feature f_{di}, which is skip-connected to f_{ei}, tends to complement and enhance existing f_{ei} to make the fused encoder-decoder feature f_{edi} dense, resulting in the decoder feature f_{di} exhibits **sparse**. However, existing ED-Nets obtain the **sparse** f_{di} from the **dense** $f_{ed(i+1)}$ at the previous stage, where the "**dense⇒sparse**" process will lose partial information of established dense features at multiple scales.

Inspired by the visualization, we propose a small plug-and-play module, DFU containing **only 1.67M**, to improve existing ED-Nets, which explicitly utilizes these dense features f_{edi} before being destroyed. Fig. 1 (b) and (c) show that the completeness of features is maintained regardless of whether DFU uses the basic addition or sophisticated CGM (proposed in this paper) to fuse f_{edi} (actually f_{edi} will be downsampled by the channel to obtain the guidance feature f_{gi}) and f_{ddi}. Moreover, the output feature of DFU is denser and smoother. The quantitative results in Fig. 1 (on KITTI) also verify the effectiveness of our method.

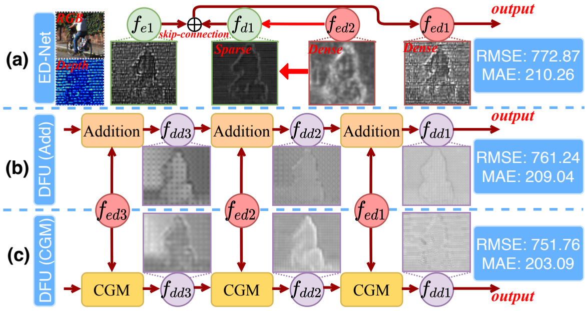


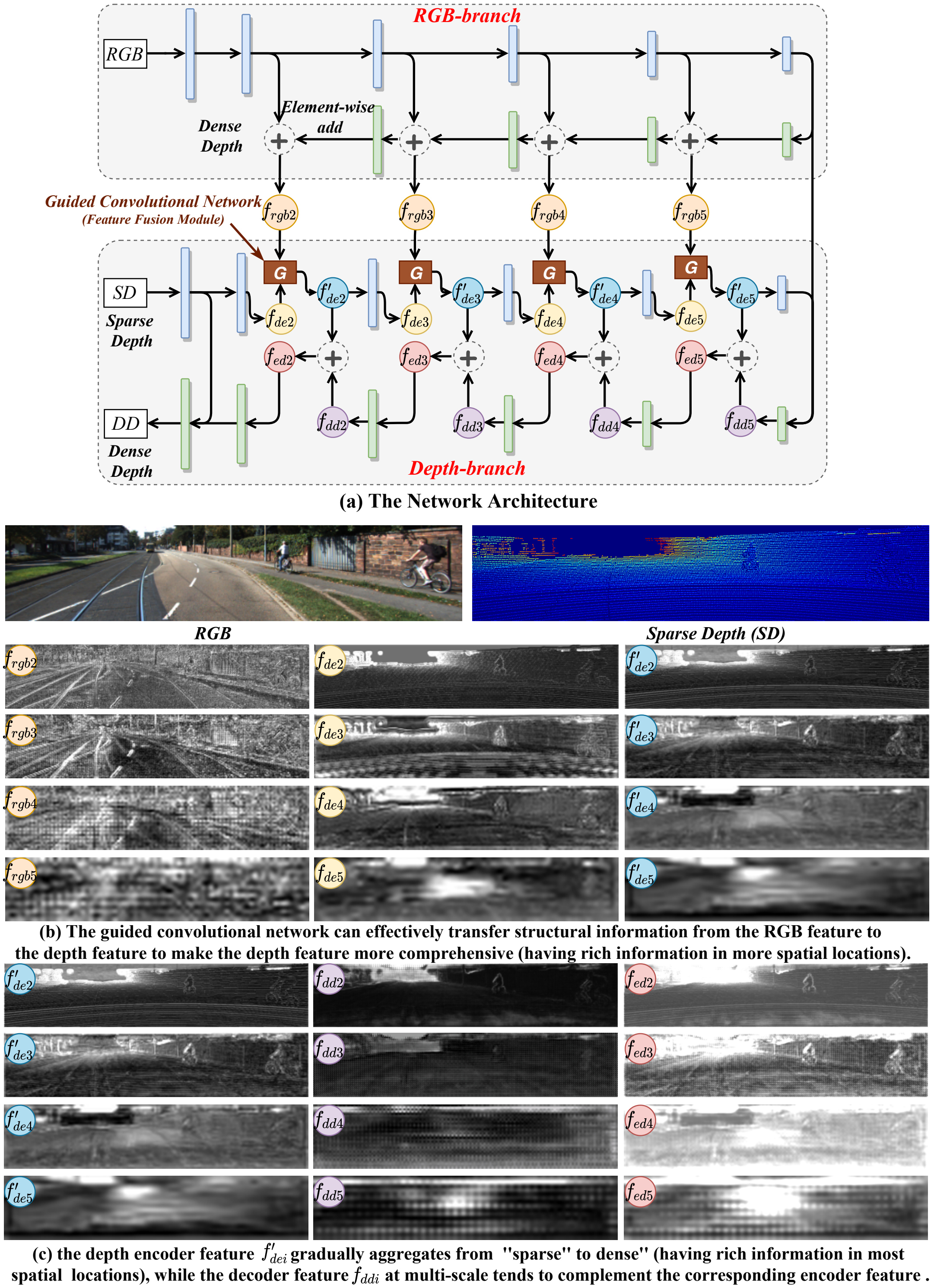
Figure 1. Activation maps and results. **Input depth is dilated for show**.

