A Two-stage hybrid CNN-Transformer Network for RGB Guided Indoor Depth Completion

Yufan Deng
School of Cyber
Science and Technology
Beihang University
Beijing, China
yfdeng@buaa.edu.cn

Xin Deng*
School of Cyber
Science and Technology
Beihang University
Beijing, China
cindydeng@buaa.edu.cn

Mai Xu
School of Electronic
Information Engineering
Beihang University
Beijing, China
maixu@buaa.edu.cn

***Abstract*—The indoor captured raw depth images usually contain large in-homogeneous missing regions. Most existing methods are designed for the outdoor sparse depth completion, which struggle in completing the indoor depth with large holes. In this paper, to solve this problem, we propose a hybrid CNN-Transformer network for RGB guided indoor depth completion. The proposed network is composed of two stages to achieve depth completion in a coarse-to-fine manner. In the first stage, we propose a CNN based self-completion module (SCM) with cross scale attention to restore a coarse depth image. In the second stage, we further refine the completed depth image with the guidance of RGB image by proposing a guided completion module (GCM). To fully explore the guidance from the RGB image, we design a cross-modal Transformer (CMT) block to fuse the features from the depth and RGB modalities at different scales. Extensive experiments on NYUv2 and SUN RGB-D datasets demonstrate the superior performance of the proposed method over other state-of-the-art methods both quantitatively and qualitatively. The code is available at https://github.com/eecoder-dyf/ICME-2023-depth-completion.**

***Index Terms*—Indoor depth completion, RGB-D images, Transformer network.**

I. Introduction

Recent years have witnessed the widespread use of depth sensors on various devices, from smart phones to autonomous cars and robots. Generally, the depth sensors can be divided into two categories, the optical sensors (such as structured light) and non-optical sensors (such as LiDAR). The non-optical sensors are usually used in outdoor scenes which can capture sparse depth point clouds, while the optical depth sensors are usually used in indoor scenes which capture depth maps with more dense distributions but large missing regions. Compared to the sparse depth, the depth captured by the optical sensors is also called semi-dense depth [1]. Due to physical limitations, the raw depth images are usually with various missing regions, which cannot work well in applications such as 3D reconstruction. Recently, due to the rapid growth of autonomous cars equipped with LiDAR sensors, there has been an increasing amount of research focusing on completing sparse depth. [2]–[4]. Nevertheless, the research on semi-dense depth completion in indoor scenes is still rare.

Compared with sparse depth completion based on LiDAR sensors, the completion of semi-dense depth is more challenging since its missing regions can be very large with even a whole object missing. The traditional methods for semi-dense depth completion [5], [6] are usually designed for stereoscopic conditions, which requires the stereo image dataset. The recent deep learning based methods [7], [8] require surface normal as the ground truth, which is often difficult to calculate in practical applications. In this paper, we focus on RGB guided semi-dense depth completion with no need of the surface normal. Specifically, we propose a two-stage depth completion network to achieve high accurate depth maps in a coarse-to-fine manner. The first stage is named as self completion module (SCM), in which the raw depth image is the only input and the network attempts to complete the depth image by its own information. The SCM is a CNN based encoder-decoder network with our proposed cross scale attention (CSA) block to achieve mutual guidance between multi-scale features in both encoder and decoder. The second stage is named as guided completion module (GCM), in which the network aims to further refine the completed depth image with the guidance of RGB image. The GCM is a hybrid CNN-Transformer network, in which we propose cross modal transformer (CMT) block to achieve better feature fusion between depth and RGB modalities. The contributions of this paper are as follows:

- We design a two-stage hybrid CNN-Transformer network to achieve RGB guided depth image completion in a coarse-to-fine manner.
- We propose a cross scale attention block which significantly improves the performance of self-completion, and a cross modal transformer block which improves the fusion performance between two modalities.
- Experiments on NYUv2 and SUN RGB-D datasets demonstrate that our method presents better completion results than other state-of-the-art methods.

II. Related Work

Based on the sensor type, the existing depth completion methods can be broadly classified into two categories, the

Corresponding author: Xin Deng.

This work is supported by National Natural Science Foundation of China under Grants 62001016, 62231002, Young Elite Scientists Sponsorship Program 2022QNRC001 by CAST, and Beijing Natural Science Foundation under Grant JQ20020.

