WSSIC-Net: Weakly-Supervised Semantic Instance Completion of 3D Point Cloud Scenes

Zhiheng Fu, Yulan Guo, *Senior Member, IEEE*, Minglin Chen, *Graduate Student Member, IEEE*, Qingyong Hu, Hamid Laga, Farid Boussaid, and Mohammed Bennamoun, *Senior Member, IEEE*

***Abstract*— Semantic instance completion aims to recover the complete 3D shapes of foreground objects together with their labels from a partial 2.5D scan of a scene. Previous works have relied on full supervision, which requires ground-truth annotations, in the form of bounding boxes and complete 3D objects. This has greatly limited their real-world application because the acquisition of ground-truth data is very costly and time-consuming. To address this bottleneck, we propose a Weakly-Supervised Semantic Instance Completion Network (WSSIC-Net), which learns real-world partial point cloud object completion without requiring the ground truth of complete 3D objects. Instead, WSSIC-Net leverages 3D ground-truth bounding boxes, partial objects of a raw scene, and unpaired synthetic 3D point clouds. More specifically, a 3D detector is used to encode partial point clouds into proposal features, which are then fed into two branches. The first branch uses fully supervised box prediction based on proposal features. The second branch, hereinafter called instance completion, leverages the proposal features as partial object features to achieve weakly-supervised instance completion. A Generative Adversarial Network (GAN) completes the partial features of the 2.5D foreground objects of real-world scenes using only unpaired but semantically-consistent complete synthetic point clouds. In our experiments, we demonstrate that the fully-supervised 3D detection and the weakly-supervised instance completion complement one another. The qualitative and quantitative evaluations on the ScanNet v2 dataset demonstrate that the proposed "weakly-supervised" approach consistently achieves comparable performance to the state-of-the-art "fully supervised" methods.**

***Index Terms*— Weak supervision, point cloud, 3D completion.**

Received 2 December 2022; revised 17 December 2023 and 27 August 2024; accepted 9 December 2024. Date of publication 27 March 2025; date of current version 1 April 2025. This work was supported in part by Australian Research Council under Grant ARC DP210101682 and Grant ARC DP220102197, in part by the National Natural Science Foundation of China under Grant U20A20185 and Grant 62372491, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2022B1515020103 and Grant 2023B1515120087, and in part by Shenzhen Science and Technology Program under Grant RCYX20200714114641140. The associate editor coordinating the review of this article and approving it for publication was Prof. Kui Jia. *(Corresponding author: Yulan Guo.)*

Zhiheng Fu and Mohammed Bennamoun are with the Department of Computer Science and Software Engineering, The University of Western Australia (UWA), Perth, WA 6009, Australia.

Yulan Guo and Minglin Chen are with the School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen Campus, Shenzhen 518107, China (e-mail: guoyulan@sysu.edu.cn).

Qingyong Hu was with the Department of Computer Science, University of Oxford, OX1 3PG Oxford, U.K. He is now an Independent Researcher.

Hamid Laga is with the School of IT, Murdoch University, Murdoch, WA 6150, Australia.

Farid Boussaid is with the Department of Electrical, Electronic and Computer Engineering, The University of Western Australia (UWA), Perth, WA 6009, Australia.

Digital Object Identifier 10.1109/TIP.2024.3520013

I. Introduction

3D SCENE understanding from partial scans is a complex problem due to occlusions, viewpoint constraints, and poor lighting [1], [2], [3], [4], [5]. Based on partial observations (*e.g.*, depth images or point clouds) of the scene, semantic instance completion aims to recover the semantic labels of voxels/points, as well as the poses and 3D geometries of objects (Fig. 1 (a)). Semantic scene completion is particularly important in applications such as robot navigation, human-robot interaction, and object grasping [6], [7], [8], [9], [10].

