# Learning Long-range Information with Dual-Scale Transformers for Indoor Scene Completion

Ziqi Wang[1], Fei Luo * [1], Xiaoxiao Long † [1], Wenxiao Zhang[2], and Chunxia Xiao *[1]

[1]School of Computer Science, Wuhan University
[2]ISTD, Singapore University of Technology and Design

## Abstract

*Due to the limited resolution of 3D sensors and the inevitable mutual occlusion between objects, 3D scans of real scenes are commonly incomplete. Previous scene completion methods struggle to capture long-range spatial context, resulting in unsatisfactory completion results. To alleviate the problem, we propose a novel Dual-Scale Transformer Network (DST-Net) that efficiently utilizes both long-range and short-range spatial context information to improve the quality of 3D scene completion. To reduce the heavy computation cost of extracting long-range features via transformers, DST-Net adopts a self-supervised two-stage completion strategy. In the first stage, we split the input scene into blocks and perform completion on individual blocks. In the second stage, the blocks are merged together as a whole and then further refined to improve completeness. More importantly, we propose a contrastive attention training strategy to encourage the transformers to learn distinguishable features for better scene completion. Experiments on datasets of Matterport3D, ScanNet, and ICL-NUIM demonstrate that our method can generate better completion results, and our method outperforms the state-of-the-art methods quantitatively and qualitatively.*

## 1. Introduction

Indoor 3D reconstruction is an essential part of many applications like AR/VR [19, 1], building information modeling (BIM) [24, 32], automatic robot indoor navigation [14] and so on [36, 38]. Due to the limited resolution of 3D sensors and the inevitable occlusion between objects in the scene, there always exist incomplete surfaces in the reconstruction results.

Recently, scene 3D completion based on deep learning has made great progress. Dai *et al.* [6] proposed a self-supervised method called SG-NN to complete the indoor scene with incomplete real-world scan data, which can get a more complete result compared to the training data. Dai *et al.* [8] further proposed SPSG to complete the 3D surface and texture. Most of the current methods use convolutional neural networks (CNN) for completion and achieve impressive progress. Nonetheless, receptive fields in CNN remain local at a certain resolution, limiting its ability to capture long-range information. This is very important for the scene completion task with incomplete inputs.

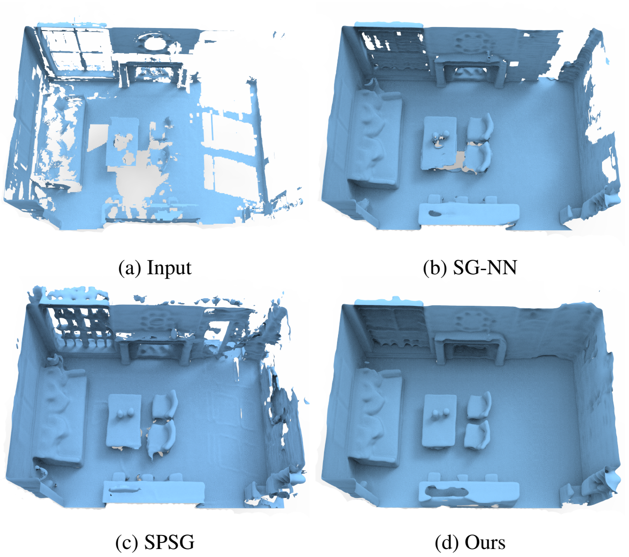

(a) Input
(b) SG-NN
(c) SPSG
(d) Ours

Figure 1: Indoor scene completion on one case in the Matterport3D dataset. Comparison with the state-of-the-art methods SG-NN [6] and SPSG [8].

Recently, Vision Transformer (ViT) [27, 18, 10] has made great stride in the computer vision field, owing to its ability to sense long-range information. For the task of single object 3D completion, some methods [30,

[*]Fei Luo and Chunxia Xiao are co-corresponding authors
[†]Xiaoxiao Long was interning at Wuhan University

33, 34] achieved better point cloud completion by using transformer-based architecture to extract global information with a spatial attention mechanism. However, there is few attempt to apply transformer to the task of large-scale scene completion, since the large-scale scenes contain complex layouts and various objects.

In this work, to use long-range information to handle large missing areas and precise completion of geometric shape in the scene, we propose a Dual-Scale Transformer Network (DST-Net) and apply it in a two-stage manner. Specifically, in the first stage, we use DST-Net to complete the blocks split from an incomplete scene, then merge the complete geometric blocks to get the entire scene output. However, the generated results still have some small holes in local areas. To tackle it, in the second stage, we utilize scene-level information to refine the output scene of stage 1. In this stage, we fuse the ground truth and the first stage output as the supervision signal to focus on learning problematic areas and ensure the consistency of learning.

It is difficult to obtain satisfactory performance when applying transformer for scene completion, as the long-range and sparse incomplete information is more difficult to learn. To solve this problem, our proposed DST-Net involves specific transformer modules for different-level scene information. Furthermore, we propose a contrastive attention training strategy to make the DST-Net efficiently learn similar and distinguishable shape features. Our method includes a structure loss to improve the accuracy of geometric structure, and a CIoT (Cube Intersection over Target) loss to ensure complete output voxels. One comparison case between the state-of-the-art methods and ours is illustrated in Fig. 1.

In summary, our contributions are as follows :

1. We propose a novel Dual-Scale Transformer Network (DST-Net) for indoor scene surface completion. We conduct completion operations from the block level to the scene level to achieve better completion performance.

2. We propose a contrastive attention training strategy to make the transformer work robustly in scene completion. In addition, we propose a structure loss to improve the accuracy of geometric shapes and a CIoT loss to make the scene more complete.

## 2. Related Work

Completing the 3D surface from the 2D or 3D sparse inputs is an ill-posed task. This task greatly relies on the prior knowledge extracted from other similar scenes. We briefly introduce related traditional methods and deep learning-based methods in this field. More extensive reviews can be found in two recent works [23, 16].

### 2.1. Traditional Completion

Due to lacking powerful 3D scene encoding and inferring models, most of the early traditional methods used the interpolating or optimization strategy. Davis *et al.* [9] proposed a diffusion process with the SDF function to fill the holes in the surface. Kawai *et al.* [15] proposed an energy minimization surface completion method, which could complete indoor models with holes. Previtali *et al.* [22] designed a flexible pipeline to perform outdoor and indoor reconstruction from occluded point clouds and proposed corresponding modules for outdoor and indoor scan completion. Silberman *et al.* [25] proposed a probabilistic model of contour completion random fields, which could complete the boundaries of occluded surfaces. Xiao *et al.* [31] integrated point cloud completion and surface connectivity relation inference to obtain complete 3D models and surface connections. The traditional methods depend on certain assumption and their generalization is poor when the prerequisite is not met.