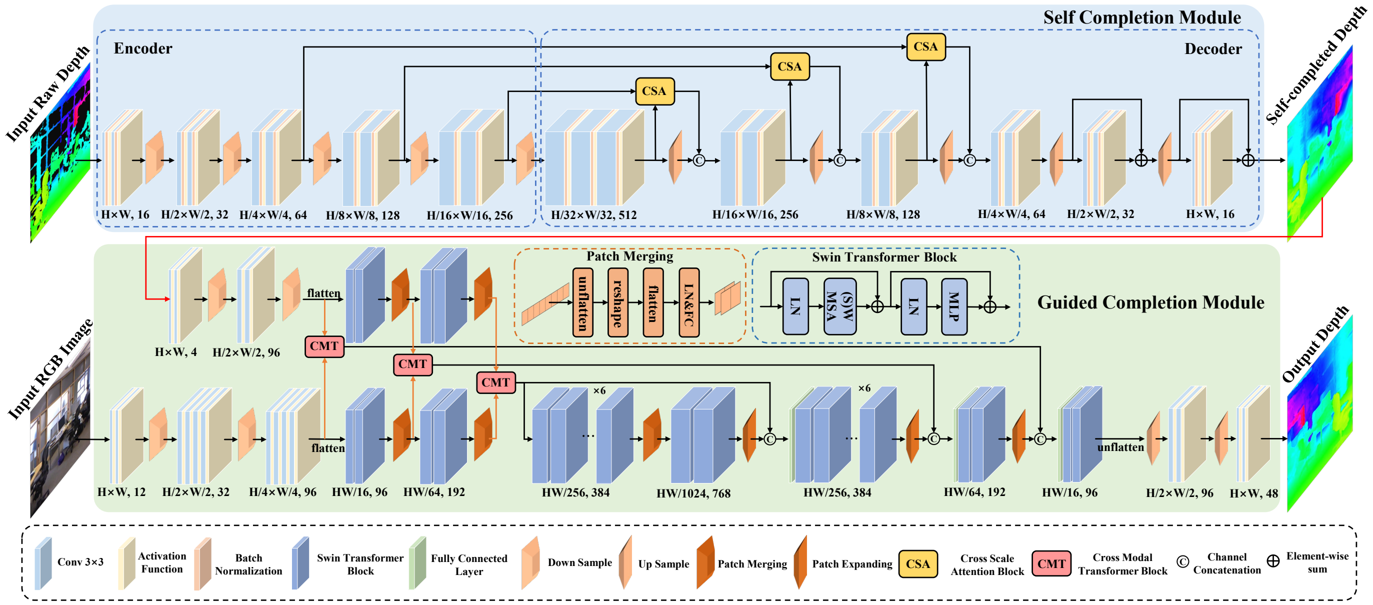

Fig. 1. The overall architecture of the proposed hybrid CNN-Transformer network for RGB guided depth completion. The network is composed of two important modules, including a CNN based self-completion module for coarse completion and a Transformer based guided completion module for finer completion.

optical sensor based indoor semi-dense depth completion and the non-optical sensor based outdoor sparse depth completion.

Sparse Depth Completion. The sparse depth completion aims to generate a dense depth map from a LiDAR captured sparse depth map. The missing regions are usually with sparse and homogeneous distribution in the depth map. Most sparse depth completion methods are based on the encoder-decoder architecture, including single encoder-decoder style [2], [3], [9] , dual-encoder style [10]–[12] and double encoder-decoder style [4], [13], [14] networks. In the single encoder-decoder networks, the RGB and depth images are usually concatenated as one input to be encoded and decoded. The dual-encoder networks encoded the RGB and depth image separately and sent the fused features to decoder to obtain the completed depth image. In the double encoder-decoder networks, the RGB and depth images have their own encoder and decoder, but their features are interactive with each other in the decoder to improve the completion accuracy.

Semi-Dense Depth Completion. Compared to sparse depth completion, the semi-dense depth completion is more challenging, since the missing regions are usually very large and not homogeneous. The traditional methods are usually designed for multi-view conditions [5], [6]. For a single RGB guided semi-dense depth image completion, Zhang *et al.* [7] first proposed a deep neural network to learn the surface normal and occlusion boundary from the RGB image to guide the completion of the depth image. Later, Huang *et al.* [8] introduced boundary consistency for sharper depth structures, and employed self-attention blocks to enhance the completion accuracy. To cope with the large holes, Senushkin *et al.* [1] proposed a modulation branch to consider the statistical difference between the dense and empty regions. Recently, Wang et al. [15] proposed a generative adversarial network (GAN) based method for indoor semi-dense depth completion.

