SSR-2D: Semantic 3D Scene Reconstruction from 2D Images

Junwen Huang[1] Alexey Artemov[1] Yujin Chen[1] Shuaifeng Zhi[2*] Kai Xu[2] Matthias Nießner[1]

[1]Technical University of Munich [2]National University of Defense Technology

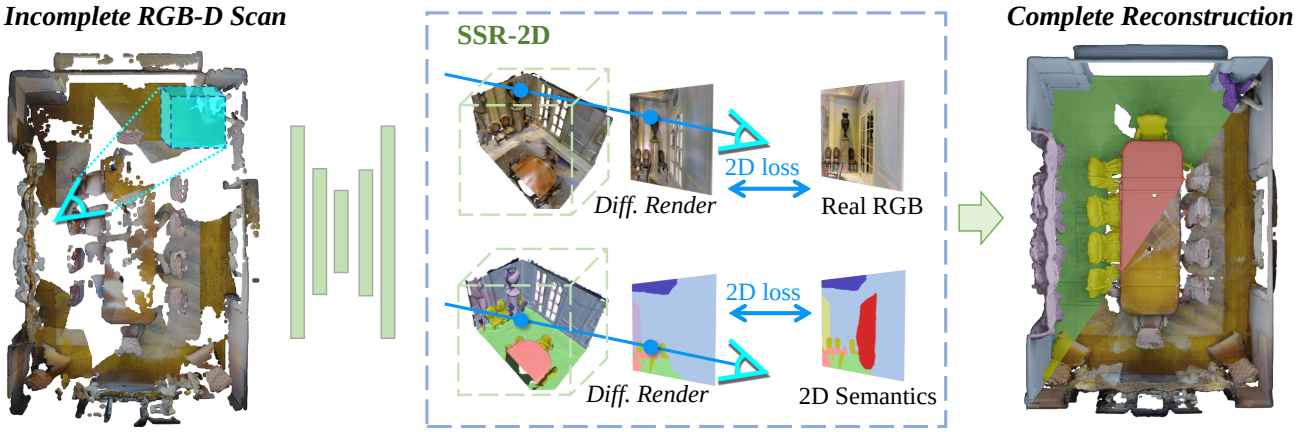


Figure 1: Given sparse RGB-D images, our method is capable of jointly predicting complete geometry, appearance, and semantic labels without access to in-place 3D ground-truth annotations during training.

Abstract

Most deep learning approaches to comprehensive semantic modeling of 3D indoor spaces require costly dense annotations in the 3D domain. In this work, we explore a central 3D scene modeling task, namely, semantic scene reconstruction without using any 3D annotations. The key idea of our approach is to design a trainable model that employs both incomplete 3D reconstructions and their corresponding source RGB-D images, fusing cross-domain features into volumetric embeddings to predict complete 3D geometry, color, and semantics with only 2D labeling which can be either manual or machine-generated. Our key technical innovation is to leverage differentiable rendering of color and semantics to bridge 2D observations and unknown 3D space, using the observed RGB images and 2D semantics as supervision, respectively. We additionally develop a learning pipeline and corresponding method to enable learning from imperfect predicted 2D labels, which could be additionally acquired by synthesizing in an augmented set of virtual training views complementing the original real captures, enabling more efficient self-supervision loop for semantics. In this work, we propose an end-to-end trainable solution jointly addressing geometry completion, colorization, and semantic mapping from limited RGB-D images, without relying on any 3D ground-truth information. Our method achieves state-of-the-art performance of semantic scene reconstruction on two large-scale benchmark datasets MatterPort3D and ScanNet, surpasses baselines even with costly 3D annotations. To our knowledge, our method is also the first 2D-driven method addressing completion and semantic segmentation of real-world 3D scans.

1. Introduction

There seems to exist an agreement that accurate modeling of geometry, appearance, and semantics are three essential ingredients for the construction of comprehensive digital replicas for real-world 3D scenes. Indeed, downstream applications such as the creation and manipulation of 3D assets, virtual-real interactions, or free-viewpoint virtual tours all require complete 3D shapes and faithful color information of objects [3, 33]. Others, supporting autonomous systems to localize themselves, to navigate environments, or to perform tasks like grasping, involve understanding semantics and accurate modeling of scene geometry [7, 15].

*Shuaifefgng Zhi is the corresponding author.

Digitizing real-world 3D scenes such as entire indoor areas, however, is widely recognized as a challenging endeavor; limitations of range 3D scanning and physical constraints such as occlusions disable acquiring complete 3D geometry (see, *e.g.*, [12, 22]). Moreover, even resorting to human experts for either scanning, semantic annotation, or artistic editing is unlikely to deliver flawless, complete digital 3D assets while being notoriously labour-intensive [3, 9]. This view has motivated much research on automatic (particularly, learning-based) approaches for completion, colorization, and semantic segmentation of raw single-view [2, 14, 35, 37], multi-view [11, 24], and fused 3D acquisitions [12, 13, 21, 22].

Despite these efforts brought in methods capable of recovering one or two ("geometry + semantics" or "geometry + color") of the target quantities from raw RGB-D scans independently, a comprehensive joint understanding of all three complementary signals has not been yet demonstrated. To address this gap, in this work we propose to jointly predict geometry, appearance, and semantics using a three-branch 3D CNN architecture, and empirically validate the feasibility of this scheme; to our best knowledge, our system is the first to predict the three complementary targets using a single trained model.

Training a multi-modal, highly parameterized 3D network in the 3D domain in a densely supervised manner is challenging as it depends on vast amounts of diverse, high-quality, complete, and labeled 3D data. Learning from purely synthetic data (*e.g.*, [32, 35]) is unlikely to well generalize to real 3D scans; on the other hand, ground-truth 2D/3D data and semantic annotation masks in common real-world RGB-D datasets (see, *e.g.*, [1, 3, 9]) are incomplete and imperfect. Instead, we opt for a number of design choices enabling our learning algorithm to leverage the original RGB-D images along with 2D semantics only.