### 2.2. Deep Learning based Completion

**Supervised based completion.** With the increment of 3D data resources [2, 4] and the development of deep learning, more and more methods use deep learning neural network to learn various priors to complete the scene. Some methods [13, 17, 28] focus on depth completion to fulfill surface completion. Surface completion is always combined with semantic segmentation provides both geometry and object label information. Song *et al.* [26] built up a synthetic indoor dataset SUNCG with dense occupancy and semantic annotation, and proposed a fully convolutional end-to-end model to solve both tasks of semantic segmentation and scene completion. Dai *et al.* [7] further proposed a coarse-to-fine fully convolutional model called ScanComplete based on SUNCG to solve semantic segmentation and scene completion tasks for high-resolution indoor scenes. However, it is difficult to have a complete ground truth for 3D indoor scenes due to device and data capturing imperfections.

**Self-supervised based completion.** Self-supervised completion learns the geometric features from incomplete 3D surface ground truth and then expands such learned completion principle to boost the incomplete 3D surface ground truth itself [34, 37, 20, 35]. Dai *et al.* [6] proposed a self-supervised fully convolutional approach SG-NN that can be trained on incomplete, real-world scan data and produced a better geometric model than the training target. SPSG [8] used a similar self-supervised idea to infer a complete scene geometry with color and used 2D data to supervise the color of the output 3D scene for generating a complete color and more accurate scene. It is a promising technical route, as it does not need a large amount of expensive GT data. But it requires the model to have a more robust capacity to observe a wider scope and a stronger completion ability.

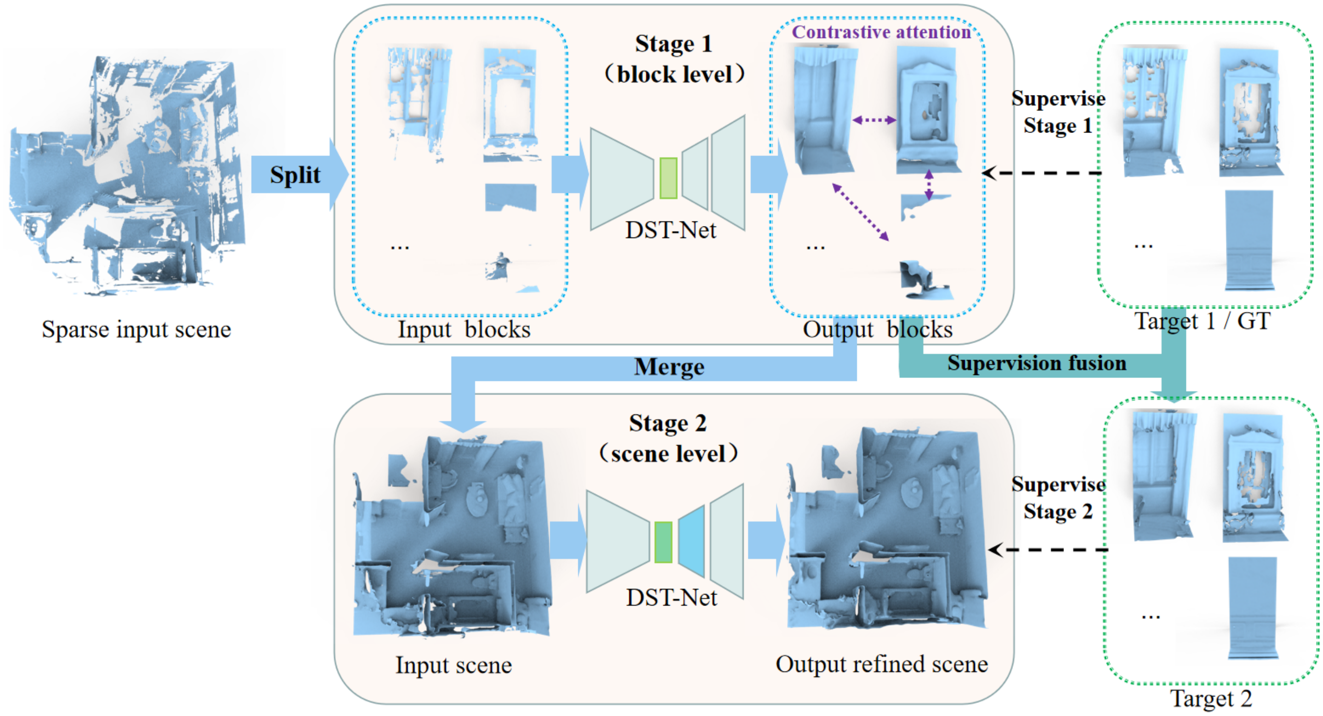


Figure 2: Overview of our completion method. Stage 1 uses DST-Net to complete blocks of the incomplete input, and merge them to get the coarse scene output. Stage 2 proceeds to refine the coarse scene. We use different colors to represent the differences between the two-level network structures in DST-Net.

# 3. Method

## 3.1. Overview

The diagram of our proposed method is illustrated in Fig. 2. Given a training dataset with a series of incomplete indoor scans, our proposed method has two stages. Stage 1 trains DST-Net to learn completion rules on small blocks split from the incomplete scene. Due to memory limitation, in the test, the input and output of stage 1 are same-sized voxel blocks split from a scene, then we merge the output blocks to get the stage 2 input. Then stage 2 aims to address the inadequate inferring result caused by the limited scope of block-level completion by processing the scene-level information, while the training target is obtained by fusing the output and target of stage 1.