2. Experiments on multi-branch networks.

For the single-branch ED-Net, we have shown that the encoder feature gradually aggregates from "sparse" to "dense", while the decoder feature at multi-scale tends to complement the corresponding encoder feature, such as in the areas that lack input depth information and object boundary (more details in the "Introduction" section of the main paper). To demonstrate that the multi-branch ED-Net performs similarly to the single-branch ED-Net, we conduct additional experiments on the representative approach GuideNet [5], which is based on the multi-branch ED-Net. The network employed by GuideNet [5] consists of an RGB-branch and a Depth-branch, which uses a whole encoder-decoder sub-network. The RGB-branch extracts RGB features at multiple scales f_{rgbi}. Then, these RGB features are gradually injected into the Depth-branch to effectively integrate RGB and depth information by the proposed guided convolutional network.

As shown in Fig. 2, we visualize intermediate features of the GuideNet [5] through the feature heatmap [7]. We observe that the intermediate features of the RGB-branch have

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



(a) The Network Architecture

(b) The guided convolutional network can effectively transfer structural information from the RGB feature to the depth feature to make the depth feature more comprehensive (having rich information in more spatial locations).

(c) the depth encoder feature f'_{dei} gradually aggregates from "sparse" to dense" (having rich information in most spatial locations), while the decoder feature f_{ddi} at multi-scale tends to complement the corresponding encoder feature .

Figure 2. Typical method (GuideNet [5]) based on the multi-branch ED-Net and **its intermediate features are visualized by the heatmap [7]**. Note that the encoder feature gradually aggregates from "sparse" to "dense", while the decoder feature of the Depth-branch at multi-scale f_{ddi} tends to complement the corresponding encoder feature f'_{dei}, such as in the areas that lack input depth information and object boundary.

rich information in most spatial locations since the RGB image is dense and the RGB-branch works independently. However, the shallow feature of the Depth-branch, such as f_{de2}, primarily focuses on few regions where the sparse depth map has values. Since the guided convolutional network can effectively transfer structural information from the RGB feature to the depth feature, the depth feature after the feature fusion module based on the guided convolutional network f'_{de2} has rich information in more spatial locations. Then, through multiple downsampling and guided convolutional networks, the feature f'_{dei} becomes more "dense".

However, like the single-branch EDNet, the decoder of the multi-branch ED-Net also tends to obtain a complementary feature for the corresponding encoder feature, which undergoes a "**sparse⇒dense⇒sparse**" procedure. For example, the learned decoder feature f_{dd3} focuses more on the areas that lack input depth information and object boundary, which are not well considered by the corresponding encoder feature f'_{de3}. Then, a dense feature f_{ed3} is obtained by fusing the paired encoder and decoder feature. However, the "dense" feature is utilized to obtain a "sparse" decoder feature f_{dd2} at the next stage, which destroys the completeness of features and loses information.

