




Easing 3D Pattern Reasoning with Side-view Features for Semantic Scene Completion

Linxi Huan¹, Mingyue Dong¹, Linwei Yue², Shuhan Shen³, and Xianwei Zheng¹

¹ The State Key Lab. LIESMARS, Wuhan University

² School of Geography and Information Engineering, China University of Geosciences

³ Institute of Automation, Chinese Academy of Sciences

Abstract. This paper proposes a side-view context inpainting strategy (SidePaint) to ease the reasoning of unknown 3D patterns for semantic scene completion. Based on the observation that the learning burden on pattern completion increases with spatial complexity and feature sparsity, the SidePaint strategy is designed to decompose the complex 3D pattern learning into easier 2D context inpainting with dense feature volumes. Specifically, our approach densely lifts image features into 3D volume space with distance-aware projection, and reasons missing patterns in 2D side-view feature maps sliced from feature volumes. With the learning burden relieved by decreasing pattern complexity in 2D space, our SidePaint strategy enables more effective semantic completion than directly learning 3D patterns. Extensive experiments demonstrate the effectiveness of our SidePaint strategy on several challenging semantic scene completion benchmarks.

Keywords: Semantic scene completion · 3D pattern reasoning · Side-view feature learning

1 Introduction

Semantic scene completion (SSC) requires to recover complete 3D scenes from partially captured surface information by semantic volumetric occupancy inference. Pioneering works leveraged volumetric depth features for 3D pattern completion [36, 40, 41], while later efforts were devoted to exploiting RGB-D data for SSC by two-stream deep models with delicate fusion mechanisms [14, 23, 39]. With the utilization of complementary multi-modality clues, the performance of SSC has greatly advanced over the years.

Despite the progress achieved with extensive research, semantic scene completion is still hindered by the ill-posed problem of inferring unobserved areas with partial observations. To complete the complex 3D spatial relationships, existing works tend to adopt a head-on but clumsy solution of expanding model capacity with bulky multi-modality network architectures [5, 23, 24]. However, the reasoning difficulty brought by the naturally high complexity of 3D context patterns is seldom truly addressed, and the signal sparsity of 3D feature volumes

derived by commonly-used depth-guided 2D-3D projection further toughens the completion of dense 3D context.

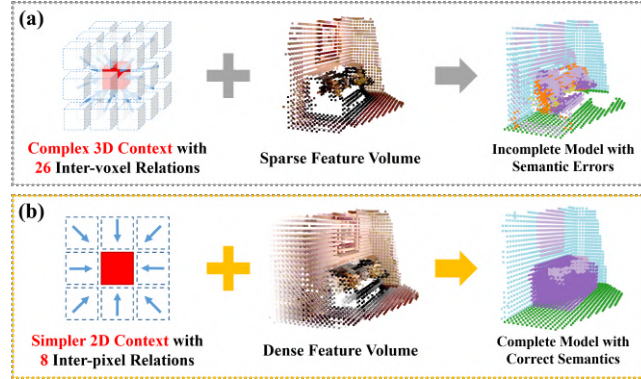


Fig. 1: Comparison between directly completing 3D context with sparse signals and our SidePaint strategy. (a) Reasoning complex 3D semantic patterns from a sparse feature volume. (b) SidePaint strategy models a semantic scene by reasoning simpler 2D context with a denser feature volume.

As illustrated in Figure 1, compared to the 2D image, the spatial relations are more sophisticated in 3D space. When only considering the neighbors of a sample in a 3D or 2D grid, there are at most 26 different inter-voxel relations for a 3D voxel, but only 8 inter-pixel ones for a 2D pixel. The 3D context becomes more complicated w.r.t long-range relationships, inevitably resulting in the increasing reliance on highly complicated deep SSC frameworks for more powerful learning ability. Due to the relation complexity in 3D space, semantic completion errors easily occur especially with sparse volumetric features as shown by Figure 1 (a).

Motivated by the observation above, we introduce a side-view context inpainting strategy (SidePaint) to ease the reasoning burden for SSC. The SidePaint strategy is proposed to decompose the challenging 3D pattern completion to simpler side-view 2D context inpainting, where spatial relation complexity is reduced with a lower dimension. Concretely, our SidePaint strategy works by regarding a feature volume as groups of 2D slices and propagating features to fill undetected context in each 2D slice w.r.t three different side views. To guarantee sufficient information in each slice for side-view inpainting, the SidePaint strategy generates dense volumetric features with a distance-aware 2D-3D projection (Figure 1 (b)). In this case, occluded areas are initialized according to their distance to observed object surfaces instead of being kept empty as in prior works. With SidePaint strategy, we build a single-stream deep model that is free of complicated dual-branch multi-modality learning design for effective SSC.

The main contributions are summarized as follows:

- We introduce a side-view context inpainting (SidePaint) strategy to ease the 3D pattern learning burden by reasoning the simpler spatial relations in 2D space with dense volumetric features.
- We build a single-stream framework with the SidePaint strategy to facilitate semantic scene completion.
- Extensive experiments demonstrate the effectiveness of our SidePaint strategy, and our SidePaint-based model achieves state-of-the-art performance on three challenging semantic scene completion benchmarks.

2 Related Works

2.1 Semantic Scene Completion

Semantic scene completion is introduced by [36] to unify geometry completion and semantic parsing into a problem of single-view dense voxel-wise labelling. [36] pioneer to propose an end-to-end deep solution (SSCNet) to model semantic 3D scenes with depth features embedded by modified truncated signed distance function (TSDF), which is also adopted in followers for supporting SSC with already known geometry priors [6,15,41]. The depth-only SSC is studied with not only TSDF but also other geometry representations (*e.g.*, point cloud) [7,42,48], but these methods stumble at inferring large-area unknown semantic context. Increasing attention is thus turned to the joint utilization of the semantic cues in RGB images and the geometric information in depth data for SSC.