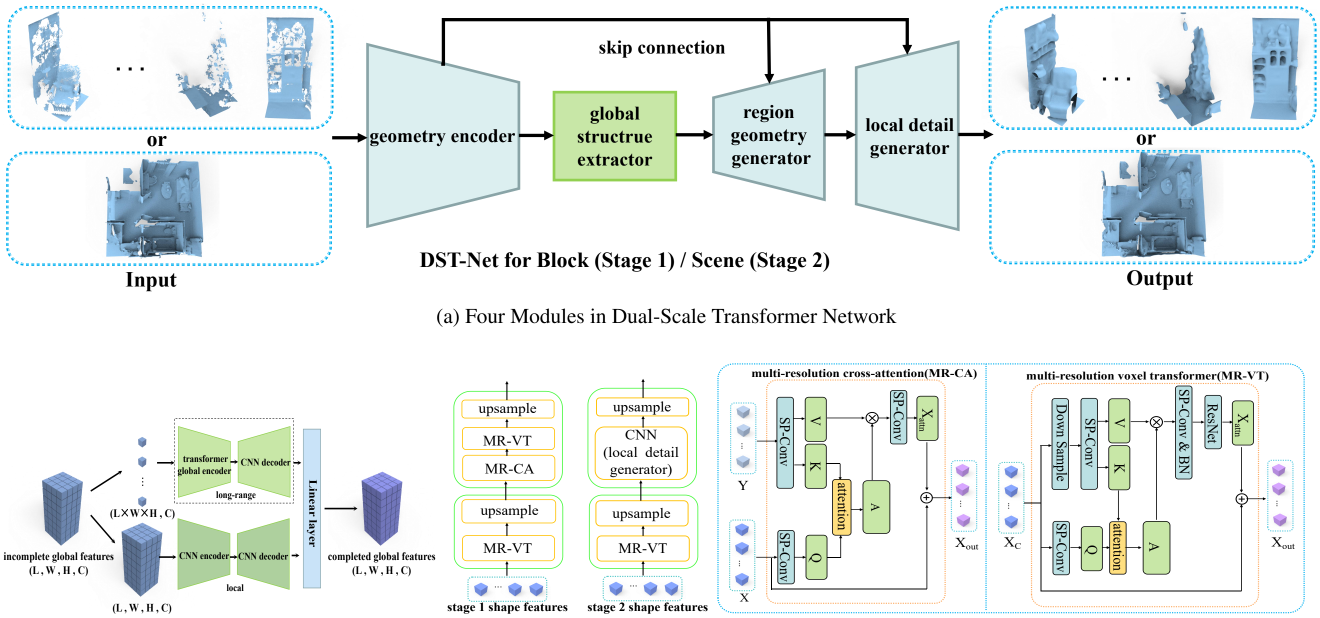
In the training progress, we propose a contrastive attention training strategy that uses the information between different blocks to encourage the transformers to learn distinguishable features for better scene completion. We propose the structure loss and CIoT loss to ensure completed scenes precise and complete.

## 3.2. DST-Net Structure

Our DST-Net is an encoder-decoder structure, as shown in the Fig. 3(a), The DST-Net basically consists of four modules: geometry encoder, global structure extractor, region geometry generator, and local detail generator. The geometry encoder and local detail generator are similar to [6], consisting of 3D Sparse Convolutions [11] ($SP\text{-}Convs$) to encode and decode local geometry features. We propose a global structure extractor to extract and complete global-level features, as well as proposing the region geometry generator to decode region-level geometry shapes with long-range information. Considering the proportion of valid data in the scene, the modules at the two ends of our network process sparse voxels (shown in blue color), and the middle module processes dense voxels (shown in green color). Next, we will introduce the global structure extractor and region geometry generator in detail.

### 3.2.1 Global Structure Extractor

To extract similar high-dimensional features from similar structures, we introduce a global structure extractor. The global structure extractor has two branches, as shown in Fig. 3(b). Inspired by PVT [29], we design a 3D multilayer transformer pyramid structure as one branch to reshape structured voxels into non-sequential voxels and encode them, which mainly focuses on long-range global information. In order not to lose details, we also reserve CNNs as the other branch to encode local information.



(a) Four Modules in Dual-Scale Transformer Network

(b) Global Structure Extractor

(c) Region Geometry Generator

(d) Network structure of MR-CA and MR-VT

Figure 3: Network structure and details. (a) The diagram of our DST-Net structure and its four main modules. The input for stage 1 are blocks and the input for stage 2 is a scene, respectively. Arrows indicate the flow of data in the network. The difference in DST-Net between stage 1 and stage 2 lies in the global structure extractor and region geometry generator. (b) Global structure extractor contains two branches of transformer and CNN. There is no transformer branch (marked with a dot line rectangle) in stage 2. (c) The component of region geometry generator in stage 1 and stage 2. (d) The network structure of two major components in region geometry generator : MR-CA (left column), and MR-VT (right column).

Then the two branches are decoded with CNNs. We use skip-connection to connect the information of the same resolution in both branches. Finally, the global features completed by these two branches are fused by a linear layer. At stage 2, we remove the transformer branch, as the scene is relatively complete, and the need for long-range information capture is reduced.

### 3.2.2 Region Geometry Generator

We propose a region geometry generator to hierarchically generate accurate region shapes with long-range shape similarity. The key components of this module are the multi-resolution cross-attention (MR-CA) and the multi-resolution voxel transformer (MR-VT). The organization ways of region geometry generator in stage 1 and 2 are shown in Fig. 3(c).

**Multi-resolution cross-attention.** We propose the MR-CA module to connect shape features $Y \in \mathbb{R}^{N \times C}$ in the geometry encoder to high-resolution features $X \in \mathbb{R}^{N \times C}$ in the region geometry generator for shape completion, as shown in the left column of Fig. 3(d). $N$ is the number of sparse voxel features, and $C$ is the number of channels. In this module, $Q$ is obtained from $X$ through $1 \times 1$ *SP-Conv*, $K$ and $V$ are obtained from $Y$ through $1 \times 1$ *SP-Conv*. We conduct dot-product between $Q$ and $K^T$ to get the attention map $A$.

$A$ stores the similarity of the shape information between $Y$ and $X$, we use $A$ to select similar shape features from $V$:

$$X_{attn} = SP\text{-}Conv(A \times V) \quad (1)$$

We add the result $X_{attn}$ selected from $Y$ and the input feature $X$ to get the completed feature $X_{out}$ :

$$X_{out} = X + X_{attn} \quad (2)$$

**Multi-resolution voxel transformer.** Unlike MR-CA, which uses the encoder's shape information, MR-VT is proposed to perform region shape completion with the long-range complete shape feature generated in the generator. We first connect the sparse voxel features of the same resolution in the encoder and the decoder, to obtain the feature $X_C \in \mathbb{R}^{N \times C}$. $Q$ is projected by $X_C$. $K$ and $V$ are projected by $X_C$ after down-sampling. The latter attention operation and more details are shown in the right column of Fig. 3(d). By using MR-VT, each voxel with incomplete shape information can capture long-range information from all other generated voxels for completion.