III. Proposed Method

A. Framework

Fig. 1 shows the overall framework of the proposed hybrid CNN-Transformer network for RGB guided depth image completion. The proposed network is composed of two stages, which achieves depth completion in a coarse-to-fine manner. In the first stage, with only the raw depth image as input, we propose a CNN based self-completion module (SCM) with cross scale attention (CSA) block to restore a coarse version of the depth image. This process can be formulated as follows,

$$\boldsymbol{D}_{self} = f_{\text{SCM}}(\boldsymbol{D}_{raw}) \odot (\boldsymbol{I} - \boldsymbol{M}) + \boldsymbol{D}_{raw} \odot \boldsymbol{M}, \quad (1)$$

where f_{SCM} denotes the function of SCM module. The $\boldsymbol{D}_{raw}$ and $\boldsymbol{D}_{self}$ represent the input raw depth image and the completed coarse depth image after SCM module, respectively. The $\odot$ denotes the operation of element-wise multiplication. $\boldsymbol{I}$ is the matrix with all elements as one. The $\boldsymbol{M}$ is a mask generated from $\boldsymbol{D}_{raw}$ by the following rule, which indicates whether the pixel needs to be completed,

$$\boldsymbol{M}(i,j) = \begin{cases} 1, & \boldsymbol{D}_{raw}(i,j) > 0, \\ 0, & \boldsymbol{D}_{raw}(i,j) = 0. \end{cases} \quad (2)$$

where (i, j) represents the position of the pixel. The value 0 indicates that this place should be completed. Then, in the second stage, with the guidance of RGB image, we propose a guided completion module (GCM) with cross-modal

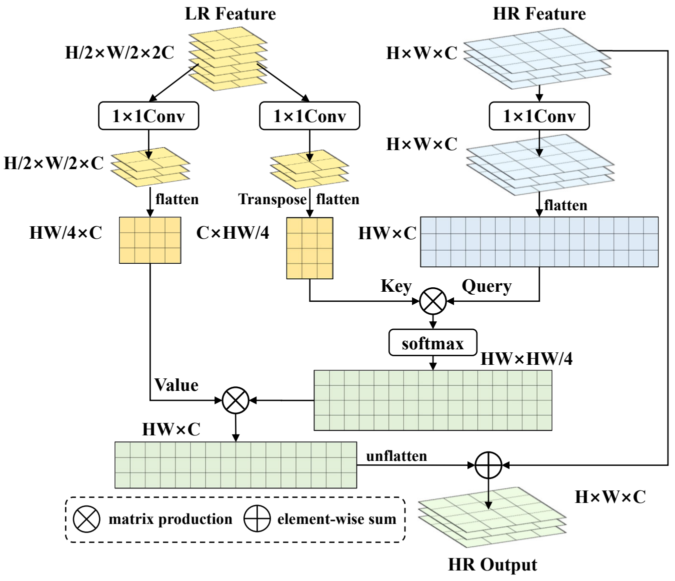

Fig. 2. The structure of the cross scale attention (CSA) bock.

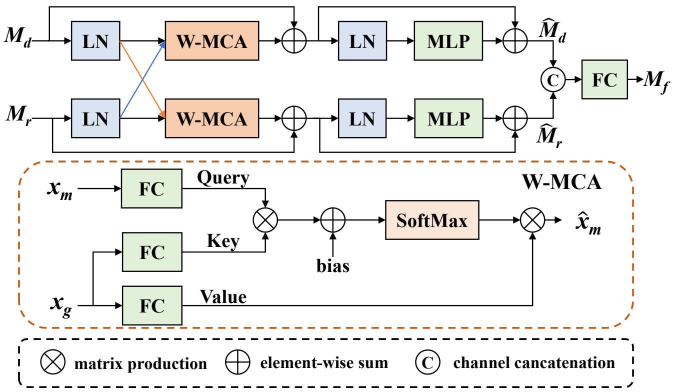

Fig. 3. The structure of cross modal Transformer block. The W-MCA denotes window multi-head cross attention.