Typical multi-modality approaches mainly work under dual-branch deep frameworks with different feature fusion schemes and utilize the depth information for 2D-3D feature projection. These methods generally first derive 2D features from data of different modalities with two 2D feature extractors and then use depth-guided 2D-3D projection to obtain two sparse 3D feature volumes for subsequent cross-voxel 3D context propagation [24–27]. To leverage real 3D geometries rather than the 2.5D information in depth images, some works choose to extract observed geometric cues from TSDF for later fusion with projected volumetric RGB-based features [5,22,28]; while others deploy additional 3D object detectors to refine coarse completion results with instance-level clues [2,13]. With the focus on effective semantic guidance, works like SATNet [29], SPAwN [11], and FFNet [39] leverage RGB-D semantic segmentation algorithms to provide explicit semantic priors for dense voxel-wise semantic inference. Promoted by the success gained in single-view RGB-D SSC, the exploration of SSC has also been extended to more data types, including monocular RGB images [3], scans [8], and outdoor lidar data [1,34].

Some Tri-plane feature learning methods also leverage 2D clues for 3D learning [9,19]. However, the 2D clues are taken as 3D spatial context information that needs to be processed by cross-view hybrid networks for complex 3D semantic pattern recovery. In contrast to prior arts that developed complex deep network models for extracting and merging multi-view or multi-modality features, this paper aims to tackle the learning difficulty brought by the high complexity of 3D context patterns for SCC. To this end, we introduce the side-view context

painting strategy to decompose complex 3D semantic patterns into simpler 2D relations for easier pattern completion.

2.2 Complete 3D Reconstruction

Complete 3D Reconstruction has been studied with the purpose to recover intact 3D models with partial observations. The complete 3D models are generally recovered without semantics [4, 12, 43], but increasing research attention has been witnessed recently in building scene models with semantics in an end-to-end manner.

For semantic-aware complete scene reconstruction, one popular way works by the SSC framework mentioned above, while the other one relies on multi-task systems. Different from the SSC framework that unifies geometry completion and semantic understanding, multi-task reconstructors factorize the semantic reconstruction into several sub-tasks, such as object detection and scene-aligned instance shape completion [31, 38]. The multi-task system is intuitively designed to align CAD shapes to real scene space based on the information offered by object bounding boxes or semantic segments derived from images [18, 20]. To be free of the dependence on an off-line CAD pool, current studies prefer to insert an instance shape completion branch into their deep end-to-end multi-task networks [17, 21, 44]. Among these approaches, some systems are constructed for scene-aligned instance completion only [16, 32, 37], while some are built with additional consideration of the layout estimation [10].

Whether it is the single-task SSC framework or the multi-task semantic-aware reconstruction system, each of the two solutions has its own merits for different application requirements. The target of this paper is to realize SSC with a single-task deep learning method, and we concentrate on addressing the fundamental issue of 3D context reasoning burden in semantic scene completion.

3 Method

3.1 Overview

In Figure 2, we present the overview of a semantic scene completion system deployed with the proposed side-view context inpainting strategy (SidePaint). Given the 2D features extracted from an image, distance-aware projection is first adopted to convert the 2D clues into a dense 3D feature volume. The 3D features are next fed into a feature enrichment module, and then a side-view context inpainter is applied to the enriched features for reasoning the unknown semantic patterns in 2D space. With the inpainted features, voxel-wise semantics are predicted by a linear classifier at the end. In the following, we will elaborate on the two components of SidePaint: the distance-aware 2D-3D projection and the side-view context inpainter.

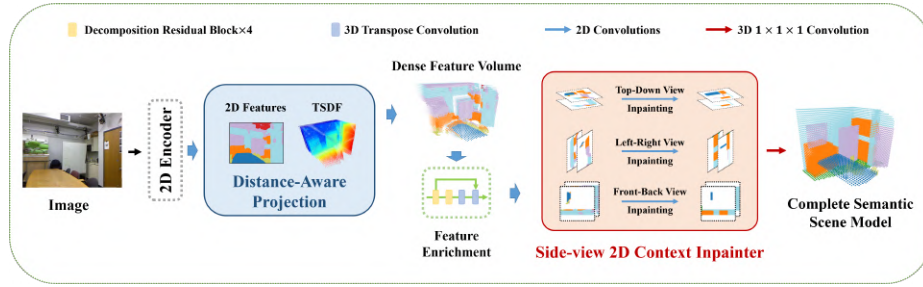


Fig. 2: Overview of our SidePaint-based SSC framework.

3.2 Distance-aware 2D-3D Projection

The common depth-guided 2D-3D feature projection leaves unobserved area voxels empty and thus results in highly sparse 3D feature volumes, where most 2D slices hardly contain information for effective side-view context completion (Figure 3 (a)). Therefore, our SidePaint strategy first adopts a distance-aware 2D-3D projection to provide dense feature volumes by computing the feature vector f_v for any given voxel v according to Equation (1).

$$f_v = w[v] \times f_{2d}(v), \text{ where}$$

$$w[v] = \begin{cases} 1 - \text{sign}(v_{tsdf}) \times v_{tsdf}, & \text{if } v_{tsdf} \leq \delta; \\ 0, & \text{else.} \end{cases} \quad (1)$$

In Equation (1), $f_{2d}(v)$ denotes the corresponding 2D feature vector of v on the image plane, v_{tsdf} is the TSDF value at v , δ is a threshold hyper-parameter that determines whether v should be initialized with non-empty features, and $w[v]$ controls how much information will be transferred to v . In practice, $w[v]$ is set to 1 if v is occupied to assign observed surfaces with non-filtered information.

The design of distance-aware projection takes inspiration from the phenomenon that missing areas often share similar semantic patterns with front observed surfaces at a certain distance. Since the correlation between observable and unobservable areas decreases with distance, the distance-aware projection initializes occluded voxels with intensity-decreasing signals. The attenuation degree of signal intensity is controlled by weights $\{w[v]\}$ that are negatively correlated with the distance, as shown by Equation (1).