3. Details of improving LRRU.

To verify the effectiveness of the proposed DFU for SPN-based methods, we apply it to LRRU [6], which introduces a flexible SPN model and achieves top-ranking performance on the KITTI benchmark. LRRU explicitly employs intermediate dense features of the guided-feature extraction network to guide the SPN model. However, these features are utilized individually, and the features at different scales can not be aggregated to improve the robustness and effectiveness of the SPN model. As shown in Fig. 3, we embed 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.

In the training process of the improving LRRU, we first load the pre-trained LRRU model provided by the authors and use the same settings as LRRU. For the LRRU-Base model, training the improving model by using one-layer DFU with four 3090 GPUs requires five days.

4. Visualization of intermediate features of improving ED-Net.

By visualizing the intermediate features of the model, we have shown that the decoder of existing ED-Nets tends to obtain a complementary feature for the corresponding encoder feature, which undergoes a "**sparse⇒dense⇒sparse**" procedure. Therefore, the intermediate dense features are not fully utilized by existing ED-Nets, thereby restricting the performance of methods based on such networks. To address this issue, we propose a depth feature upsampling network (DFU) that effectively utilizes these dense features to guide the upsampling of a low-resolution (LR) depth feature to a high-resolution (HR) one. As shown in Fig. 4, we observe that the completeness of depth features is maintained throughout the upsampling process, thus avoiding information loss.

5. Comparison with existing multi-scale guidance methods.

The multi-scale guidance strategy has been widely studied in existing multi-branch ED-Nets. These methods employ two separate branches to extract features from RGB images and sparse depth, respectively. Then, extracted RGB features are used to guide the extraction process of depth features in multiple scales. Unlike existing methods, we propose to effectively utilize intermediate dense features of the pre-trained ED-Net to guide the upsampling of the depth feature, where the intermediate dense features contain rich information across most spatial locations. The benefits are two-fold. First, we explicitly employ multi-scale intermediate dense features of the ED-Net whose completeness is destroyed in existing methods, thus avoiding information loss. Second, we propose a confidence-aware guidance module (CGM) to fully exploit the potentiality of these dense features as guidance. The features at different scales are aggregated to improve the robustness and effectiveness of the guidance process. In addition, the proposed network can be extended to multi-layer to achieve better results.

6. Qualitative comparison with SOTA

In Fig. 5, we show the qualitative results on the KITTI depth completion online benchmark, including CSPN [1], NLSPN [4], DySPN [2], LRRU-Base [6], and improving LRRU-Base model by using one-layer DFU. Our DFU effectively uses intermediate dense features of ED-Nets that cover comprehensive scene depth information. Therefore, the dense depth predicted by the improving LRRU-Base model has better results in fine and small structures, such as the gap area between two adjacent objects.

7. Evaluation metrics

Following exiting depth completion methods[2, 3], we employ the Root Mean Squared Error (RMSE[mm]), Mean Absolute Error (MAE[mm]), Root Mean Squared Error of the Inverse depth (iRMSE[$1/\mathrm{km}$]), Mean Absolute Error of the Inverse depth (iMAE[1/km]), mean absolute relative error (REL), and percentage of pixels satisfying δ_τ for quantitative evaluation. Eq. (1) shows the detailed definitions, where d^{gt} denotes the ground truth depth map, d^{pred} denotes the predicted dense depth map, and $\mathcal{V}$ is the set of available points in the ground truth.