Several strategies have been explored in the literature to recover missing 3D shapes: **(1) Completion-from-scans:** A 3D convolutional encoder-decoder architecture utilizing 2.5D partially scanned surfaces is used to predict the voxel occupancy and the object category of voxels of observed partial objects and occluded regions [13], [14], [15], [16], [17], [18]. However, due to their reliance on 3D convolutions at the scene level, they can only recover the 3D scene with a low resolution. **(2) Completion-from-retrieval:** Shape retrieval methods search for CAD models that can match the incomplete object [19], [20], [21], [22], [23]. The computation efficiency and accuracy depend on the size of the model dataset. **(3) Completion-from-objects:** In this category, the scene is regarded as a collection of multiple objects, and a reconstruction-from-detection strategy is used: 3D object detection (where objects are labeled, and their poses and positions are determined) and instance completion (where the 3D models of the objects are fully reconstructed) [11], [24]. By using 3D detection of objects, this class of methods completes the partial scan and reconstructs the details of those objects. As opposed to completion-from-scans, completion-from-objects methods require ground truth for bounding boxes. Additionally, most existing methods are **fully supervised** approaches, which limits their use in real-world settings.

Rendering synthetic but highly realistic 3D structures is much easier than capturing them in a real-world setting. A widely accepted observation is that 3D objects with the same category tend to have similar geometrical 3D shapes. In addition, a labelling process requiring 3D bounding boxes and semantic labels of objects is often simpler in practice than producing a complete 3D point cloud of the scene for ground truths. In this regard, we aim to construct "pseudo" ground truths of 3D shapes using the same number of complete synthetic objects of the same category as the real-world partial scans to enable semantic scene completion only under the supervision of partial objects, bounding boxes, and "pseudo"

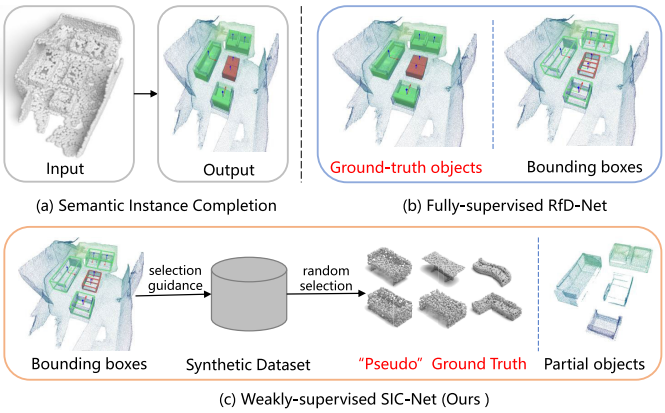


Fig. 1. Overview of semantic instance completion and existing supervision strategies. **(a)** In semantic instance completion, the input is a partial 2.5D point cloud with missing data points, and the output consists of fully recovered 3D foreground objects with their labels, positions, and poses, which can be observed from any viewpoint. **(b)** Existing methods like RfD-Net [11] require both corresponding complete objects and ground-truth bounding boxes, involving a complex and time-consuming labeling process. **(c)** In contrast, our proposed WSSIC-Net introduces three supervision signals: **(1)** ground-truth bounding boxes, representing a more reasonable labeling process, are used for 3D box prediction and classification (with different colors representing different categories); **(2)** "pseudo" ground truth, randomly selected from a synthetic dataset based on the number and category of ground-truth bounding boxes (e.g., two tables and four sofas), is employed for coarse 3D shape completion; and **(3)** partial objects, segmented using ground-truth bounding boxes, are incorporated in the Hausdorff distance loss (HL) [12] for 3D shape refinement. Our WSSIC-Net alleviates the need for the ground truth of complete 3D objects.

ground truth (Fig. 1 (c)). Thus, the proposed method does not require the ground truth of complete 3D objects. In contrast, existing state-of-the-art methods require complete RGB-D scans as ground truth (RevealNet [24]) or use watertight meshes as ground truth (RfDNet [11]) (Fig. 1 (b)).