With the distance-aware projection, many occluded voxels are initialized as non-empty, making the 2D slices in a 3D volume more informative than those yielded by general depth-guided projection as shown in Figure 3.

To enrich and further densify the volumetric features, a feature enrichment module is subsequently set with an encoder-decoder architecture, which is composed of dimensional decomposition residual blocks [24] and 3D transpose convolutions (Figure 2). By linking neighboring volumetric features, the feature

enrichment module provides reliable context priors for constraining multi-view consistency and rich semantic pattern priors to support the following 2D context inpainting.

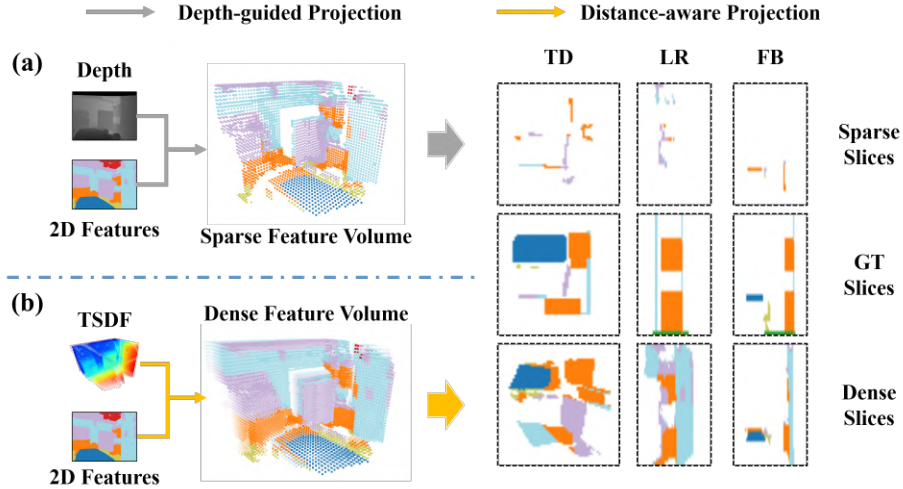


Fig. 3: Comparison between the 2D slices generated with the depth-guided projection and our distance-aware projection. (a) The depth-guided projection outputs a sparse 3D volume that leaves only a little information in 2D slices. (b) Our distance-aware projection ensures a dense 3D volume to provide rich semantic pattern priors in 2D slices.

3.3 Side-view Context Inpainter

Considering a feature volume can be regarded as a group of concatenated 2D maps, we propose a 2D context inpainter to perform semantic pattern completion in 2D space, alleviating the learning difficulty caused by directly reasoning complex 3D relationships. As depicted by Figure 4, the 3D volume can be split into groups of 2D feature slices in terms of three side views, and the side-view 2D context inpainter is constructed to complete the unknown 3D semantic patterns in 2D space with three side-view reasoning branches.

Specifically, the side-view reasoning branches independently process the 3D feature volume in terms of top-down (TD), left-right (LR), and front-back (FB) views. For example, the reasoning branch for the top-down view is only in charge of inferring the 2D semantic patterns of feature maps sliced along the height dimension. Each branch is made of three consecutive 3×3 convolution layers with dilation values set to 1, 2, and 3. The 2D features inpainted from different side views are fused with a $1 \times 1 \times 1$ convolution for the final voxel-wise semantic prediction. The detailed mechanism is formulated as

$$F^{TD/LR/FB} = CBR_{3 \times 3}^3 \left(\left\{ f_i^{td/LR/FB} \right\} \right); \quad (2)$$

$$\hat{F} = CBR_{1 \times 1 \times 1} ([F^{TD}, F^{LR}, F^{FB}, F]).$$

where $F^{TD/LR/FB}$ refers to volumes obtained by inpainting 2D maps $\{f_i^{TD/LR/FB}\}$ generated from the enriched feature volume input F w.r.t three views, CBR is a combination of convolution, batch normalization and relu activation, and CBR^3 is a sequence of three CBRs.

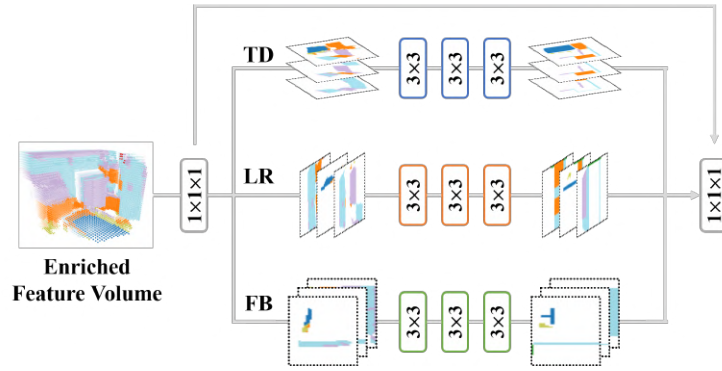


Fig. 4: Mechanism of the side-view 2D context inpainter. TD: top-down side view; LR: left-right side view; FB: front-back side view.

The side-view context inpainter is deployed after the last 3D transpose convolution of the feature enrichment module and before the final linear classifier. Such deployment benefits the optimization of the side-view context inpainter with a short transmission distance of supervision signal, which also strengthens the multi-view consistency in the 2D in-painting process.

3.4 Training Loss

We use a cross entropy loss L_{ce} and a pattern matching loss L_{pm} for model optimization. As revealed by Equation (3), L_{ce} works with the prediction V^{pred} and the 3D semantic annotation V^{gt} , while L_{pm} computes the mean squared errors between the intermediately encoded features F^{enc} and the real semantic pattern priors F^{sem} .

$$L = L_{ce}(V^{pred}, V^{gt}) + L_{pm}(F^{enc}, F^{sem}). \quad (3)$$

In L_{pm} , F^{enc} comes from the two encoding stages of the feature enrichment module, and F^{sem} is derived from the same stages of a copy module that is fed and trained with the 3D semantic ground-truths. L_{pm} explicitly guides the

learning of intermediate features, thereby guaranteeing reliable semantic pattern priors for consistent side-view context inpainting.