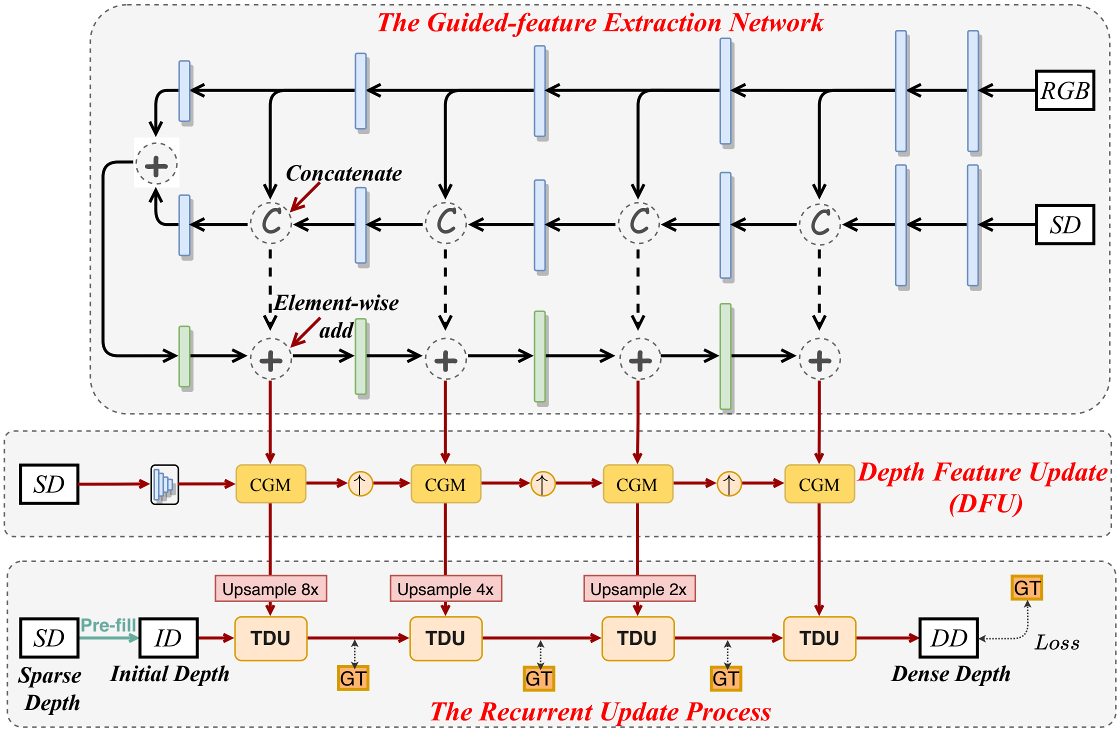


Figure 3. The network architecture of improving LRRU by using one-layer DFU.

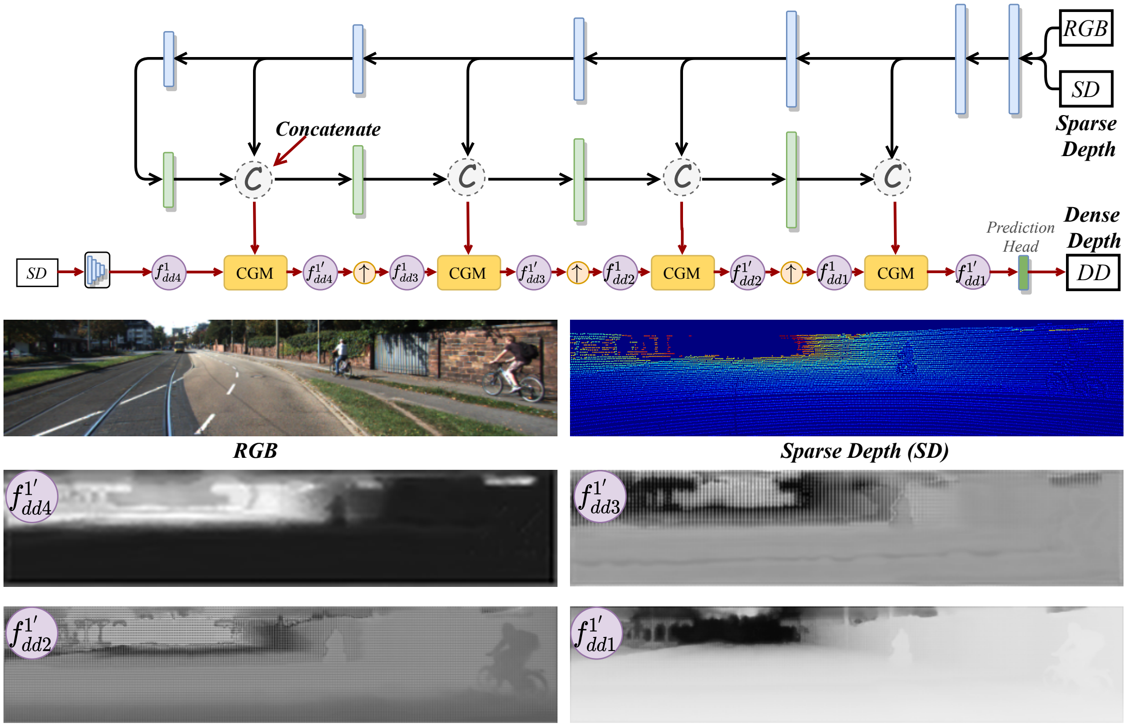


Figure 4. The network architecture of improving S2D [3] by using one-layer DFU, and its intermediate features are visualized by the heatmap [7]. The completeness of depth features is maintained throughout the upsampling process, thus avoiding information loss.