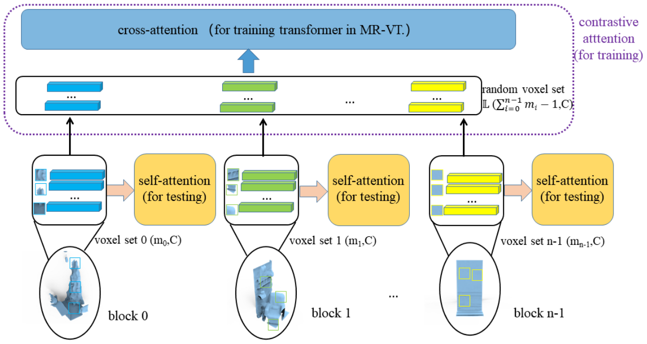

cross-attention (for training transformer in MR-VT.)
contrastive atttention (for training)
random voxel set $\mathbb{L}$ ($\sum_{i=0}^{n-1} m_i - 1$,C)
self-attention (for testing)
self-attention (for testing)
self-attention (for testing)
voxel set 0 ($m_0$,C)
voxel set 1 ($m_1$,C)
voxel set n-1 ($m_{n-1}$,C)
block 0
block 1
block n-1

Figure 4: The diagram of CAT working mechanism.

### 3.3. Contrastive attention training

Contrast learning enhances the performance of vision tasks by contrasting samples against each other, which learns common attributes from similar samples and discriminative attributes from distinguished ones. Inspired by the idea of contrastive learning, we propose a contrastive attention training (CAT) strategy to empower the transformers for scene completion by efficiently learning common and distinguishable shape features from the training samples.

Our method randomly takes a batch containing $b$ blocks as input in training, and the input $X_C$ for MR-VT is a voxel set $\{L_j, j = 0, ..., m_i - 1\}$ processed from each block $i$, where $m_i$ means the number of the voxels in block $i$ and each voxel $L_j$ contains a shape feature within a cubic space. Unlike the general self-attention approach, where the attention operation is done between $\{L_j, j = 0, ..., m_i - 1\}$, we further randomly select the voxels sets from other blocks, and use all voxels in them to form a random voxel set $\mathbb{L} : \{L_j, j = 0, ..., \sum_{i=0}^{n-1} m_i - 1\}$, where $n$ means the number of all the selected blocks. The physical meaning of $\mathbb{L}$ is shown in Fig. 4. In brief, we use M to denote $\sum_{i=0}^{n-1} m_i$. Similar to MR-VT, the $Q_j$, $K_j$ and $V_j$ are projected from $L_j$. In this way, we perform a cross-attention operation between n blocks in a training batch. For a certain voxel $L_j$, we obtain the similarity set $\{A_{j,k}, k = 0, ..., M - 1\}$ between $L_j$ and all M voxels in $\mathbb{L}$ by computing the dot-product of $Q_j$ and $K_k$ and then normalized via softmax function:

$$A_{j,k} = \frac{exp(Q_j \cdot K_k)}{\sum\limits_{k=0}^{M-1} exp(Q_j \cdot K_k)}, \tag{3}$$

After obtaining the similarity set $\{A_{j,k}, k = 0, ..., M - 1\}$, we can use it to select features $V_k$ derived from $L_k$ in $\mathbb{L}$ to complete the voxel $L_j$:

$$X_{attn} = SP\text{-}Conv(\sum_{k=0}^{M-1} A_{j,k} \cdot V_k) \tag{4}$$

The underlying motivation is that unlike self-attention applied to individual blocks with a single layout, our contrastive attention operation requires the transformer to extract useful information from a set of random other blocks. Due to the randomness when selecting blocks in a training batch, the blocks with different layouts may share similar properties or have different attributes, which may be useful or detrimental. By explicitly adding extra information, transformers are required to identify whether the extra information is useful or not by attention scores. The learning strategy enables transformers to learn geometric features with a higher degree of discrimination, thereby improving the completion ability. We evaluate different values for $n$ with 2, 4, and 8. The experimental results indicate that the completion performance reaches its best for $n$ equal to 8.

### 3.4. Loss Function and Data Generation

**Geometry loss.** We train the network with three geometry loss items: depth loss $L_D$, 3D normal loss $L_N$, and structure loss $L_S$. Given a voxel $v$, depth loss $L_D$ is used to calculate the difference between the predicted depth $D_v^P$ and the target depth $D_v^T$ in $v$. The 3D normal loss $L_N$ constrains the predicted normal $N_v^P$ and the target normal $N_v^T$ in $v$. $V$ is the set of voxels to be calculated:

$$L_D = \frac{1}{V} \sum_{v \in V} ||D_v^P - D_v^T||_1 \tag{5}$$

$$L_N = \frac{1}{V} \sum_{v \in V} ||N_v^P - N_v^T||_1 \tag{6}$$

Normal loss is used to constrain the geometry structures by calculating local depth variations. However, for more complex geometry structures, like various concave and convex shapes, it is difficult to complete them well only with normal loss. Inspired by normal loss, we introduce the structure loss $L_S$ to further constrain the geometry structures by calculating local normal variations.

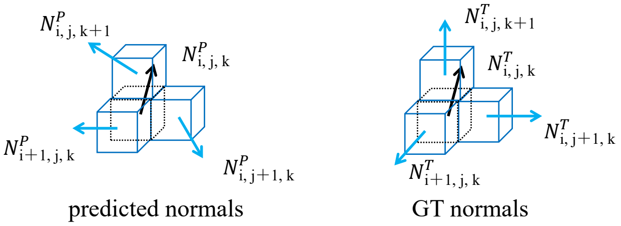


Figure 5: Visual comparison of two different local structures.

As shown in Fig. 5, $N^P_{i,j,k}$ denotes the normal of a predicted voxel, $N^P_{i+1,j,k}$, $N^P_{i,j+1,k}$, and $N^P_{i,j,k+1}$ denote the normals of its three adjacent voxels, where $i, j, k$ represent voxel coordinates in three directions respectively. The GT normals are represented in a similar fashion. In a brief, we use $N_v^P$, $N_{vx}^P$, $N_{vy}^P$ and $N_{vz}^P$ to denote $N_{i,j,k}^P$, $N_{i+1,j,k}^P$, $N_{i,j+1,k}^P$ and $N_{i,j,k+1}^P$, as well as the GT normals. The normal loss of a single predicted voxel is small when calculated with GT, while the local structures of the predicted case and GT are totally different due to the accumulation of normal variations. Therefore, we use the dot-product similarity between $N_v^P$ and $N_{vx}^P$, $N_{vy}^P$, and $N_{vz}^P$ to get $S_{vx}^P$, $S_{vy}^P$, and $S_{vz}^P$ as structure descriptors, as shown in Equation (7), to describe the local variation of normals. The structure descriptors of GT are calculated in the same way. The loss $L_S$ between predicted and target structure descriptors enable the whole predicted geometry structure to represent GT more accurately.