First, inspired by recent methods [10, 13], we remove a subset of RGB-D frames from the input and learn to predict a complete semantic scene from an incomplete reconstruction; for this, we design a three-branch deep 3D CNN to jointly output geometry, color, and semantics in each voxel. Second, to support end-to-end optimization of our learning-based algorithm, we develop an extended differentiable rendering method, enabling us to render 3D volume data to depth, RGB and semantic images through raycasting directly. We design our training algorithm to reproduce the original RGB-D data; as a key ingredient for learning semantics, we learn segmentations provided by either manual annotations or a generic neural predictor trained on diverse, multi-domain data. Third, we additionally adapt the virtual view augmentation scheme inspired by the recent work [30] to further improve training performance given machine-generated imperfect labels.

The direction we chose is in line with recent approaches where 2D view information is used to supervise 3D predictions [13, 17]. Our method can additionally be viewed as generalizing upon several recent works [10, 12, 13] by integrating 2D RGB and semantic inputs as supervision; in several instances, we compare to these prior arts.

To summarize, our key contributions are as follows:

- To the best of our knowledge, our approach is the first to address the challenging task of semantic scene reconstruction from incomplete observations of challenging real-world indoor 3D scenes, without requiring manual 3D annotations.
- We achieve state-of-the-art semantic scene completion performance on two large-scale benchmarks, namely Matterport3D [3] and ScanNet [9].
- We demonstrate the practicability of our approach in an important special case by supervising it with generic, proxy segmentation labels, without access to expensive human 3D annotations.

2. Related Work

We briefly review closely related approaches that target analysis of large-scale, volumetric 3D scenes, mentioning works for other 3D representations where possible.

Semantic Scene Segmentation and its variants [19, 27] generally serve as an initial stage in scene analysis and continue to be extensively researched, in particular for RGB images (see, *e.g.*, [19, 26, 44] for a review). Yet, transferring 2D image segmentation to 3D world is non-straightforward; to this end, specialized point-, voxel-, and mesh-based methods were proposed (surveyed in [40]); we particularly note volumetric approaches [5, 9, 12, 38] as our network architecturally is a 3D CNN defined on a voxel grid. Seeking to empower 3D scene segmentation with image-based features, recent approaches propose multiple schemes to leverage 2D and 3D data in parallel [11, 17, 21, 24]. Among these, un-projecting per-pixel appearance features from nearby RGB-D images for 3D semantic [11] and 3D instance [21] segmentation, and employing bidirectional view-voxel projection to mutually reinforce 2D and 3D features [24] have proved to be effective. For 3D scenes represented as meshes, a more direct approach is to render and segment diversely sampled virtual 2D views using a pre-trained image-based network [30]; we adapt their view sampling scheme to our approach. All these approaches are fully supervised and need dense 3D annotation for training.

For reducing labeling costs and aiding generalization for 3D scene understanding, recent works sought to learn representations in a self-supervised fashion in a separate pre-training step (*e.g.*, [4, 23, 39]); however, these methods still require 3D annotations for fine-tuning. Most similarly to our approach, 2D3DNet [17] obtains 2D features in each image using a pre-trained segmentation model, projects

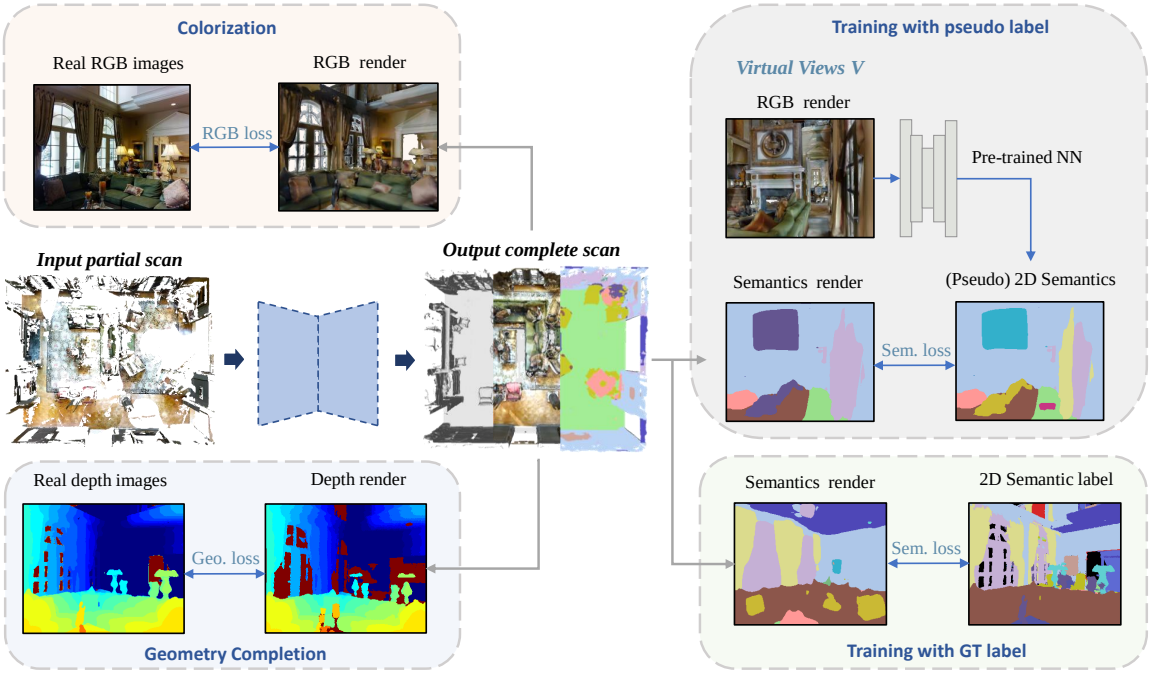


Figure 2: Our method accepts a fused but incomplete TSDF reconstruction as input and convolves it with a 3D encoder-decoder CNN, producing complete 3D geometry, colorization, and semantic segmentation. In the general case (Section 3.4), we generate 2D depth, color, and semantic images using either the original viewpoints U, or arbitrary virtual viewpoints V, by a differentiable rendering technique. These synthesized views are used to supervise training w.r.t. the original RGB-D images in a pseudo-supervised training loop, or w.r.t. multi-view consistency in a self-supervised training loop.