Transformer (CMT) block to further refine the completion of the depth image. This process can be formulated as follows,

$$\boldsymbol{D}_{out} = f_{\text{GCM}}(\boldsymbol{D}_{self}, \boldsymbol{R}_{guide}) \odot (\boldsymbol{I} - \boldsymbol{M}) + \boldsymbol{D}_{raw} \odot \boldsymbol{M}, \quad (3)$$

where f_{GCM} denotes the function of GCM module. The $\boldsymbol{R}_{guide}$ is the guided RGB image and $\boldsymbol{D}_{out}$ represents the final completed depth image by the network.

B. Self-Completion Module

The architecture of the SCM module is shown in the upper branch of Fig. 1. In this module, we aim to recover a coarse depth image only from the information of the raw depth image. To achieve this goal, we design an encoder-decoder style network based on U-Net [16]. In the encoder part, we employ a sequence of downsampling convolution layers to generate depth features with different scales. The continuous downsampling operations make it easier for completion since the missing area becomes much smaller. After that, in the decoder part, we employ a sequence of upsampling layers to gradually fill the missing areas from the smallest scale. In the completion process, the cooperation of features from different scales is very important to improve the completion accuracy. Thus, we propose a cross scale attention (CSA) block to correlate the features from different scales, as shown in Fig. 1. Next, we introduce the details of the CSA block.

Cross Scale Attention (CSA) Block. Fig. 2 shows the structure of the CSA block. As can be seen in this figure, the CSA block has two inputs, including one large-scale high-resolution (HR) feature $\boldsymbol{X} \in \mathbb{R}^{H \times W \times C}$ from encoder and one small-scale low-resolution (LR) feature $\boldsymbol{Y} \in \mathbb{R}^{\frac{H}{2} \times \frac{W}{2} \times 2C}$ from decoder. The HR feature $\boldsymbol{X}$ first passes through a 1×1 convolution and then is flattened as the query weight $\boldsymbol{Q}_X \in \mathbb{R}^{HW \times C}$. The LR feature $\boldsymbol{Y}$ passes through two different 1×1 convolutions separately, which align the number of channels the same as HR feature. Then, the LR features are flattened to get the key weight $\boldsymbol{K}_Y \in \mathbb{R}^{\frac{HW}{4} \times C}$ and the value weight $\boldsymbol{V}_Y \in \mathbb{R}^{\frac{HW}{4} \times C}$, separately. Finally, the cross scale attention result can be generated as follows,

$$\text{CSA}(\boldsymbol{X}, \boldsymbol{Y}) = \text{Softmax}(\boldsymbol{Q}_X \boldsymbol{K}_Y^T) \boldsymbol{V}_Y + \boldsymbol{X}. \quad (4)$$

The main idea behind our CSA block is the mutual guidance between the input features with different scales. On one hand, the HR feature from the encoder side has large missing areas, while the LR feature from decoder side is more completed with smaller missing areas. Thus, it can be effective to guide the completion of the HR feature by the LR feature. On the other hand, the HR feature can also guide the upsampling of LR feature for the reason that the HR feature holds more depth information than the LR feature. Therefore, we design the CSA block to calculate the attention between feature maps with different scales. In SCM, we only apply the CSA block in the three deeper features (as shown in Fig. 1), considering the memory cost of attention mechanism.

C. Guided Completion Module

The architecture of our GCM module is shown in the lower branch of Fig. 1. The inputs to the GCM module are the self-completed image $\boldsymbol{D}_{self}$ and the RGB image $\boldsymbol{R}_{guide}$, and the output is the final completed depth image $\boldsymbol{D}_{out}$. In this module, we first extract the shallow features from the depth image $\boldsymbol{D}_{self}$ and RGB image $\boldsymbol{R}_{guide}$ through several convolutional layers separately for each modality. Here, the RGB image needs deeper layers since it has more redundant information to be removed. After that, considering the outstanding ability of global feature extraction of Transformers, we further encode these two modalities by the Swin Transformer blocks [18]. In this process, these two modalities are required to share the same parameters so that more depth related information can be learned from the RGB modality. The downsampling operation is performed between different Swin Transformer blocks to extract features with different scales. In the GCM module, the effective guidance from the RGB modality is important for the depth completion. Thus, we design a cross modal Transformer (CMT) block to fuse the features from

TABLE I

THE DEPTH COMPLETION RESULTS OF DIFFERENT METHODS IN TERMS OF MAE, RMSE, REL, SSIM AND $\delta_{1.05}$ ON NYUV2 AND SUN RGB-D DATASETS. THE BEST RESULTS ARE MARKED IN BOLD AND SECOND BESTS ARE UNDERLINED.