4 Experiments

4.1 Experiment Settings

Datasets. We evaluated our method on three public semantic scene completion benchmarks, including **NYUv2** [35], **NYUCAD** [12] and **SUNCG-RGBD** [36]. The NYUv2 benchmark includes 1,449 RGB-D images, each of which has a complete 3D semantic annotation offered by [12]. The NYUCAD dataset replaces the depth data in NYUv2 with higher-quality depth maps derived from the 3D annotations. The SUNCG-RGBD benchmark used here is generated from synthetic data by [29] and contains 13011 pairs of images and 3D annotations for training while 499 for testing. As there are some wrongly generated 3D annotations, we cleaned the SUNCG-RGBD dataset, leaving 10980 pairs for training and 426 for testing.

Implementation Details. We implemented our method with Pytorch [33] and conducted experiments on one NVIDIA RTX3090TI GPU. The 2D feature encoder in our model is built with a swin transformer backbone [30] pretrained on NYUv2 and SUNCG-RGBD with semantic segmentation, and the threshold δ in Equation (1) is set to 0.3 for NYUv2 while 0 for NYUCAD and SUNCG-RGBD. Our model is trained by an SGD optimizer with momentum of 0.9, weight decay of $5e-4$, and batch size of 8. The learning rate is initialized as 0.01 and multiplied by $\left(1 - \frac{iter}{max_iter}^{0.9}\right)$. For NYUv2 and NYUCAD, the feature enrichment module for generating F^{sem} is trained once for 50 epochs and the whole model is trained for 500 epochs. As for SUNCG-RGBD, the former training period is 20 epochs while the latter one is 50 epochs.

Evaluation metrics. We use the scores of category-wise intersection over union (IoU) and mean IoU (mIoU) to evaluate the performance of semantic scene completion (SSC). The IoU metric is also leveraged to measure the scene geometry completion (SC) by only categorizing voxels as occupied or empty.

4.2 Quantitative Comparison with Other Methods

Quantitative results on NYUv2 and NYUCAD datasets. Table 1 and Table 2 present the numeric comparison between our model and existing methods on NYUv2 and NYUCAD. Our model outperforms prior arts under most metrics with state-of-the-art performance. For example, our model gains over other methods by at least 8.1% and 7.2% mIoU on NYUv2 and NYUCAD for semantic scene completion, respectively. It should be noted that our method only uses TSDF for distance-aware projection while other multi-modality learning approaches generally rely on a sub-network to independently extract geometry features from

depth, TSDF, or point cloud data. In this case, our model works via a single-stream framework without a special design for geometry-focused learning but still achieves competitive performance for scene geometry completion.

Table 1: Numeric comparison on NYUv2 dataset. Inputs: the letter D denotes the depth or HHA images, and TSDF † means that the TSDF is not fed into a sub-network for feature extraction. The best is **bolded**, while the second is underlined.

Methods	Inputs	SC IoU(%)	SSC Category-wise IoU(%) and mIoU(%)											
			ceil.	floor	wall	win.	chair	bed	sofa	table	TVs	furn.	objs.	avg.
SSCNet [36]	TSDF	55.1	15.1	<u>94.7</u>	24.4	0.0	12.6	32.1	35.0	13.0	7.8	27.1	10.1	24.7
ESSCNet [45]	TSDF	56.2	17.5	75.4	25.8	6.7	15.3	53.8	42.4	11.2	0	33.4	11.8	26.7
ForkNet [41]	TSDF	63.4	36.2	93.8	29.2	18.9	17.7	<u>61.6</u>	52.9	23.3	19.5	45.4	20.0	37.1
TS3D [14]	RGB+D	60.0	9.7	93.4	25.5	21.0	17.4	55.9	49.2	17.0	27.5	39.4	19.3	34.1
SATNet [29]	RGB+D	60.6	17.3	92.1	28.0	16.6	19.3	57.5	53.8	17.2	18.5	38.4	18.9	34.4
CCPNet [46]	TSDF	63.5	23.5	96.3	35.7	20.2	25.8	61.4	<u>56.1</u>	18.1	<u>28.1</u>	37.8	20.1	38.5
DDRNet [24]	RGB+D	61.0	21.1	92.2	33.5	6.8	14.8	48.3	42.3	13.2	13.9	35.3	13.2	30.4
AICNet [23]	RGB+D	59.2	23.2	90.8	32.3	14.8	18.2	51.1	44.8	15.2	22.4	38.3	15.7	33.3
PALNet [25]	TSDF+D	61.3	23.5	92.0	33.0	11.6	20.1	53.9	48.1	16.2	24.2	37.8	14.7	34.1
IPF-SPCNet [48]	RGB+Point	39.0	32.7	66.0	41.2	17.2	34.7	55.3	47.0	21.7	12.5	38.4	19.2	35.1
AFMNet [28]	RGB+D	57.2	20.0	78.7	27.3	20.5	21.8	56.5	53.9	19.5	18.8	40.1	19.5	34.2
IMENet [22]	RGB+D	72.1	43.6	93.6	<u>42.9</u>	<u>31.3</u>	<u>36.6</u>	57.6	48.4	<u>32.1</u>	16.0	<u>47.8</u>	<u>36.7</u>	44.2
3D-Sketch [5]	RGB+TSDF	71.3	43.1	93.6	40.5	24.3	30.0	57.1	49.3	29.2	14.3	42.5	28.6	41.1
FFNet [39]	RGB+D+TSDF	71.8	<u>44.0</u>	93.7	41.5	29.3	36.2	59.0	51.1	28.9	26.5	45.0	32.6	<u>44.4</u>
MFF [13]	RGB+TSDF+Point	73.1	45.4	92.3	41.1	25.6	32.6	58.3	49.8	30.5	17.1	44.1	33.9	42.8
Ours	RGB+TSDF†	<u>72.2</u>	42.6	93.9	45.5	38.4	46.9	66.2	66.2	37.2	41.4	53.7	45.3	52.5

Table 2: Numeric comparison on NYUCAD. Inputs: the letter D denotes the depth or HHA images, and TSDF † means that the TSDF is not fed into a sub-network for feature extraction. The best is **bolded**, while the second is underlined.