("lifts") these to 3D points, and refines them by a 3D network trained without 3D labels, bypassing the need for 3D annotations during training. [34] predicts semantic segmentation for a target viewpoint by rendering a volumetric 3D representation of projected semantics predicted by a pre-trained segmentation model. Similarly to the latter two works, one of our experiments uses a pre-trained generic segmentation network. All these approaches entirely leave out geometry completion or refinement, instead relying on raw scanned geometry.

Scene Completion. A common requirement in applications is to infer semantic labels not only in directly observed but also in occluded space; to this end, semantic scene completion (SSC) [35] seeks to address both scene occupancy completion and semantic object labeling jointly. Single-view depth images can be viewed as minimal input data [18, 35, 36, 37, 42, 43]; alternatives [2, 12, 14, 22] (including our method) tackle completing fused reconstructions of entire 3D spaces. [12] jointly predicts a truncated, unsigned distance field and per-volume semantics in a series of hierarchy levels ranging from low to high resolution. Leveraging RGB-D image back-projection, [22] combines geometry and appearance features to infer semantics and refine 3D geometry; [14] adopts a view of RGB image segmentation as a prior and computes the final semantic scene completion using a 3D CNN. [2] exploits an interplay between the scene- and instance-level completion tasks and alternates between semantic scene completion and detection of object instances. All these methods require dense 3D semantic labels and complete 3D reconstructions during training, making them dependent on synthetic data; in contrast, our algorithm is able to (1) use incomplete, real-world scenes and (2) train from photometric losses obtained via appearance synthesis and segmentations in the 2D domain.

Self-Supervised Scene Analysis. SG-NN [10] learns a scene completion network in a self-supervised learning task where a more complete sparse truncated signed distance field (TSDF) volume is used to guide prediction with inputs from a less complete one. Most similarly to our work, SPSG [13] diverts from using imprecise and incomplete 3D volumetric TSDF values as targets and leverages 2D appearance view synthesis where 2D image-based losses are used for supervising both completion and colorization. Our work presents an important extension of this line of research by enabling semantic segmentation in addition to generating complete and photometrically realistic scenes. Recently proposed semantics-enabled variants of neural radiance fields (NeRF) [16, 29, 45] strive to represent 3D scene geometry and semantics by a unified neural encoding; unlike these methods, our approach does not require expensive

per-scene optimization and can generalize across hundreds of unseen scenes.

3. Method

3.1. Method Overview

The goal of our method is to train a generalizable network g to perform semantic geometry completion, appearance (color) reconstruction, and semantic labeling without having access to any 3D ground-truth (GT) annotations during training. The input to our method is a set of RGB-D frames $\{(I_u, D_u), u \in U\}$ and their respective estimated camera poses. To generate input reconstructions, we select a subset of views $\widetilde{U} \subset U$, fusing these into a truncated signed distance field (TSDF) representation d_{in} through volumetric fusion [8], projecting color c_{in} into each voxel. As an output, our model predicts a corrected TSDF value $\widehat{d}$, color $\widehat{c}$, and semantic label $\widehat{s}$ in each voxel of the input grid.

Our network g follows a 3D U-shaped encoder-decoder architecture with two encoders processing geometry and color, and three decoder branches to output geometry, appearance and semantic labels for each voxel, respectively (Section 3.2). Next, having computed predictions, we synthesize depth $\widehat{D}_v$, appearance $\widehat{I}_v$, and semantic $\widehat{S}_v$ views via a raytracing-based differentiable rendering process for TSDF volumes [13] (Section 3.3). To enable self-supervised training, we minimize a set of 2D losses involving the ground-truth image I_v, a synthesized colour image $\widehat{I}_v$, a semantic map S_v^{P} (either a reference, or produced by a generic semantic segmentation model), and a reconstructed synthetic semantic map $\widehat{S}_v^{\text{R}}$ generated via rendering (Section 3.4). For geometry completion, we additionally self-supervise training using the incomplete scan to produce a more complete scan. The overall pipeline of our training procedure is shown in Figure 2.

3.2. Semantic Scene Reconstruction Architecture

We base our network architecturally on the variant proposed previously for photometric scene generation [13]. As input, our algorithm accepts a 4D tensor (*i.e.*, 3D volume with per-voxel TSDF RGB values). To extract geometry and color features from the input volumes, we use a dense 3D U-Net [6] backbone comprised of two 3D encoder branches with 5 ResNet-type [20] convolutional blocks each, and three 3D decoder branches with an equal number of convolutional blocks in each. The architecture of our 3D CNN is shown in Figure 3, where we summarise the data flow between the layers in our convolutional model. The full architecture is presented in the supplemental.