Dataset	NYUv2					SUN RGB-D				
Metric	MAE ↓	RMSE ↓	REL ↓	SSIM ↑	$\delta_{1.05}$ ↑	MAE ↓	RMSE ↓	REL ↓	SSIM ↑	$\delta_{1.05}$ ↑
CSPN [3]	1428.8	4049.6	0.0161	0.9802	0.9186	1526.9	4562.6	0.0272	0.9796	0.9140
S2D [2]	1618.7	4510.5	0.0190	0.9737	0.9052	1575.0	4622.4	0.0275	0.9808	0.9075
Van *et al.* [17]	1120.8	4121.5	0.0115	0.8588	0.9365	1012.0	5074.5	0.0201	0.7890	0.9479
GuideNet [13]	1393.7	3802.0	0.0134	0.8354	0.9193	1088.1	3447.9	0.0188	0.7530	0.9394
DM-LRN [1]	1268.2	3681.1	0.0137	0.9828	0.9249	1088.8	3838.8	**0.0178**	0.9889	0.9480
FCFR-Net [11]	1062.0	3024.6	0.0110	0.9875	0.9388	1069.7	3458.3	0.0206	0.9890	0.9505
Ours	**946.4**	**2773.7**	**0.0094**	**0.9890**	**0.9495**	**939.3**	**3369.2**	0.0191	**0.9912**	**0.9586**

depth and RGB modalities at different scales. In addition to the feature fusion, the CMT block also participates in the decoder for better depth completion, as shown in Fig. 1. Next, we introduce the details about the CMT block.

Cross Modal Transformer (CMT) Block. The structure of CMT Block is shown in Fig. 3. As shown in this figure, the CMT block is composed of two symmetrical and interactive branches. The input to the upper branch is the feature $\boldsymbol{M}_d$ extracted from the depth image, while the input to the lower branch is the feature $\boldsymbol{M}_r$ extracted from the RGB image. The final result is obtained by fusing the outputs of the two branches. Inspired by the shift-window attention mechanism [18], we apply window multi-head cross attention (W-MCA) to correlate the features from depth and RGB modalities. The two modalities $\boldsymbol{M}_d$ and $\boldsymbol{M}_r$ are first fed into the Layernorm (LN) layer, and then we calculate the W-MCA results for the two branches, respectively. For example, in the upper branch where the depth modality is regarded as the main modality, the W-MCA is calculated as follows,

$$\begin{aligned} &\text{W-MCA}(\boldsymbol{x}_m, \boldsymbol{x}_g) = \text{Softmax}(\frac{\boldsymbol{Q}_m \boldsymbol{K}_g^T}{\sqrt{u}} + \boldsymbol{B})\boldsymbol{V}_g, \\ &where \quad \boldsymbol{Q}_m = \boldsymbol{W}^Q \boldsymbol{x}_m, \boldsymbol{K}_g = \boldsymbol{W}^K \boldsymbol{x}_g, \boldsymbol{V}_g = \boldsymbol{W}^V \boldsymbol{x}_g. \end{aligned} \tag{5}$$

In Eq. (5), $\boldsymbol{x}_m = \text{LN}(\boldsymbol{M}_d)$ and $\boldsymbol{x}_g = \text{LN}(\boldsymbol{M}_r)$ represent the features after the LN layer. The $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V$ are the weight matrices of query, key and value. $\boldsymbol{B}$ is the bias and u is the number of heads. The output of the upper branch can be formulated as follows,

$$\begin{aligned} \boldsymbol{M}_d^{'} &= \text{W-MCA}(\boldsymbol{x}_m, \boldsymbol{x}_g) + \boldsymbol{M}_d, \\ \hat{\boldsymbol{M}}_d &= \text{MLP}(\text{LN}(\boldsymbol{M}_d^{'})) + \boldsymbol{M}_d^{'}, \end{aligned} \tag{6}$$