$$\begin{aligned} S_{vx}^P &= N_{vx}^P \cdot N_v^P \\ S_{vy}^P &= N_{vy}^P \cdot N_v^P \\ S_{vz}^P &= N_{vz}^P \cdot N_v^P \end{aligned} \tag{7}$$

$$\begin{aligned} L_S = \frac{1}{V}\sum_{v\in V}\{\frac{1}{3}(||S_{vx}^P - S_{vx}^T||_1 + \\ ||S_{vy}^P - S_{vy}^T||_1 + ||S_{vz}^P - S_{vz}^T||_1)\} \end{aligned} \tag{8}$$

**CIoT loss.** Considering incomplete real-world data, we propose CIoT (Cube Intersection over Target) loss to make output voxel distribution fit the target. Due to the voxel sparsity, the same IoT may represent a completely different distribution. To address it, we divide the entire output into $N$ small cubes of the same size. In a cube, $C_i^P$ and $C_i^T$ represent the number of predicted voxels and the number of target voxels, respectively. We only calculate the loss in the cube where $C_i^T > 0$. CIoT loss is defined as:

$$L_{CIoT} = 1 - \frac{1}{N}\sum_{i=1}^{N}\frac{C_i^P \cap C_i^T}{C_i^T} \tag{9}$$

**Data Generation.** Both input and target in training are the representation of sparse TSDF voxel with depth information, obtained from RGB-D frames by the method of voxel fusion [3]. Following the self-supervised learning approach of SG-NN [6], the training target (GT) in stage 1 is generated using all RGB-D frames, while a certain proportion is used in getting input. The training target in stage 2 is obtained by fusing the output and target of stage 1. To focus on learning problem regions and minimize error accumulation, we select the TSDF value of target 1 voxel as the value for target 2 voxel when the coordinates of output and target of stage 1 coincide.

# 4. Experiments and Results

To validate our proposed method, we use the Matterport3D [2] dataset as the training data and conduct the comparison experiments on both real-world data and synthetic data. The training dataset includes 1788 rooms. We train our method on a single NVIDIA GeForce RTX 2080. The learning rate is 0.001, and the batch size is 8. The resolution of all voxels is 2 cm. The training blocks are all 128 × 64 × 64 size cutting from the scene, and parameter $n$ in contrastive attention training strategy is 8. The four terms in the loss function have equal weights. Our stage 1 model training takes about 72 hours, and the stage 2 model training takes about 24 hours. Three state-of-the-art methods of ConvOccNet [21], SG-NN [6] and SPSG (only with geometry) [8] are used to compare with ours on the metrics of Chamfer Distance (CD) in metric space, Recall, and Precision. For real-world unobserved space, we ignore it for the CD evaluation.

| Method | CD($\times 10^{-1}$)↓ | Recall↑ | Precision↑ |
|---|---|---|---|
| ConvOccNet | 1.48 | 0.51 | 0.54 |
| SG-NN | 0.65 | 0.69 | **0.62** |
| SPSG | 0.35 | 0.74 | 0.53 |
| Ours | **0.23** | **0.78** | 0.61 |

Table 1: Quantitative comparison results on Matterport3D.

## 4.1. Completion on Matterport3D

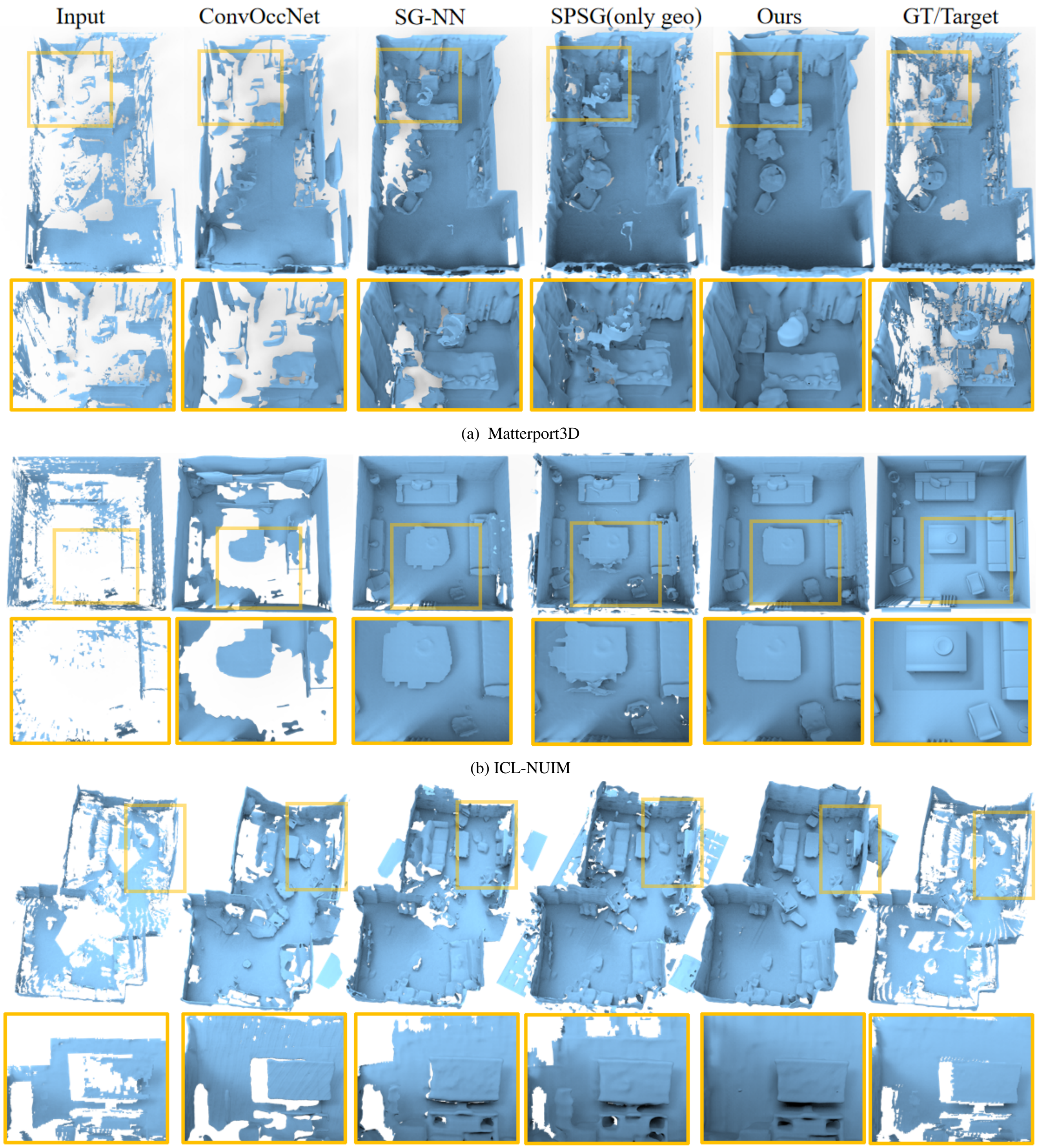
First, we compare four methods on 394 other rooms from the Matterport3D dataset. We provide the same incomplete TSDF input, except for ConvOccNet, which requires the corresponding mesh to be pre-computed as input. Some scene completion results are shown in Fig. 6(a). In visual effect, our method completes more parts in the scenes than the other three methods, which is consistent with the recall metric in Tab. 1. The highest score on CD proves that our method's shape accuracy is the best. As the GT is incomplete and the more completion areas from our method have no corresponding GT, our precision is less than the SG-NN under this special condition.