Input and Network Varieties. We note a few important architectural differences w.r.t. the original SPSG prototype [13], subject to available inputs. In some instances, raw captured inputs cannot include color images but provide

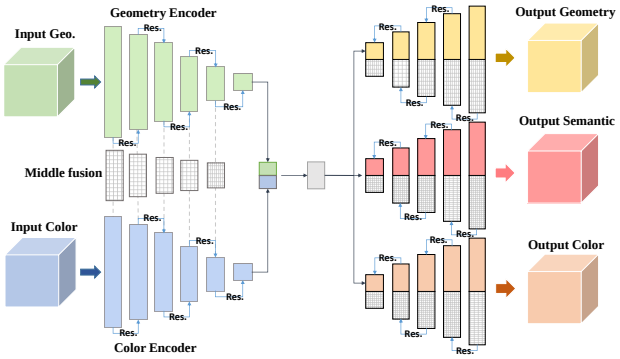


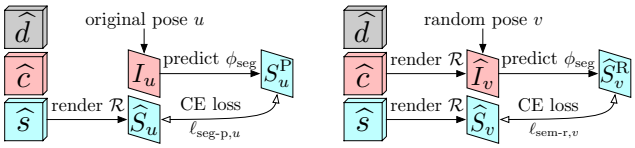
Figure 3: Our 3D CNN architecture comprises two encoders for geometry and color, and three decoders for completed geometry, semantics, and color, respectively.

depth only $\{D_u\}$. Importantly, in this regime, colors can potentially assume arbitrary values (*e.g.*, walls in a room may be painted white or vividly textured); however, hallucinating realistic appearance from geometry alone is a difficult generative problem and would require non-trivial complications to our model's architecture (*e.g.*, including an adversarial component in [13]). To verify, we modify our model and make it entirely "color-blind" by removing 2D and 3D color encoders and decoders, disabling appearance synthesis during rendering and excluding RGB loss term. Overall, we have found depth-only inputs to considerably decrease performance compared to RGB-D inputs. We leave detailed specification on training each variant for Section 3.4.

3.3. Differentiable Rendering of Depth, Color and Semantics

Our key design choice is to train a network defined in 3D space, but leverage the information contained in the original 2D RGB-D images (possibly, augmented with semantic information) instead of relying on 3D annotation directly. We thus require a 3D-to-2D conversion to enable gradient flow from pixelwise to voxelwise representations. Such operations, known as *differentiable volumetric rendering* [25], have proven essential in multiple tasks (*e.g.*, scene generation [13] or surface reconstruction [41]). Among these ops, we opted to extend a straightforward, efficient raycasting-based rendering approach for TSDF volumes [13] with a subroutine to render semantic maps.

Our differentiable rendering algorithm $\mathcal{R}$ accepts a predicted TSDF volume $\widehat{d}$ with per-voxel predicted colors $\widehat{c}$ and semantics $\widehat{s}$ and produces a set $\{\widehat{D}_v = \mathcal{R}(\widehat{d}; v), \widehat{I}_v = \mathcal{R}(\widehat{c}; v), \widehat{S}_v = \mathcal{R}(\widehat{s}; v)\}$ of depth, color and semantic images, respectively. To this end, we select RGB-D images taken from the viewing directions $\{v\}$ with the most overlap w.r.t. the chunk surface (top 5 views each with at least 5% depth samples within 2 cm to the near-surface voxels). For rendering semantics, we use a binary mask to represent each semantic class through raycasting, obtaining n_{sem}-channel



(a): pseudo-supervised (b): self-supervised

Figure 4: To train with pseudo-GT labels, we optimize: (a) for the original views, deviations between segmented RGBs $\phi_{\text{seg}}(I_u)$ and same-view semantic renderings $\mathcal{R}(\widehat{s}; u)$; (b) for randomly sampled views, deviations between semantic predictions of RGB renderings $\phi_{\text{seg}}(\mathcal{R}(\widehat{c}; v))$ and direct semantic renderings $\mathcal{R}(\widehat{s}; v)$ in the same view.

one-hot semantics image where n_{sem} equals the number of semantic classes. Depth and color rendering are obtained using the original process from [13]. The resulting RGB color images contain three channels, and the semantic images contain n_{sem} channels, representing binary semantic masks for each of the semantic classes in the class taxonomy of the respective collection [3, 9]. Image resolution of synthesized views is 320×256 pixels.

Varying View Synthesis. Our system by default synthesizes depth, color, semantics views of the scene; however, to explore depth-only training, we disable color rendering and only produce depth and semantics. We note that as the predictions are assumed to be complete in 3D, one may use arbitrary viewing directions; we explore this idea in our view augmentation technique in Section 4.1.

3.4. End-to-End Training with 2D Supervision

Our learning algorithm is conceptually inspired by SPSG [13], but differs from it in a number of ways, most importantly, by injecting semantic supervision and excluding adversarial components, which considerably simplifies our system. Similarly to SPSG though, our final scheme involves the objectives formulated in the 2D domain only, is end-to-end differentiable, and results in a 3D CNN model.

Geometry Completion. To self-supervise geometry completion, we use rendered depth images and penalize deviations from captured depth via pixel-wise L_1 loss

$$\ell_{\text{geo},u} = \sum_p ||D_u(p) - \widehat{D}_u(p)||_1. \tag{1}$$

We additionally use an L_1 3D loss term $\ell_{\text{geo-3d}}$ to self-supervised geometry reconstruction on the predicted 3D TSDF distances directly.

Appearance Reconstruction using Raw RGB Images. SPSG [13] heavily emphasized the need for adversarial training and optimizing a visual loss for achieving perceptually compelling RGB synthesis. In contrast, we opted to train without an adversarial part with either color or normal maps, hence bypassing the need for training a discriminator; likewise, to make the training task easier for the optimizer, we additionally exclude the perceptual loss term. Overall, we have found these modifications to have limited effect on achieving high-quality completion and semantic segmentation while simplifying our system and bringing down the number of trainable parameters. As a result, to enable faithful color synthesis using our model, we simply minimize per-pixel L_1 distances between the synthesized appearance view $\widehat{I}_u$ and the target view I_u

$$\ell_{\text{app},u} = \sum_p ||I_u(p) - \widehat{I}_u(p)||_1. \tag{2}$$

Method	Matterport3D [3]		ScanNet [9]	
	mAcc	*mIoU*	*mAcc*	*mIoU*
BPNet [24]	12.3	10.9	29.5	22.7
VMFusion [30]	24.9	17.6	12.3	10.9
ScanComplete [12]	34.9	28.2	26.9	22.7
Ours	**44.8**	**28.6**	**50.8**	**31.2**

Table 1: Our method outperforms three strong baselines and sets the new state-of-the-art on the semantic scene completion on two challenging, real-world benchmarks.