where MLP is the multi-layer perception. In the lower branch where the RGB modality is regarded as the main modality, the W-MCA is calculated with $\boldsymbol{x}_m = LN(\boldsymbol{M}_r)$ and $\boldsymbol{x}_g = LN(\boldsymbol{M}_d)$ by Eq. (5). Then, the output of the lower branch can be formulated as follows,

$$\begin{aligned} \boldsymbol{M}_r^{'} &= \text{W-MCA}(\text{LN}(\boldsymbol{M}_r), \text{LN}(\boldsymbol{M}_d)) + \boldsymbol{M}_r, \\ \hat{\boldsymbol{M}}_r &= \text{MLP}(\text{LN}(\boldsymbol{M}_r^{'})) + \boldsymbol{M}_r^{'}, \end{aligned} \tag{7}$$

The final output of the CMT block is the fusion of $\hat{\boldsymbol{M}}_d$ and $\hat{\boldsymbol{M}}_r$ by fully connected (FC) layer as follows,

$$\boldsymbol{M}_f = \text{FC}([\hat{\boldsymbol{M}}_d, \hat{\boldsymbol{M}}_r]). \tag{8}$$

where $[\cdot]$ indicates the concatenation operation. With the CMT block, we can achieve better fusion of the depth and RGB information, which significantly helps the completion of the depth image.

D. Loss Function

The network is trained in a multi-stage manner. In the first stage, we only train the self-completion module via the following loss function,

$$L_{SCM} = \lambda_1 \|\boldsymbol{D}_{gt} - \boldsymbol{D}_{self}\|_1 + \lambda_2 \sqrt{\|\boldsymbol{D}_{gt} - \boldsymbol{D}_{self}\|_2}, \tag{9}$$

where λ_1 and λ_2 are the weights of the mean absolute error (MAE) and the root mean square error (RMSE), respectively. The $\boldsymbol{D}_{gt}$ represents the ground-truth depth image. In the second stage, we fix the self-completion module to train the guided completion module via the following loss function,

$$L_{GCM} = \lambda_1 \|\boldsymbol{D}_{gt} - \boldsymbol{D}_{out}\|_1 + \lambda_2 \sqrt{\|\boldsymbol{D}_{gt} - \boldsymbol{D}_{out}\|_2}. \tag{10}$$

Finally, the whole network is slightly fine-tuned in an end-to-end manner to yield the trained model.

IV. EXPERIMENTS

A. Experiment Setup

Datasets and Metrics. We conduct experiments on two RGB-D datasets of indoor scenes, including NYUv2 dataset [19] and SUN RGB-D dataset [20]. Following [1], [7], we use RMSE, MAE, structural similarity (SSIM), relative error (REL), and δ_t as the evaluation metrics. The δ_t metric denotes the percentage of pixels where the relative error is less than a threshold t and we set t=1.05 in our experiments. The higher value of SSIM and δ_t and lower value of the other three metrics indicate higher depth completion accuracy.

Implementation Details. The proposed method was implemented by Pytorch and trained with 2 NVIDIA GeForce RTX 3090 GPUs. The ADAM optimizer is adopted with β_1=0.9, β_2=0.99, and the learning rate is 1×10^{-4}. The hyper-parameters in Eq. (9) and Eq. (10) are set as λ_1=1 and λ_2=1. For self completion module, we trained 1,500 and 2,000 epochs on SUN RGB-D and NYUv2 datasets, respectively. For guided completion module, we trained 500 epochs on SUN RGB-D dataset and 1,000 epochs on NYUv2 dataset.