In this paper, we adopt a reconstruction-from-detection strategy to first locate the positions of objects using bounding boxes (*i.e.*, based on a supervised approach), then perform a weakly-supervised mesh reconstruction of the objects. More specifically, using a 3D detector backbone, we first extract proposal features of partial objects from a raw point cloud of the scene. The proposal features are then fed to two branches. The first branch uses a dedicated detection module to detect 3D objects in a supervised manner. The second branch uses a weakly-supervised approach to achieve instance completion. In the instance completion branch, proposal features of partial objects are completed under the supervision of "complete" latent features that are learned by a pre-trained encoder on synthetically complete point cloud objects using a Generative Adversarial Network (GAN) and the unidirectional Hausdorff distance loss (HL) [12]. The completed proposal features are then fed to a pre-trained decoder in order to obtain a complete meshes of objects. Additionally, we propose an Objectness-aware Chamfer Distance (OCD) loss to jointly optimize the 3D detection module and the instance completion module. In the OCD loss, objectness scores are predicted in the detection branch to identify the probability of having an object in a proposal and the chamfer distance is widely used in point cloud completion. Experiments on the ScanNet v2 Dataset [25] demonstrate that the proposed weakly-supervised semantic instance completion method achieves comparable performance to the **fully supervised** state-of-the-art RfD-Net [11] both in the case of 3D detection and instance reconstruction tasks on raw point clouds.

The main contributions of this paper are as follows:

- We propose a Weakly-Supervised Semantic Instance Completion Network (WSSIC-Net) for scene-level point cloud completion. To minimize the reliance on manually generated ground truth, we use synthetically completed point clouds of objects as "pseudo" ground truth for real-world partial point clouds of objects.
- We propose an OCD loss for joint task optimization of 3D detection and point cloud instance completion. Our results demonstrate that the instance completion task, supervised by unpaired complete synthetic point clouds of objects, can significantly boost object detection performance in real-world environments.

The rest of this paper is organized as follows. We briefly review related works in Section II. We describe the proposed network in detail in Section III. Extensive experimental results are presented in Section IV. We conclude this paper and present future work in Section V.

II. Related Works

In this section, we provide a brief overview of the related literature for 3D point cloud object detection, 3D object shape completion, generative adversarial networks for 3D completion, and semantic 3D scene completion.

A. 3D Point Cloud Object Detection

The goal of 3D point cloud object detection is to detect bounding boxes around objects and identify their classes. Due to the unstructured and irregular nature of 3D point clouds, early detection approaches [26], [27], [28], [29], [30], [31], [32] first project the 3D point clouds onto regular 2D grids, before applying mature 2D detection pipelines. The incorporation of associated RGB images was investigated as a way to improve the detection performance [28], [33], [34]. However, most of these approaches rely on predefined anchors and a two-stage region proposal network, leading to high computational costs [35]. Without relying on anchors, PointRCNN [36] can learn to detect objects via foreground point segmentation, while VoteNet [37] can detect objects via point feature grouping, sampling, and voting. Shi et al. [38] combined voxel representation and point representation for 3D object detection. Zhang et al. [39] introduced an instance-aware downsampling strategy for efficient 3D object detection. Yang et al. [40] directly regressed 3D object bounding boxes from the compact global features through a single forward pass. Feng et al. [41] proposed a relation graph network that comprises a 3D object proposal generation module and a 3D relation module, making it an end-to-end trainable network for detecting 3D objects in point clouds. More recently, Misra et al. [42] proposed the end-to-end transformer (3DTR) with non-parametric queries and Fourier positional embeddings. Liu et al. [43] proposed the transformer-based detection framework Group-free to group foreground points for object detection. Mao et al. [44] proposed the Voxel Transformer to effectively capture the long-range relationships between voxels for 3D object detection. Cai et al. [45] proposed a cascade architecture,

named 3D Cascade RCNN, that applies a completeness-aware re-weighting strategy to multiple detectors based on the voxelized point clouds in a cascade paradigm. Wang et al. [46] proposed the Bridged Transformer (BrT) for 3D object detection with both RGB images and point clouds as input. Recently, 3D object detection with limited annotation is also explored in [47], [48], [49], and [50].