Semantic Segmentation with 2D Supervision. Our semantic loss follows a general intuition requiring that the segmentation $\widehat{s}$ inferred in the 3D domain generates plausible 2D semantic maps $\{\widehat{S}_v\}$ under a certain set of 2D views $\{v\}$. We consider the segmentation labels $\{S^{\text{P}}_u\}$ available for the *original captured* RGB-D images $\{(I_u, D_u), u \in U\}$ (we elaborate on an important special case below) and compute a cross-entropy (CE) loss $\ell_{\text{seg-p}}$ between each pair of rendered $\widehat{S}_u$ and reference S^{P}_u semantic views:

$$\ell_{\text{seg-p},u} = \sum_p L_{\text{CE}}(\widehat{S}_u(p), S^{\text{P}}_u(p)). \tag{3}$$

Summary: Supervised Formulation. Our final training objective integrates geometry, color, and semantic terms

$$L = \sum_{u \in U} \frac{1}{n_u} \left[\ell_{\text{geo},u} + \ell_{\text{app},u} + \ell_{\text{seg-p},u}\right], \tag{4}$$

over the set of n_u valid pixels in u (*i.e.*, pixels where surface geometry was predicted in $\widehat{d}$), and the 3D term of the form $\sum_{\text{chunks}} \ell_{\text{geo-3d}}$. To calculate the loss terms in the 2D domain, we use a 3D volumetric mask of the form $\{x : \widehat{d}(x) < \varepsilon\}$ (we use $\varepsilon = 3\,\text{cm}$) corresponding to the generated geometry, available upon completing geometry in each voxel of the input volumetric grid.

A Pseudo-Supervised Formulation. We mention a separate, relevant for applications, instance of our task, which consists in complete absence of ground-truth semantic segmentation labels for the original RGB-D data. Indeed, while obtaining reasonable RGB-D captures is increasingly

Method	*Geo. -Recall*	*Geo. -IoU*
SGNN [10]	57	28
SPSG [13]	64	39
ScanComplete [12]	36	30
Ours	**67.9**	**40.2**

Table 2: Geometry completion results on Matterport3D.

cheap, their semantic labelling (particularly, manual) is equally expensive; the question is then: can we still obtain semantic reconstructions in 3D?

We opted to address this question by using a generic semantic predictor (either a pre-trained neural network or an untrained model such as CRFs [28]).

More formally, let ϕ_{seg} denote a function that maps an observed RGB image I to per-pixel semantic labels S. We use ϕ_{seg} to obtain a set of *pseudo-ground-truth* semantic labels, via $S_u^{\text{P}} = \phi_{\text{seg}}(I_u)$ (see Figure 4 (a)). However, as the finite set of source views potentially limits the volume of supervision available to our model, the generic predictor gives us the flexibility to generate pseudo label on more (synthesized) RGB images from arbitrary poses, we thus create a self-supervised training loop (see Figure 4 (b)) using supervision from these virtual views. As a result,during training under this, we include an extra supervision from this training loop (see Figure 4) with an additive term semantic segmentation cost. We leave more details for supplemental.

4. Experiments

4.1. Experimental Setup

Benchmark Datasets. To evaluate our system and validate our design choices, we conduct a series of experiments using the challenging large-scale real 3D scans in the Matterport3D [3] and ScanNet dataset [9] collections. Both collections provide sufficient amounts of real-world training and testing data; following the official guidelines, we use 1788 spaces for training and validation, and 394 for testing on Matterport3D; on ScanNet, we use 1201 spaces for training and 312 for validation. To obtain highly detailed reconstructions, we use fine voxels with a 2 cm resolution during TSDF fusion; to enable memory efficient training, we extract $64 \times 64 \times 128$ subvolumes from fused scans like [13]. Overall, we use 77,581 and 88,420 chunks for training with Matterport3D and ScanNet respectively.

Evaluation Metrics. Following [10, 12, 13, 24, 30], we report mean intersection-over-union (mIoU) and mean voxel-wise accuracy (mAcc) for semantic segmentation, and geometric IoU as well as Recall rate for geometry completion.

Training with Virtual View Selection. For training with pseudo-GT labels, we model these using a pre-trained MSeg semantic segmentation network [31], treating it as a generic semantic predictor (ϕ_{seg} in Section 3.4); we stress that this model is applied in test mode, on data unseen during its training. Following [12], we map the 196 universal classes of MSeg into the most frequent categories in our test data (11 for Matterport3D and 15 for ScanNet). Quantitatively, our generic predictor demonstrates a mean IoU performance of 36.9% and 37.4% for Matterport3D and ScanNet, respectively (on the training split without any fine-tuning).

Method	Supervision	*mAcc*	*mIoU*
ScanComplete [12]	3D GT	46.6	35.8
BPNet [24]	3D GT+2D GT	47.7	33.3
Ours	2D GT	**48.3**	**36.7**

Table 3: Semantic segmentation results on Matterport3D. In this context, we provide the first self-supervised semantic segmentation results.

During training, we on-the-fly generate additional, virtual views independent to the original views in our data, to provide auxiliary semantic supervision. A similar technique is proposed in [30], yielding improved performance of a semantic 3D mesh segmentation by sampling views with otherwise unusual directions and fields of view. We give more details in the supplemental.