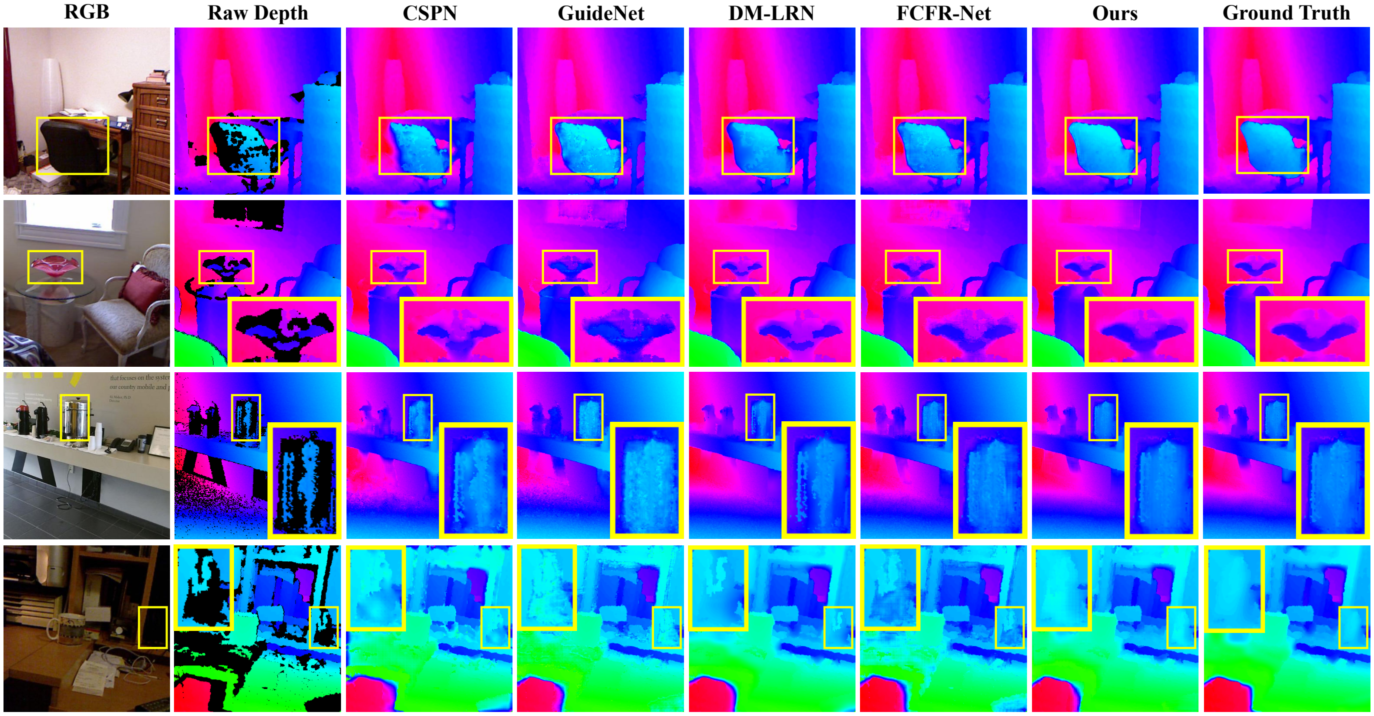

Fig. 4. The visual comparisons of the depth images completed by different methods on NYUv2 ($1^{st}, 2^{nd}$ rows) and SUN RGB-D ($3^{rd}, 4^{th}$ rows) datasets. The yellow blocks indicate the objects with large missing areas.

B. Performance Evaluation

We compare our method with six state-of-the-art methods, including CSPN [3], S2D [2], Van *et al.* [17], GuideNet [13], DM-LRN [1] and FCFR-Net [11]. For all the comparison methods, we trained and evaluated them with the same settings as ours on NYUv2 and SUN RGB-D datasets.

Quantitative Evaluation. Table I compares the performance of our method with the six comparison methods on NYUv2 and SUN RGB-D datasets. As can be seen from this table, our method achieves the highest completion accuracy in terms of all the five metrics on NYUv2 dataset. On the SUN RGB-D dataset, our MAE, RMSE, SSIM and $\delta_{1.05}$ are the best among all methods. These results demonstrate the superior ability of our method in depth completion.

Qualitative Evaluation. Fig. 4 visualizes the completed depth images by our method and other comparison methods on the two datasets. As shown in this figure, our completion result is the closest to the ground-truth, especially on the objects with large missing areas. In contrast, the comparison methods either lead to high frequency noises [11], [13], or overly smooth depth boundaries [1], [3].

C. Ablation Study

The proposed network has two modules, the SCM module for self-completion and the GCM module for guided completion. Thus, we do several ablation studies to investigate the effects of these two modules, together with the CSA block in the SCM module and CMT block in the GCM module. We have carried out four ablation studies denoted by Case 1 to Case 4. In Case 1 and Case 2, we remove the GCM and SCM modules from the network, respectively. In Case 3, we remove the CSA block based on Case 1. In Case 4, we remove the CMT block from the original network. Table II presents the ablation study results in these four cases on both NYUv2 and SUN RGB-D datasets. As can be seen, the completion results become worse in all the four ablation cases, indicating the effectiveness of the corresponding components.