Methods	Inputs	SC IoU(%)	SSC Category-wise IoU(%) and mIoU(%)											
			ceil.	floor	wall	win.	chair	bed	sofa	table	tv	furn.	objs.	avg.
SSCNet [36]	TSDF	73.2	32.5	92.6	40.2	8.9	33.9	57.0	59.5	28.3	8.1	44.8	25.1	40.0
DDRNet [24]	RGB+D	79.4	54.1	91.5	56.4	14.9	37.0	55.7	51.0	28.8	9.2	44.1	27.8	42.8
AICNet [23]	RGB+D	80.5	53.0	91.2	57.2	20.2	44.6	58.4	56.2	36.2	9.7	47.1	30.4	45.8
TS3D [14]	RGB+D	76.1	25.9	93.8	48.9	33.4	31.2	66.1	56.4	31.6	<u>38.5</u>	51.4	30.8	46.2
CCPNet [47]	TSDF	82.4	56.2	94.6	58.7	35.1	44.8	68.6	65.3	37.6	35.5	53.1	35.2	53.2
3D-Sketch [5]	RGB+TSDF	84.2	<u>59.7</u>	94.3	<u>64.3</u>	32.6	51.7	72.0	68.7	45.9	19.0	60.5	38.5	55.2
PALNet [25]	TSDF+D	80.8	54.8	92.8	60.3	15.3	43.1	60.7	59.9	37.6	8.1	48.6	31.7	46.6
FFNet [39]	RGB+D+TSDF	85.5	62.7	94.9	67.9	<u>35.2</u>	<u>52.0</u>	<u>74.8</u>	69.9	<u>47.9</u>	27.9	<u>62.7</u>	35.1	<u>57.4</u>
MFF [13]	RGB+TSDF+Point	<u>84.8</u>	54.5	<u>94.8</u>	63.3	29.3	50.9	73.6	<u>70.9</u>	56.4	31.7	61.3	<u>42.0</u>	57.2
CasFusionNet [42]	Point	-	-	-	-	-	-	-	-	-	-	-	-	49.5
Ours	RGB+TSDF†	83.9	59.2	94.6	63.7	51.9	63.9	80.1	77.4	<u>47.9</u>	51.3	67.2	53.0	64.6

Quantitative results on SUNCG dataset. We compare our method with three multi-modality semantic scene completion models, including DDRNet, AICNet, and 3D-Sketch, and present the results in Table 3. Our SidePaint-based model gains over other approaches by at least 19.5% mIoU for semantic scene completion, as well as at least 3.5% IoU improvement for scene geometry recovery.

Table 3: Numeric results on SUNCG-RGBD. The best is **bolded**, while the second is underlined.

Methods	Inputs	SC IoU	SSC mIoU
DDRNet	RGB+D	71.8	38.5
AICNet	RGB+D	61.8	35.0
3D-Sketch	RGB+TSDF	<u>87.5</u>	<u>70.4</u>
Ours	RGB+TSDF†	91.0	89.9

4.3 Qualitative Analysis with Other Methods

We provide some visualization results on NYUv2, NYUCAD, and SUNCG-RGBD datasets for perceptual understanding in Figure 5. Our SidePaint-based model recovers the scenes with cleaner geometry structures and more consistent semantic completion than other compared methods.

4.4 Ablation Studies

For a deep understanding of the proposed SidePaint strategy, we separately study the distance-aware projection and the 2D context inpainter on NYUv2 and NYUCAD datasets. The baseline is a model that shares a similar architecture with our SidePaint-based framework. The only difference is that the baseline model is equipped with depth-guided projection and a 3D inpainter, where the 2D kernels in the 2D inpainter are replaced with 3D $3 \times 3 \times 3$ counterparts. We also evaluated the performance improvement brought by the pattern matching loss L_{pm} . The quantitative results are reported in Table 4.

Table 4 shows the effectiveness of the dense distance-aware projection and the 2D context inpainter in our SidePaint strategy. The numeric results reveal that the 2D context inpainter exceeds the 3D counterpart used in the baseline by 1.1%-2.9% mIoU on NYUv2 and NYUCAD datasets, under the cases with or without the distance-aware projection. This phenomenon suggests the superiority of reasoning context in 2D space over in 3D space. With the assistance of L_{pm} , the 2D context inpainter is further strengthened with reliable semantic pattern priors and finally outperforms the baseline by at least 4.7% mIoU on the two datasets. By comparing the results, it is interesting to find that the dense

● ceil. ● floor ● wall ● win. ● chair ● bed ● sofa ● table ● tvs. ● furn. ● objs.