4.2. Comparisons to State-of-the-Art

Semantic Scene Completion. As a semantic scene completion baseline, we use ScanComplete [12], a supervised method operating on 3D TSDF volumes. For geometry completion, we additionally compare to the self-supervised SG-NN [10] and SPSG [13] that do not perform semantic segmentation. We note that among these methods, ScanComplete and SPSG perform complex multi-modal training while SG-NN focuses on completion only.

We present the results statistically in Tables 1–2. For both semantic segmentation and geometry completion constituents of the SSC task, our method outperforms all existing baselines, showing the effectiveness of our framework that jointly performs appearance prediction, geometry completion, and semantic segmentation, in a unified manner. Detailed visual comparisons of semantic scene completion of both datasets are presented in Figure 5.

Semantic Segmentation (without Completion). For reference, we additionally evaluate semantic labeling in isolation, without performing geometry completion, operating instead on a given (possibly, incomplete) scene. We provide semantic segmentation results on Matterport3D in Table 3 and Figure 6. Note that, in all instances, our network is trained for both completion and segmentation jointly; in this section, we only evaluate it with a setting focused on segmentation. For a strictly fair comparison against BPNet [24], we used the official implementation of BPNet and trained it to predict segmentation labels on input 3D scenes without scene completion; we report semantic segmentation performance of supplying it with the same training input to ours. Though only 2D labels are leveraged, our method out-

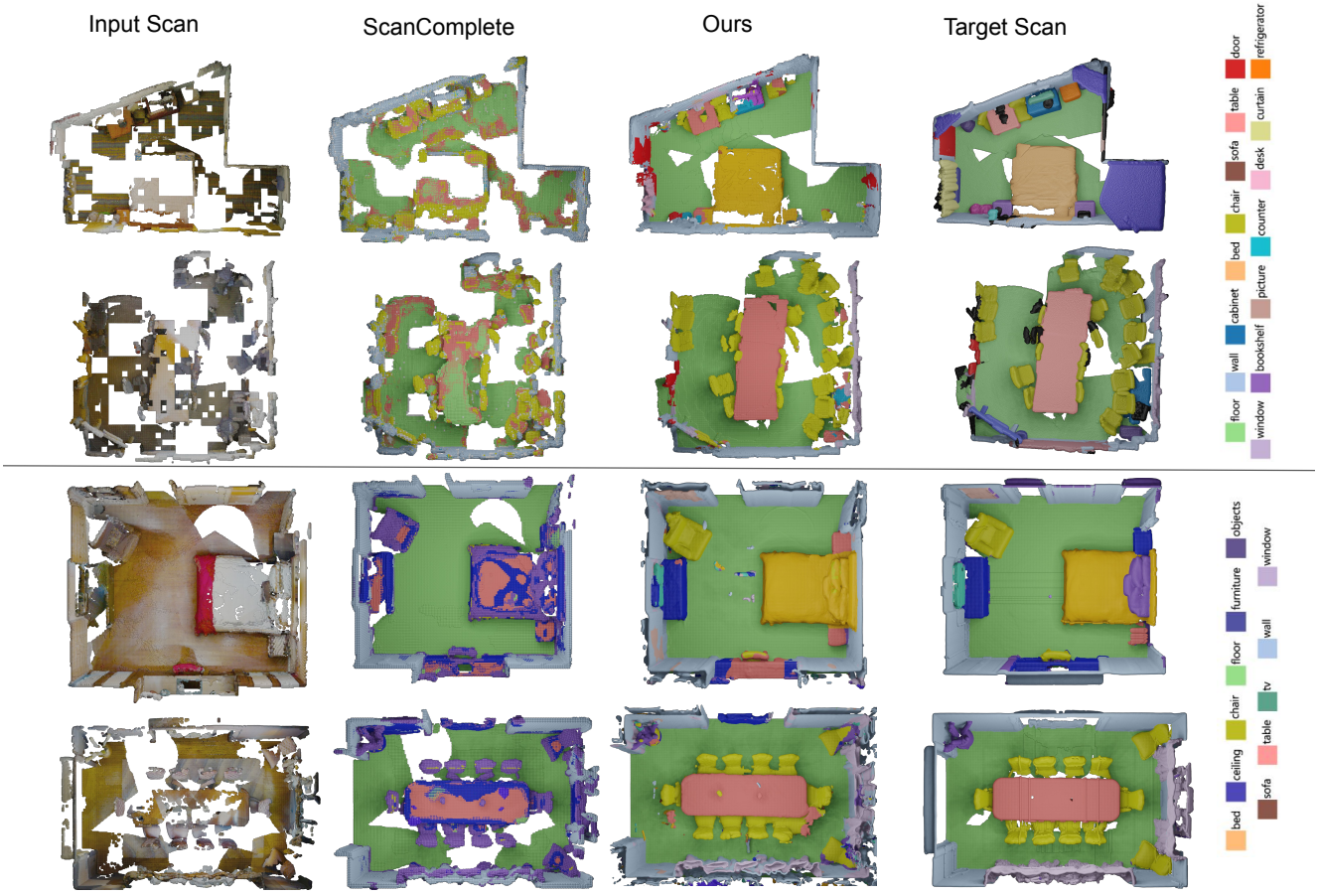


Figure 5: Qualitative SSC results using our approach and [12] on ScanNet (top rows) and Matterport3D (bottom) datasets.