TABLE II
THE RESULTS OF DIFFERENT ABLATION STUDIES.

	NYUv2 Dataset			
Case	MAE ↓	RMSE ↓	REL ↓	$\delta_{1.05}$ ↑
Case1	1211.5	3563.0	0.0126	0.9337
Case2	968.8	2797.0	0.0098	0.9481
Case3	1288.2	3810.8	0.0135	0.9259
Case4	1009.7	3064.4	0.0102	0.9488
Ours	**946.4**	**2773.7**	**0.0094**	**0.9495**
	SUN RGB-D Dataset			
Case	MAE ↓	RMSE ↓	REL ↓	$\delta_{1.05}$ ↑
Case1	1139.7	3753.2	0.0208	0.9422
Case2	985.0	3375.4	0.0209	0.9537
Case3	1260.8	4477.1	0.0231	0.9393
Case4	1022.5	3540.8	0.0210	0.9529
Ours	**939.3**	**3369.2**	**0.0191**	**0.9586**

V. CONCLUSION

In this paper, we proposed a two-stage hybrid CNN-Transformer network for indoor semi-dense depth completion. The proposed network is composed of a self completion modules (SCM) and a guided completion module (GCM). A cross scale attention (CSA) block was developed in SCM to correlate features from different scales to improve the completion accuracy. Additionally, a cross modal transformer (CMT) block was developed in GCM to improve the multi-modal feature fusion performance. The quantitative and qualitative results show that our network outperforms state-of-the-art methods on two datasets with various evaluation metrics.

References

[1] D. Senushkin, M. Romanov, and et al., "Decoder modulation for indoor depth completion," in *2021 IEEE International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 2181–2188.

[2] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *2018 IEEE International Conference on Robotics and Automation*. IEEE, 2018, pp. 4796–4803.

[3] X. Cheng, P. Wang, and et al., "Depth estimation via affinity learned with convolutional spatial propagation network," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 103–119.

[4] S. Zhao, M. Gong, H. Fu, and D. Tao, "Adaptive context-aware multimodal network for depth completion," *IEEE Transactions on Image Processing*, vol. 30, pp. 5264–5276, 2021.

[5] L. Wang, H. Jin, and et al., "Stereoscopic inpainting: Joint color and depth completion from stereo images," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[6] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *IEEE robotics & automation magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[7] Y. Zhang and T. Funkhouser, "Deep depth completion of a single RGB-D image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 175–185.

[8] Y.-K. Huang, T.-H. Wu, Y.-C. Liu, and W. H. Hsu, "Indoor depth completion with boundary consistency and self-attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[9] J. Park, K. Joo, and et al., "Non-local spatial propagation network for depth completion," in *European Conference on Computer Vision*. Springer, 2020, pp. 120–136.

[10] Y. Yang, A. Wong, and et al., "Dense depth posterior (DDP) from single image and sparse range," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3353–3362.

[11] L. Liu, X. Song, and et al., "FCFR-Net: Feature fusion based coarse-to-fine residual learning for depth completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 2136–2144.

[12] H. Wu, G. Zhang, and et al., "SADG-Net: Sparse adaptive dynamic guidance network for depth completion," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 01–06.

[13] J. Tang, F.-P. Tian, W. Feng, J. Li, and P. Tan, "Learning guided convolutional network for depth completion," *IEEE Transactions on Image Processing*, vol. 30, pp. 1116–1129, 2020.

[14] M. Hu, S. Wang, and et al., "PENet: Towards precise and efficient image guided depth completion," in *2021 IEEE International Conference on Robotics and Automation*. IEEE, 2021, pp. 13 656–13 662.

[15] H. Wang, M. Wang, and et al., "RGB-Depth fusion GAN for indoor depth completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6209–6218.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[17] W. Van Gansbeke, D. Neven, B. De Brabandere, and L. Van Gool, "Sparse and noisy lidar completion with rgb guidance and uncertainty," in *2019 16th international conference on machine vision applications (MVA)*. IEEE, 2019, pp. 1–6.

[18] Z. Liu, Y. Lin, and et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[19] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *European Conference on Computer Vision*. Springer, 2012, pp. 746–760.

[20] S. Song and et al., "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 567–576.