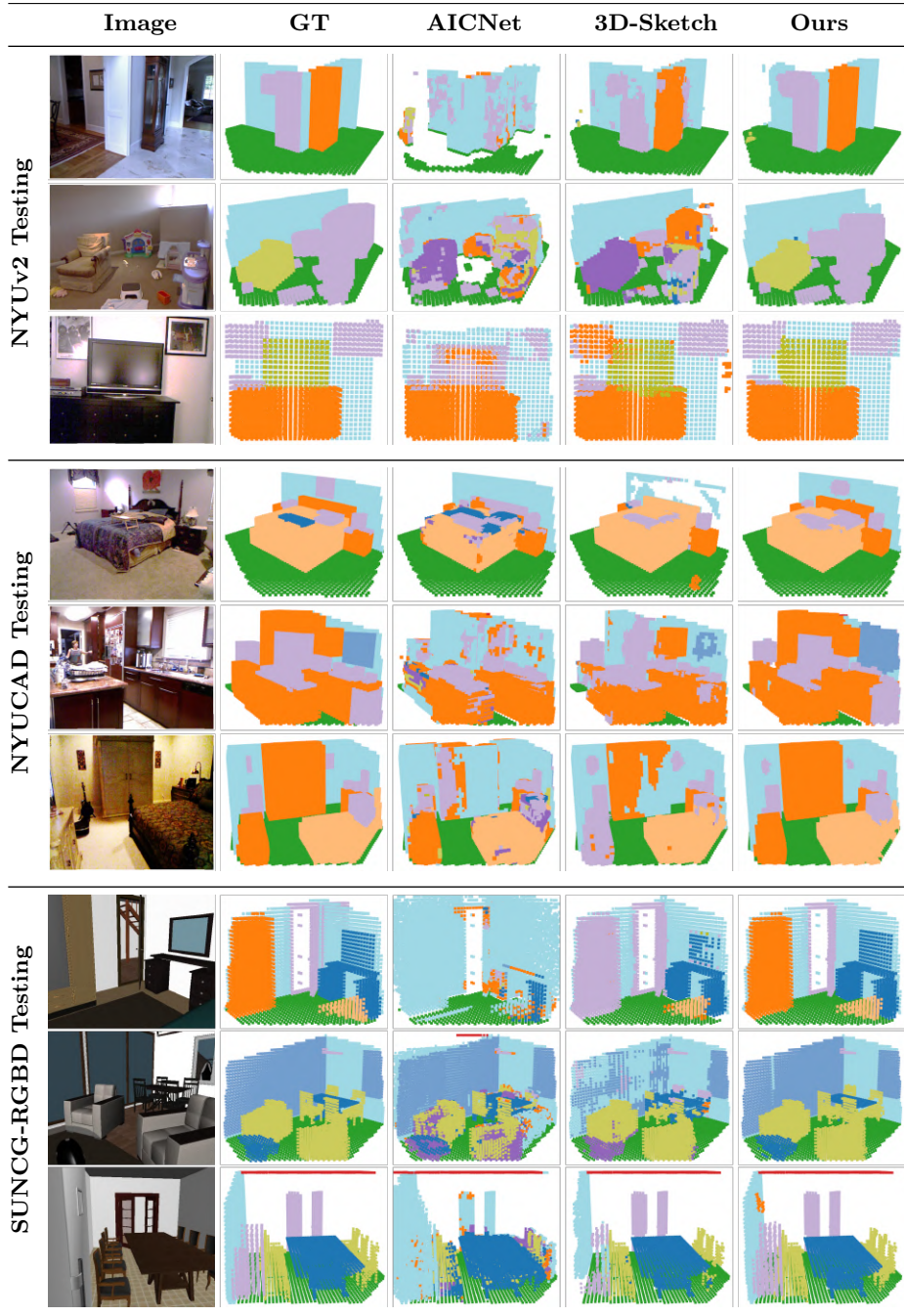


Fig. 5: Qualitative comparison for semantic scene completion on NYUv2, NYUCAD, and SUNCG-RGBD datasets.

Table 4: Ablation studies with NYUv2 and NYUCAD. Dist. Proj.: the dense distance-aware projection; 2D Inp.: the 2D context inpainter; 3D Inp.: the 3D context inpainter. The best is **bolded**, while the second is underlined.

Method	3D Inp.	2D Inp.	Dist. Proj.	L_{pm}	NYUv2		NYUCAD	
					SC_{IoU}	SSC_{mIoU}	SC_{IoU}	SSC_{mIoU}
Baseline	✓				65.3	46.3	81.7	59.9
Baseline	✓		✓		66.2	47.3	81.5	59.3
SidePaint		✓			66.5	47.4	82.1	61.3
SidePaint		✓	✓		<u>67.6</u>	<u>49.2</u>	<u>82.8</u>	<u>62.2</u>
SidePaint		✓	✓	✓	72.2	52.5	83.9	64.6

distance-aware projection fails with the baseline on NYUCAD but succeeds with the 2D context inpainter on both datasets. The distance-aware projection even works more effectively with the 2D inpainter than with the 3D one, especially for NYUv2 where the depth data quality is lower.

We additionally investigate the influence of the 2D context inpainter position in the model architecture. Table 5 shows that the 2D context inpainter performs better when it is placed near the classifier.

Table 5: Ablation studies on the position of 2D context inpainter. Before: the 2D context inpainter is inserted before the feature enrichment module; After: the 2D context inpainter is inserted after the feature enrichment module and before the classifier. The best is **bolded**.

Method	NYUv2		NYUCAD	
	SC_{IoU}	SSC_{mIoU}	SC_{IoU}	SSC_{mIoU}
SidePaint Before	69.2	50.3	82.9	63.6
SidePaint After	72.2	52.5	83.9	64.6

In Figure 6, we present visualization examples to conceptually compare the performance of the baseline and our SidePaint-based model. It can be found that the baseline suffers from semantic errors and geometry defects while our SidePaint recovers the scenes with correct semantic completion. This demonstrates that the SidePaint strategy facilitates SSC by learning context completion in 2D space with dense distance-aware feature volumes, which is much simpler than directly reasoning sophisticated 3D semantic patterns with sparse signals.