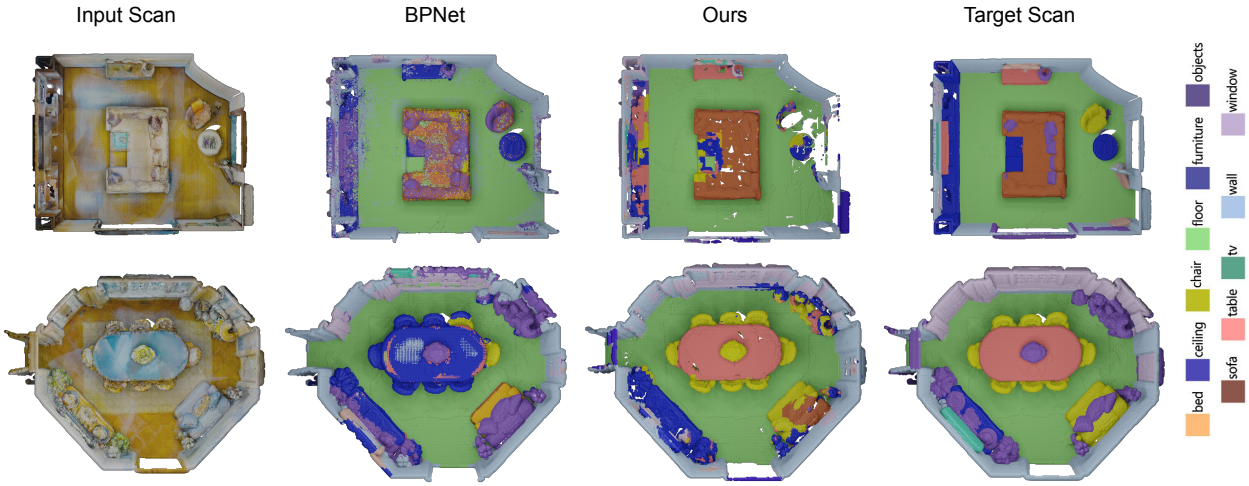


Figure 6: Qualitative semantic segmentation results using our approach and baseline method [24] on Matterport3D dataset. Compared to the baseline, our approach demonstrates robust performance.

performs all baselines in terms of segmentation accuracy, indicating that it is able to maintain high performance on seen 3D regions while generating new unobserved regions.

Semantic Fusion via Differentiable Rendering. Though our method differs conceptually from NeRF-based approaches such as [45] in a number of ways, making a direct point-to-point comparison less possible and meaningful, both methods aim to semantically reconstruct the 3D scenes from only posed 2D images and labels via differential rendering. Therefore, here we did our best to conduct

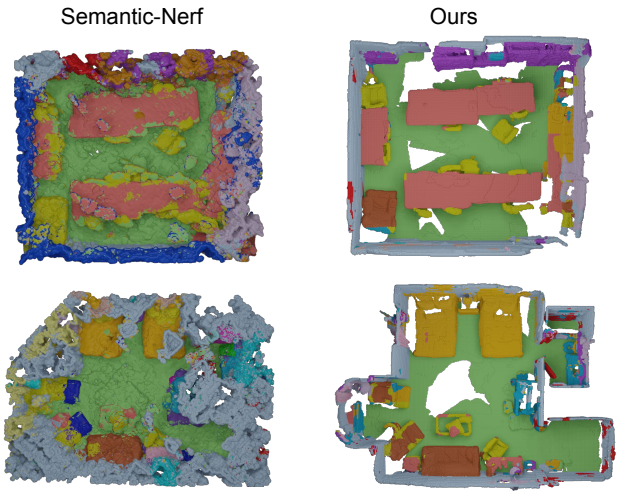


Figure 7: Qualitative semantic reconstruction results using our approach and baseline methods [45] on ScanNet dataset.

Method	ScanNet		MatterPort3d	
	Sem.-mIoU	*Geo.-IoU*	*Sem.-mIoU*	*Geo.-IoU*
Semantic-NeRF	54.6	3.9	49.5	2.5
Ours	**62.9**	**49.8**	**57.9**	**46.3**

Table 4: Comparison to Semantic-NeRF [45].

quantitative and qualitative comparisons, and want to highlight the generalization capability and crisp reconstruction of complicated indoor scenes by our approach (Figure 7).

Specifically, we have used the public implementation of Semantic-NeRF [45] to evaluate on 5 ScanNet scenes. For a fair comparison, we introduce additional depth supervision into [45] and use the same number of input views as ours. We have empirically found that Semantic-NeRF, with per-scene optimization, struggles to predict sharp 3D geometry for cluttered indoor scenes due to limited views and density-field representation. To further highlight the performance of semantic label fusion, we isolate the semantic evaluation of [45] from its underlying geometry, by projecting 2D semantic rendering to the perfect 3D geometry fused from ground-truth depths. As reported in Table 4, our approach achieves much better 3D segmentation accuracy.

4.3. Ablative Studies

The effect of pseudo-labeling. Substituting GT labels with machine-generated pseudo-GT in (3) moderately decreases performance while still outperforming [30] (Table 7; for a reference, we include a result obtained by training on 3D annotations); larger number of diverse virtual views results in narrower performance gap against GT labels (Table 8).

The effect of color supervision. As shown in Table 5, including color information in both either encoder or decoder is crucial to performance of our method.

Method	*mAcc*	*mIoU*
Ours w/o color encoder	40.9	22.4
Ours w/o color decoder	43.5	26.3
Ours	**44.8**	**28.6**

Table 5: The effect of using color encoder and decoder.

Model	*mAcc*	*mIoU*
Ours w/o completion	41.2	32.8
Ours	**48.3**	**36.7**

Table 6: The effect of using geometry completion.

Method	Matterport3d		ScanNet	
	mAcc	*mIoU*	*mAcc*	*mIoU*
Ours (3D GT)	**50.0**	**34.4**	**64.7**	**47.1**
Ours (2D GT)	44.8	28.6	50.8	31.2
VMFusion [30] (pseudo-GT)	24.9	17.6	33.9	20.5
Ours (2D pseudo-GT)	35.3	21.5	37.8	21.9

Table 7: The effect of using pseudo-GT labels.