## 4.2. Completion on Synthetic Data

The synthetic dataset ICL-NUIM [12] has 3D surface ground truth so that it can provide a more comprehensive evaluation. As the scan frames are indoors, we only select the mesh of the surface inside the room to evaluate. Qualitative and quantitative evaluations are shown in Fig. 6(b) and Tab. 2. Owing to the full GT, the results fairly reflect our method's superiority over the other three methods.

## 4.3. Completion on ScanNet

To validate the generalization of our proposed method, we also conduct the comparison experiments on a more challenging dataset ScanNet [5]. The Fig. 6(c) are the visualization results, and Tab. 3 shows the quantitative results.



(a) Matterport3D

(b) ICL-NUIM

(c) ScanNet

Figure 6: Qualitative comparisons with state-of-the-art methods on Matterport3D, ICL-NUIM and ScanNet.

### 4.4. Ablation Study

In this section, we demonstrate how the functional components in our method affect the ultimate performance.

**Stage 1 versus Stage 2.** Stage 1 in our full model has finished most of the completion. However, stage 1 may generate some inaccurate areas due to block-level scope. Stage 2 can further solve these problems by training with

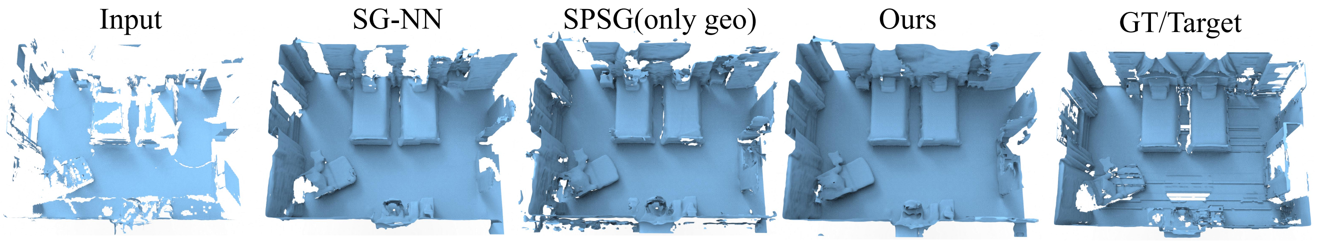


Figure 7: Qualitative comparisons with state-of-the-art methods on Matterport3D with 30% frames input.

| Method | CD($\times10^{-1}$)↓ | Recall↑ | Precision↑ |
|---|---|---|---|
| ConvOccNet | 0.61 | 0.17 | 0.28 |
| SG-NN | 0.31 | 0.31 | 0.42 |
| SPSG | 0.25 | 0.35 | 0.39 |
| Ours | **0.18** | **0.39** | **0.45** |

Table 2: Quantitative comparison results on ICL-NUIM.

| Method | CD($\times10^{-1}$)↓ | Recall↑ | Precision↑ |
|---|---|---|---|
| ConvOccNet | 0.65 | 0.79 | 0.43 |
| SG-NN | 0.36 | 0.85 | **0.50** |
| SPSG | 0.30 | 0.86 | 0.45 |
| Ours | **0.26** | **0.88** | 0.48 |

Table 3: Quantitative comparison results on ScanNet.

the fused target 2 and using the whole scan as input. The qualitative and quantitative evaluation results of stage 1 and the full model are shown in the Fig. 8 and Tab. 4. The results show that stage 2 accumulates a little distance error, but completes and corrects the results of stage 1. Tab. 4 further shows the quantitative results of training the full model with GT and using only the stage 2 model that is trained with stage 1 data to process input. The results indicate that the fused target 2 works better than GT, and the two stages are necessary.

| Method | CD($\times10^{-1}$)↓ | Recall↑ | Precision↑ |
|---|---|---|---|
| Stage1 | **0.22** | 0.77 | 0.60 |
| Full model | 0.23 | **0.78** | **0.61** |
| Full model (train with only GT) | 0.46 | 0.72 | 0.60 |
| Only Stage2 | 0.49 | 0.72 | 0.57 |

Table 4: The improvement of the full model over stage 1, the fused target 2 over GT. And the necessity of stage 1 .

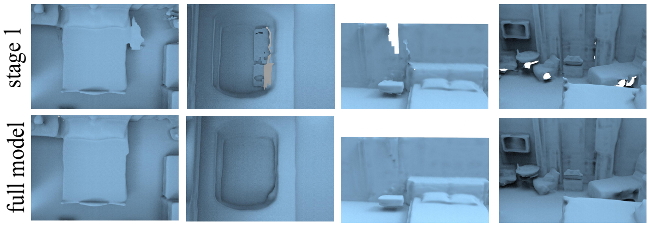


Figure 8: Completion improvement in the full model made by stage 2 over stage 1.

| Method | CD($\times10^{-1}$)↓ | Recall↑ | Precision↑ |
|---|---|---|---|
| SG-NN (40%) | 0.67 | 0.68 | **0.61** |
| SPSG (40%) | 0.36 | 0.73 | 0.53 |
| Ours (40%) | **0.23** | **0.77** | 0.60 |
| SG-NN (30%) | 0.73 | 0.67 | **0.61** |
| SPSG (30%) | 0.43 | 0.72 | 0.52 |
| Ours (30%) | **0.24** | **0.77** | 0.60 |
| SG-NN (20%) | 1.06 | 0.60 | **0.58** |
| SPSG (20%) | 0.73 | 0.65 | 0.49 |
| Ours (20%) | **0.47** | **0.72** | 0.56 |

Table 5: Quantitative comparison results on Matterport3D with 30% frames input.