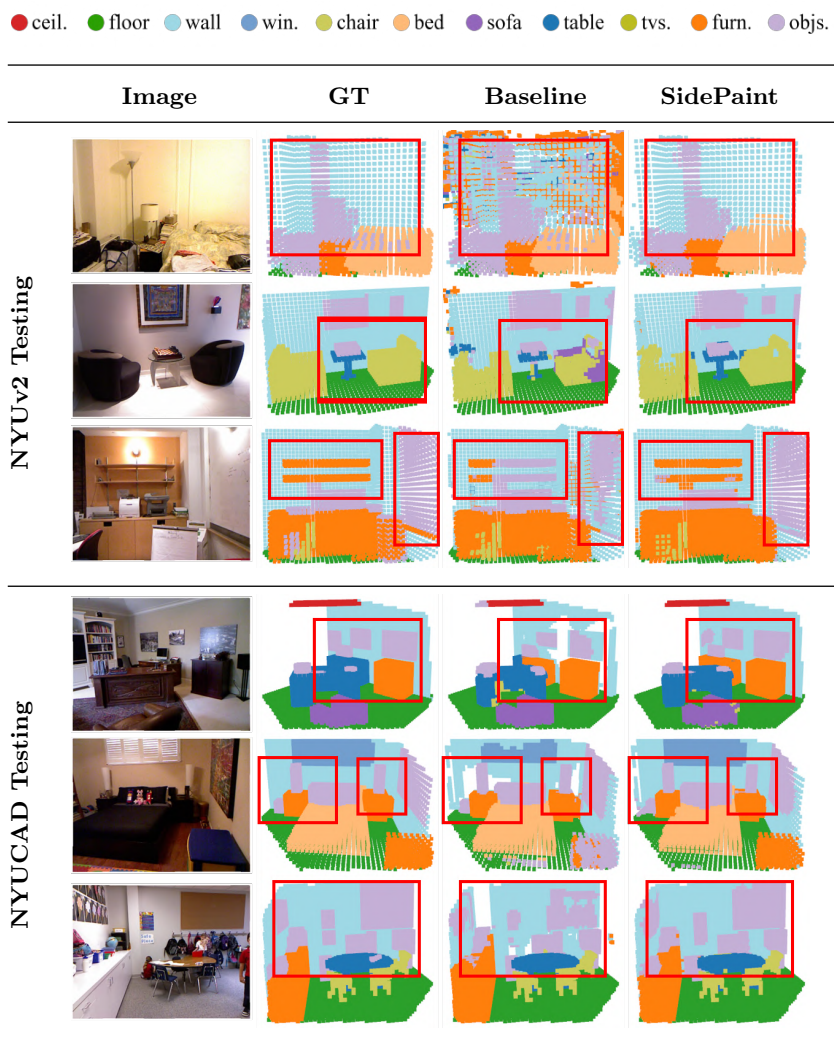


Fig. 6: Qualitative comparison for ablation studies.

5 Conclusion

In this paper, we propose a side-view context inpainting strategy (SidePaint) to alleviate the learning burden caused by reasoning highly complex 3D relationships with sparse observations for semantic scene completion. The SidePaint strategy simplifies the 3D semantic pattern completion by inpainting 2D side-view context with densely projected volumetric features. With the context complexity reduced in 2D space, the SidePaint strategy guarantees a cushy completion of missing semantic patterns with a single-branch deep model. Experi-

ments demonstrate that the spirit of recovering 3D semantic patterns in 2D space is more effective than directly reasoning 3D relationships. In comparison with state-of-the-art methods, the SidePaint-based model achieves competitive performance for not only semantic scene completion but also geometry recovery.

Acknowledgment

This research is supported by NSFC-projects under Grant 42071370 and the Fundamental Research Funds for the Central Universities of China under Grant 2042022dx0001.

References

1. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: IEEE International Conference on Computer Vision (ICCV) (2019)
2. Cai, Y., Chen, X., Zhang, C., Lin, K.Y., Wang, X., Li, H.: Semantic scene completion via integrating instances and scene in-the-loop. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
3. Cao, A.Q., de Charette, R.: Monoscene: Monocular 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3991–4001 (2022)
4. Chen, H.X., Huang, J., Mu, T.J., Hu, S.M.: Circle: Convolutional implicit reconstruction and completion for large-scale indoor scene. In: European Conference on Computer Vision. pp. 506–522. Springer (2022)
5. Chen, X., Lin, K.Y., Qian, C., Zeng, G., Li, H.: 3d sketch-aware semantic scene completion via semi-supervised structure prior. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
6. Chen, X., Xing, Y., Zeng, G.: Real-time semantic scene completion via feature aggregation and conditioned prediction. In: IEEE International Conference on Image Processing (ICIP) (2020)
7. Chen, Y.T., Garbade, M., Gall, J.: 3d semantic scene completion from a single depth image using adversarial training. In: IEEE International Conference on Image Processing (ICIP) (2019)
8. Dai, A., Ritchie, D., Bokeloh, M., Reed, S., Sturm, J., Nießner, M.: Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2018)
9. Dong, H., Ma, E., Wang, L., Wang, M., Xie, W., Guo, Q., Li, P., Liang, L., Yang, K., Lin, D.: Cvsformer: Cross-view synthesis transformer for semantic scene completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8874–8883 (2023)
10. Dong, M., Huan, L., Xiong, H., Shen, S.S., Zheng, X.: Shape anchor guided holistic indoor scene understanding. *International Journal of Computer Vision* (2023)
11. Dourado, A., Guth, F., de Campos, T.: Data augmented 3d semantic scene completion with 2d segmentation priors. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2022)