Method	*Virtual Views*	*mAcc*	*mIoU*
Ours (2D pseudo-GT)	15	**37.4**	**23.1**
Ours (2D pseudo-GT)	5	35.3	21.5
Ours (2D pseudo-GT)	0	32.1	20.5

Table 8: The effect of the number of virtual views.

The effect of completion on segmentation. We remove the geometry completion head from our network to reveal a performance drop in semantic segmentation accuracy (Table 6), which we conclude indicates mutual benefits of joint reasoning of geometry and semantics, particularly on partial 3D data.

5. Conclusion

We have presented the first to date algorithm to learn geometry completion, scan colorization, and semantic segmentation in a single, self-supervised training procedure. Our approach to self-supervised learning builds on several crucial design choices, most importantly, an efficient multi-modal deep neural U-network with residual blocks, a differentiable rendering technique augmented to produce semantic maps, and progress in universal semantic pretraining. Fundamentally, our approach enables joint geometry and color reconstruction as well as semantic labeling for unseen scenes without requiring ground-truth labels, a stepping stone in building accurate, semantic models of real-world environments. We have additionally established that adding ground-truth information where it is available considerably improves semantic reconstruction performance; in fact, leveraging 3D ground-truth enables achieving state-of-the-art results.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 2

[2] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 324–333, 2021. 2, 3

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision*, pages 667–676. IEEE, 2017. 1, 2, 5, 6

[4] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In *European Conference on Computer Vision*, 2022. 2

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 2

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016. 4

[7] Jonathan Crespo, Jose Carlos Castillo, Oscar Martinez Mozos, and Ramon Barber. Semantic information for robot navigation: A survey. *Applied Sciences*, 10(2):497, 2020. 1

[8] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual Conference on Computer Graphics and Interactive Techniques*, pages 303–312, 1996. 4

[9] 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, 6

[10] Angela Dai, Christian Diller, and Matthias Nießner. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 849–858, 2020. 2, 3, 6

[11] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 452–468, 2018. 2

[12] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2018. 2, 3, 5, 6, 7

[13] Angela Dai, Yawar Siddiqui, Justus Thies, Julien Valentin, and Matthias Nießner. Spsg: Self-supervised photometric scene generation from rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1747–1756, 2021. 2, 3, 4, 5, 6

[14] Aloisio Dourado, Frederico Guth, and Teofilo de Campos. Data augmented 3d semantic scene completion with 2d segmentation priors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3781–3790, 2022. 2, 3

[15] Guoguang Du, Kai Wang, Shiguo Lian, and Kaiyong Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, 54(3):1677–1734, 2021. 1

[16] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv preprint arXiv:2203.15224*, 2022. 3

[17] Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. Learning 3d semantic segmentation with only 2d image supervision. In *International Conference on 3D Vision*, pages 361–372. IEEE, 2021. 2

[18] Yuxiao Guo and Xin Tong. View-volume network for semantic scene completion from a single depth image. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 726–732, 2018. 3

[19] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International journal of multimedia information retrieval*, 9(3):171–189, 2020. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[21] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4421–4430, 2019. 2

[22] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2098–2107, 2020. 2, 3

[23] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021. 2

[24] Wenbo Hu, Hengshuang Zhao, Li Jiang, Jiaya Jia, and Tien-Tsin Wong. Bidirectional projection network for cross dimension scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14373–14382, 2021. 2, 5, 6, 7

[25] Hiroharu Kato, Deniz Beker, Mihai Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon.

Differentiable rendering: A survey. *arXiv preprint arXiv:2006.12057*, 2020. 4

[26] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41, 2022. 2

[27] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9404–9413, 2019. 2

[28] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *Advances in neural information processing systems*, 24, 2011. 6

[29] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12871–12881, 2022. 3

[30] Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3d semantic segmentation. In *European Conference on Computer Vision*, pages 518–535. Springer, 2020. 2, 5, 6, 8

[31] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. MSeg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 6

[32] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *British Machine Vision Conference*, 2018. 2

[33] Jenny Lin, Xingwen Guo, Jingyu Shao, Chenfanfu Jiang, Yixin Zhu, and Song-Chun Zhu. A virtual reality platform for dynamic human-scene interaction. In *SIGGRAPH ASIA 2016 virtual reality meets physical reality: Modelling and simulating virtual humans and environments*, pages 1–4. 2016. 1

[34] Shengyi Qian, Alexander Kirillov, Nikhila Ravi, Devendra Singh Chaplot, Justin Johnson, David F Fouhey, and Georgia Gkioxari. Recognizing scenes from novel viewpoints. *arXiv preprint arXiv:2112.01520*, 2021. 3

[35] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 2, 3

[36] Peng-Shuai Wang, Yang Liu, and Xin Tong. Deep octree-based cnns with output-guided skip connections for 3d shape and scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 266–267, 2020. 3

[37] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8608–8617, 2019. 2, 3

[38] Zongji Wang and Feng Lu. Voxsegnet: Volumetric cnns for semantic part segmentation of 3d shapes. *IEEE transactions on visualization and computer graphics*, 26(9):2919–2930, 2019. 2

[39] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *European conference on computer vision*, pages 574–591. Springer, 2020. 2

[40] Yuxing Xie, Jiaojiao Tian, and Xiao Xiang Zhu. Linking points with labels in 3d: A review of point cloud semantic segmentation. *IEEE Geoscience and Remote Sensing Magazine*, 8(4):38–59, 2020. 2

[41] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 4

[42] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *Proceedings of the European Conference on Computer Vision*, pages 733–749, 2018. 3

[43] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7801–7810, 2019. 3

[44] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021. 2

[45] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 3, 7, 8