**More incomplete inputs.** To illustrate the long-range advantage of our method, we use the model trained on the input consisting of 50% proportion RGB-D frames to test on the 40%, 30%, and 20% proportion input. The qualitative and quantitative evaluation results are shown in Fig. 7 and Tab. 5. The results show that as the input information gradually decreases, our method can also output a relatively complete result.

**The impact of contrastive attention training.** We validate the contribution of the contrastive attention training strategy (CAT) for training at stage 1. Its qualitative and quantitative evaluation results are shown in Fig. 9 and Tab. 6. We can find that CAT enhances the ability to restore regional shapes.

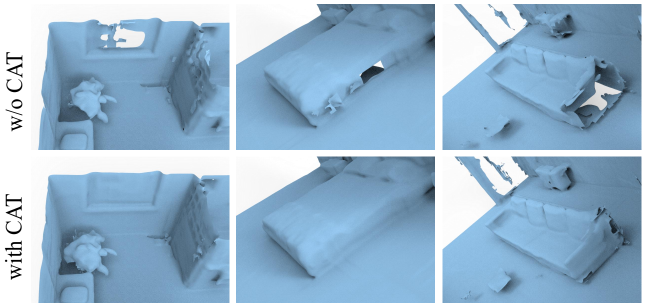


Figure 9: The visual effect of CAT on shape completion.

**The effectiveness of modules.** We replace the global structure extractor (GSE) and region geometry generator (RGG) with CNNs to evaluate their effectiveness. The Tab. 6 shows the quantitative evaluation, and Fig. 10 dis-

| Method | CD($\times 10^{-1}$)↓ | Recall↑ | Precision↑ |
|---|---|---|---|
| Stage1 | **0.22** | **0.77** | **0.60** |
| Stage1(w/o GSE) | 0.32 | 0.74 | 0.57 |
| Stage1(w/o RGG) | 0.26 | 0.75 | 0.58 |
| Stage1(w/o CAT) | 0.24 | 0.75 | 0.59 |

Table 6: Quantitative evaluation of our proposed modules.

| Method | CD($\times 10^{-1}$)↓ | Recall↑ | Precision↑ |
|---|---|---|---|
| w/o L$_{CIoT}$ | 0.25 | 0.75 | 0.60 |
| w/o L$_S$ | 0.26 | 0.75 | 0.56 |
| w/o L$_N$ | 0.26 | 0.74 | 0.56 |
| w/o L$_D$ | 0.29 | 0.75 | 0.54 |
| full loss | **0.22** | **0.77** | **0.60** |

Table 7: Ablation results of loss items in stage 1.

plays how GSE and RGG improve the completion effect in terms of global structure and region shape, respectively.

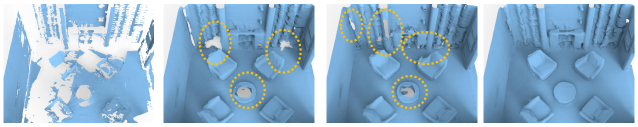

stage 1 input stage 1 w/o GSE stage 1 w/o RGG stage 1 output

Figure 10: Qualitative comparisons of how the completion results are affected by RGG and GSE.

**The impact of each loss item.** We validate the effect of $L_{CIoT}$, $L_S$, $L_N$ loss on our method, and validate the case where only $L_D$ loss is applied. Tab. 7 shows the quantitative evaluation of the loss items. Fig. 11 shows the qualitative evaluation of the loss items. We can see that with $L_{CIoT}$, the completeness of local scene completion has been improved, while $L_S$ and $L_N$ effectively constrain the accuracy of the shapes and planes, respectively. From the comparison with the results of only using $L_D$ loss, we can find that various losses acting on the connection between voxels improve the completion accuracy and completeness.

## 5. Conclusion

In this work, we have proposed a dual-scale transformer method with a two-stage completion strategy to generate fairly complete scenes from real-world incomplete RGB-D scans. Our proposed contrastive attention training strategy encourages the transformers to learn distinguishable features. We introduce two new losses to constrain geometric accuracy and completeness. However, due to the limitation of memory, currently, we only use geometric information and solve the geometric completion. In the future, we would consider scene completion with other priors such as color and semantic information.

## 6. Limitation

First, our method only completes the geometric surface, without including the texture. In some conditions, the texture is also necessary. In the near future, we will try to solve it. Second, surface completion mainly provides reasonable inference for the missing parts, rather than replacing the accurate reconstruction. When the geometric information of the input scene is not enough to infer the entire scene, our geometry completion result may not be accurate. Although our method does not need the GT of the complete scene, it still needs a large amount of indoor data to ensure its generalization.

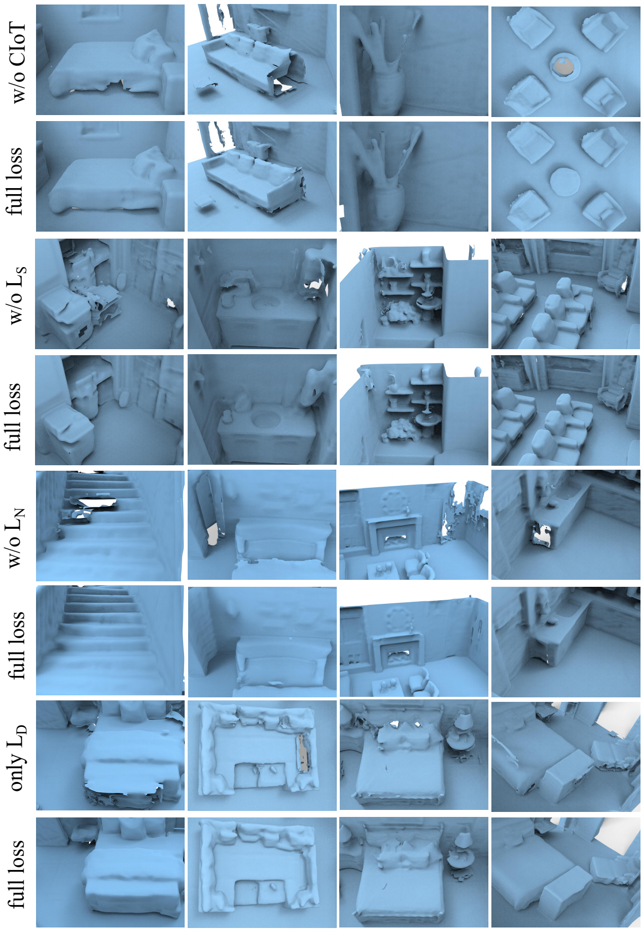


Figure 11: Effect of loss items by switching on/off them at stage 1.