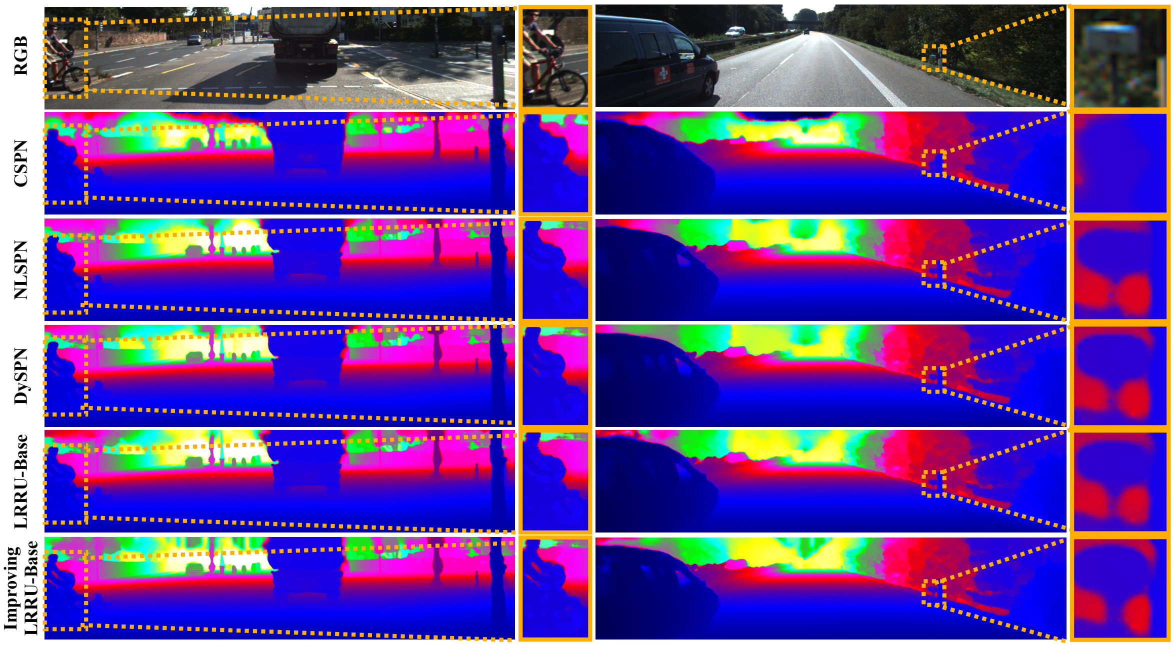


Figure 5. Qualitative results on the KITTI depth completion online benchmark, including CSPN [1], NLSPN [4], DySPN [2], LRRU-Base [6], and improving LRRU-Base model by using one-layer DFU. Some regions are zoomed-in for better visualization.

$$
\begin{aligned}
&\text{RMSE[mm]}: \sqrt{\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left|d_v^{gt}-d_v^{pred}\right|^2},\\
&\text{MAE[mm]}: \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left|d_v^{gt}-d_v^{\text{pred}}\right|,\\
&\text{iRMSE[1/km]}: \sqrt{\frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left|1/d_v^{gt}-1/d_v^{\text{pred}}\right|^2},\\
&\text{iMAE[1/km]}: \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left|1/d_v^{gt}-1/d_v^{\text{pred}}\right|,\\
&\text{REL}\ : \frac{1}{|\mathcal{V}|}\sum_{v\in\mathcal{V}}\left|\left(d_v^{gt}-d_v^{pred}\right)/d_v^{gt}\right|,\\
&\delta_\tau[\%]: \max\left(\frac{d_v^{gt}}{d_v^{pred}},\frac{d_v^{pred}}{d_v^{gt}}\right)<\tau.
\end{aligned}
\tag{1}
$$

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] 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. 3, 5

[2] 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. 3, 5

[3] 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. 3, 4

[4] 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. 3, 5

[5] 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. 1, 2

[6] 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

[7] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014. 1, 2, 4