12. Firman, M., Mac Aodha, O., Julier, S., Brostow, G.J.: Structured prediction of unobserved voxels from a single depth image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
13. Fu, R., Wu, H., Hao, M., Miao, Y.: Semantic scene completion through multi-level feature fusion. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 8399–8406. IEEE (2022)
14. Garbade, M., Chen, Y.T., Sawatzky, J., Gall, J.: Two stream 3d semantic scene completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
15. Han, X., Zhang, Z., Du, D., Yang, M., Yu, J., Pan, P., Yang, X., Liu, L., Xiong, Z., Cui, S.: Deep reinforcement learning of volume-guided progressive view inpainting for 3d point scene completion from a single depth image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 234–243 (2019)
16. Hou, J., Dai, A., Nießner, M.: Revealnet: Seeing behind objects in rgb-d scans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
17. Huan, L., Zheng, X., Gong, J.: Georec: Geometry-enhanced semantic 3d reconstruction of rgb-d indoor scenes. ISPRS Journal of Photogrammetry and Remote Sensing **186**, 301–314 (2022)
18. Huang, S., Qi, S., Zhu, Y., Xiao, Y., Xu, Y., Zhu, S.C.: Holistic 3d scene parsing and reconstruction from a single rgb image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 187–203 (2018)
19. Huang, Y., Zheng, W., Zhang, Y., Zhou, J., Lu, J.: Tri-perspective view for vision-based 3d semantic occupancy prediction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9223–9232 (2023)
20. Izadinia, H., Shan, Q., Seitz, S.M.: Im2cad. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5134–5143 (2017)
21. Kulkarni, N., Misra, I., Tulsiani, S., Gupta, A.: 3d-relnet: Joint object and relational network for 3d prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2212–2221 (2019)
22. Li, J., Ding, L., Huang, R.: Imenet: Joint 3d semantic scene completion and 2d semantic segmentation through iterative mutual enhancement. arXiv preprint arXiv:2106.15413 (2021)
23. Li, J., Han, K., Wang, P., Liu, Y., Yuan, X.: Anisotropic convolutional networks for 3d semantic scene completion. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
24. Li, J., Liu, Y., Gong, D., Shi, Q., Yuan, X., Zhao, C., Reid, I.: Rgb-d based dimensional decomposition residual network for 3d semantic scene completion. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
25. Li, J., Liu, Y., Yuan, X., Zhao, C., Siegwart, R., Reid, I., Cadena, C.: Depth based semantic scene completion with position importance aware loss. IEEE Robotics and Automation Letters **5**(1), 219–226 (2019)
26. Li, J., Song, Q., Yan, X., Chen, Y., Huang, R.: From front to rear: 3d semantic scene completion through planar convolution and attention-based network. IEEE Transactions on Multimedia (2023)
27. Li, J., Wang, P., Han, K., Liu, Y.: Anisotropic convolutional neural networks for rgb-d based semantic scene completion. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8125–8138 (2022). <https://doi.org/10.1109/TPAMI.2021.3081499>

28. Li, S., Zou, C., Li, Y., Zhao, X., Gao, Y.: Attention-based multi-modal fusion network for semantic scene completion. In: AAAI Conference on Artificial Intelligence (AAAI) (2020)
29. Liu, S., Hu, Y., Zeng, Y., Tang, Q., Jin, B., Han, Y., Li, X.: See and think: Disentangling semantic scene completion. *Advances in Neural Information Processing Systems* **31** (2018)
30. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
31. Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.J.: Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 55–64 (2020)
32. Nie, Y., Hou, J., Han, X., Nießner, M.: Rfd-net: Point scene understanding by semantic instance reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4608–4618 (2021)
33. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)* (2019)
34. Roldao, L., De Charette, R., Verroust-Blondet, A.: 3d semantic scene completion: a survey. *International Journal of Computer Vision* **130**(8), 1978–2005 (2022)
35. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgb-d images. In: European Conference on Computer Vision (ECCV) (2012)
36. Song, S., Yu, F., Zeng, A., Chang, A.X., Savva, M., Funkhouser, T.: Semantic scene completion from a single depth image. In: IEEE International Conference on Computer Vision (ICCV) (2017)
37. Tang, J., Chen, X., Wang, J., Zeng, G.: Point scene understanding via disentangled instance mesh reconstruction. arXiv preprint arXiv:2203.16832 (2022)
38. Tulsiani, S., Gupta, S., Fouhey, D.F., Efros, A.A., Malik, J.: Factoring shape, pose, and layout from the 2d image of a 3d scene. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 302–310 (2018)
39. Wang, X., Lin, D., Wan, L.: Ffnet: Frequency fusion network for semantic scene completion. In: AAAI Conference on Artificial Intelligence (AAAI) (2022)
40. Wang, Y., Tan, D.J., Navab, N., Tombari, F.: Adversarial semantic scene completion from a single depth image. In: 2018 International Conference on 3D Vision (3DV). pp. 426–434. IEEE (2018)
41. Wang, Y., Tan, D.J., Navab, N., Tombari, F.: Forknet: Multi-branch volumetric semantic completion from a single depth image. In: IEEE International Conference on Computer Vision (ICCV) (2019)
42. Xu, J., Li, X., Tang, Y., Yu, Q., Hao, Y., Hu, L., Chen, M.: Casfusionnet: A cascaded network for point cloud semantic scene completion by dense feature fusion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 3018–3026 (2023)
43. Yan, X., Lin, L., Mitra, N.J., Lischinski, D., Cohen-Or, D., Huang, H.: Shapeformer: Transformer-based shape completion via sparse representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6239–6249 (2022)

44. Zhang, C., Cui, Z., Zhang, Y., Zeng, B., Pollefeys, M., Liu, S.: Holistic 3d scene understanding from a single image with implicit representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8833–8842 (2021)
45. Zhang, J., Zhao, H., Yao, A., Chen, Y., Zhang, L., Liao, H.: Efficient semantic scene completion network with spatial group convolution. In: European conference on computer vision (ECCV) (2018)
46. Zhang, P., Liu, W., Lei, Y., Lu, H., Yang, X.: Cascaded context pyramid for full-resolution 3d semantic scene completion. In: IEEE International Conference on Computer Vision (ICCV) (2019)
47. Zhang, P., Liu, W., Lei, Y., Lu, H., Yang, X.: Cascaded context pyramid for full-resolution 3d semantic scene completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7801–7810 (2019)
48. Zhong, M., Zeng, G.: Semantic point completion network for 3d semantic scene completion. In: European Conference on Artificial Intelligence (ECAI) (2020)