## Acknowledgments

Chunxia Xiao is partially supported by NSFC (No.61972298), Bingtuan Science and Technology Program (No.2019BC008), and CAAI-Huawei MindSpore Open Fund.

## References

[1] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022. 1

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 2, 6

[3] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In John Fujii, editor, *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 303–312. ACM, 1996. 6

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2432–2443. IEEE Computer Society, 2017. 6

[6] Angela Dai, Christian Diller, and Matthias Nießner. SG-NN: sparse generative neural networks for self-supervised scene completion of RGB-D scans. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 846–855. Computer Vision Foundation / IEEE, 2020. 1, 2, 3, 6

[7] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4578–4587. Computer Vision Foundation / IEEE Computer Society, 2018. 2

[8] Angela Dai, Yawar Siddiqui, Justus Thies, Julien Valentin, and Matthias Nießner. SPSG: self-supervised photometric scene generation from RGB-D scans. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1747–1756. Computer Vision Foundation / IEEE, 2021. 1, 2, 6

[9] James Davis, Stephen R Marschner, Matt Garr, and Marc Levoy. Filling holes in complex surfaces using volumetric diffusion. In *Proceedings. First international symposium on 3d data processing visualization and transmission*, pages 428–441. IEEE, 2002. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[11] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *CoRR*, abs/1706.01307, 2017. 3

[12] A. Handa, T. Whelan, J.B. McDonald, and A.J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Hong Kong, China, May 2014. 6

[13] Yu-Kai Huang, Tsung-Han Wu, Yueh-Cheng Liu, and Winston H Hsu. Indoor depth completion with boundary consistency and self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[14] Hasan Ismail, Rohit Roy, Long-Jye Sheu, Wei-Hua Chieng, and Li-Chuan Tang. Exploration-based slam (e-slam) for the indoor mobile robot using lidar. *Sensors*, 22(4):1689, 2022. 1

[15] Norihiko Kawai, Avideh Zakhor, Tomokazu Sato, and Naokazu Yokoya. Surface completion of shape and texture based on energy minimization. In *2011 18th IEEE International Conference on Image Processing*, pages 897–900. IEEE, 2011. 2

[16] Jianwei Li, Wei Gao, Yihong Wu, Yangdong Liu, and Yanfei Shen. High-quality indoor scene 3d reconstruction with rgb-d cameras: A brief review. *Computational Visual Media*, pages 1–25, 2022. 2

[17] Hongmin Liu, Xincheng Tang, and Shuhan Shen. Depth-map completion for large indoor scene reconstruction. *Pattern Recognition*, 99:107112, 2020. 2

[18] 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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021. 1

[19] Alessandro Manni, Damiano Oriti, Andrea Sanna, Francesco De Pace, and Federico Manuri. Snap2cad: 3d indoor environment reconstruction for ar/vr applications using a smartphone device. *Computers & Graphics*, 100:116–124, 2021. 1

[20] Liang Pan, Xinyi Chen, Zhongang Cai, Junzhe Zhang, Haiyu Zhao, Shuai Yi, and Ziwei Liu. Variational relational point completion network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8524–8533, 2021. 2

[21] Songyou Peng, Michael Niemeyer, Lars M. Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 523–540. Springer, 2020. 6

[22] Mattia Previtali, Marco Scaioni, Luigi Barazzetti, and Raffaella Brumana. A flexible methodology for outdoor/indoor building reconstruction from occluded point clouds. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2(3):119, 2014. 2

[23] Luis Roldao, Raoul De Charette, and Anne Verroust-Blondet. 3d semantic scene completion: a survey. *International Journal of Computer Vision*, pages 1–28, 2022. 2

[24] Majid Seydgar, Ali Motamedi, and Erik Poirier. Deep neural networks to assist in bim creation using scanned data: A review. *Transforming Construction with Reality Capture Technologies*, 2022. 1

[25] Nathan Silberman, Lior Shapira, Ran Gal, and Pushmeet Kohli. A contour completion model for augmenting surface reconstructions. In *European Conference on Computer Vision*, pages 488–503. Springer, 2014. 2

[26] Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas A. Funkhouser. Semantic scene completion from a single depth image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 190–198. IEEE Computer Society, 2017. 2

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. 1

[28] Haowen Wang, Mingyuan Wang, Zhengping Che, Zhiyuan Xu, Xiuquan Qiao, Mengshi Qi, Feifei Feng, and Jian Tang. Rgb-depth fusion gan for indoor depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6209–6218, 2022. 2

[29] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 548–558. IEEE, 2021. 3

[30] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5479–5489. IEEE, 2021. 2

[31] Yong Xiao, Yuichi Taguchi, and Vineet R Kamat. Coupling point cloud completion and surface connectivity relation inference for 3d modeling of indoor building environments. *Journal of Computing in Civil Engineering*, 32(5):04018033, 2018. 2

[32] Mahdi Yazdanpour, Guoliang Fan, and Weihua Sheng. Manhattanfusion: Online dense reconstruction of indoor scenes from depth sequences. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 1

[33] Cheng Zhang, Haocheng Wan, Xinyi Shen, and Zizhao Wu. Patchformer: An efficient point transformer with patch attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11789–11798. IEEE, 2022. 2

[34] Wenxiao Zhang, Zhen Dong, Jun Liu, Qingan Yan, Chunxia Xiao, et al. Point cloud completion via skeleton-detail transformer. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 2

[35] Wenxiao Zhang, Chengjiang Long, Qingan Yan, Alix LH Chow, and Chunxia Xiao. Multi-stage point completion network with critical set supervision. *Computer Aided Geometric Design*, 82:101925, 2020. 2

[36] Wenxiao Zhang and Chunxia Xiao. Pcan: 3d attention map learning using contextual information for point cloud based retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12436–12445, 2019. 1

[37] Wenxiao Zhang, Qingan Yan, and Chunxia Xiao. Detail preserved point cloud completion via separated feature aggregation. In *European Conference on Computer Vision*, pages 512–528. Springer, 2020. 2

[38] Wenxiao Zhang, Huajian Zhou, Zhen Dong, Qingan Yan, and Chunxia Xiao. Rank-pointretrieval: Reranking point cloud retrieval via a visually consistent registration evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 1