In this paper, 3D object detection is an essential module for extracting features from partial point cloud objects, and instance completion is our ultimate goal. In order to fairly compare with state-of-the-art RfD-Net [11], we used VoteNet [37] (also used with RfD-Net) as our 3D detector.

B. 3D Object Shape Completion

In this task, the missing geometries of a specific object are recovered from partial scans. Yuan et al. [51] proposed an end-to-end PCNnet that directly operates on raw point clouds without any structural assumption about the underlying shape. Currently, point-based object shape completion works [52], [53], [54], [55], [56] focus on reconstructing fine-detailed 3D geometry. Groueix et al. [57] proposed AtlasNet to represent a 3D shape as a collection of parametric surface elements and to generate a shape of arbitrary resolution without memory limitations. In encoder-decoder architectures, ASFM-Net [58] and VRCNet [59] match encoded latent features with completion shape priors, leading to good completion results. To preserve the observed geometry of the partial scan for the fine reconstruction, MSN [53] and VRCNet [59] bypass the observed geometries by either using the Minimum Density Sampling (MDS) or the Farthest Point Sampling (FPS) from the observed surface and building skip connections. By embedding a volumetric sub-architecture, GRNet [60] preserves the discretized input geometries with the volumetric U-connection without sampling in the point cloud space. In more recent works, PMP-Net [61] reconstructs the entire object from the observed partial object to the nearest occluded regions in a gradual manner. By converting partial scan proxies into a set of occluded proxies, PoinTr [62] is one of the first transformer-based frameworks to achieve point cloud completion and refine reconstructions by predicting only occluded geometries. Subsequently, Wang et al. [63] proposed an adaptive downsampling and upsampling approach to achieve detailed point cloud completion.

Compared to other methods, AtlasNet [57] can produce complete 3D shape of arbitrary resolution using a fixed number of ground-truth point clouds. Besides, AtlasNet [57] can achieve mesh reconstruction from point clouds without the requirement of water-tight 3D meshes. Thus, we adopted AtlasNet [57] for instance reconstruction to transform the partial proposal features extracted from a partial point cloud scan to generate a complete 3D shape of arbitrary resolution.

C. Generative Adversarial Network for 3D Completion

Yang et al. [64] and Wang et al. [65] combined 3D-CNN and generative adversarial network to complete shapes in a supervised manner. Gurumurthy and Agrawal [66] treated the point cloud completion task as a denoising Auto-Encoder (AE) problem. Adversarial training was used in order to optimize the AE latent space. Wen et al. [67] proposed Cycle4Completion which uses two synchronous cycle transformations between the latent spaces of complete 3D shapes and incomplete 3D shapes to learn from the complete shapes. Zhang et al. [68] propose a new GAN inversion method called ShapeInversion to integrate GANs into shape completion. By searching for a latent code to reconstruct a partial scan input. ShapeInversion uses a GAN that has been pre-trained on complete shapes to reconstruct the best complete 3D shape for a given input. Li and Baciu [69] proposed HSGAN that takes a random code and hierarchically transforms it into a representation graph by incorporating both Graph Convolution Network (GCN) and self-attention. Li et al. [70] introduced PU-GAN, an up-sampling network based on GAN, for learning point distributions in latent space. Points can also be up-sampled over patches on the surface of the object using PU-GAN [71]. Through hierarchical and interpretable sampling, PC-GAN combines hierarchical Bayesian networks and implicit generative architectures. Cheng et al. [72] proposed an end-to-end generative adversarial network-based dense point cloud completion architecture (DPCG-Net). Through the development of two generative adversarial networks based modules, the DPCG-Net translates point cloud completion into mapping between global feature distributions obtained by encoding partial point clouds and ground truth.

In this work, inspired by ShapeInversion [68], we leverage an auto-encoder pre-trained on synthetically complete shapes and a GAN by searching for a latent code to find the complete shape that best reconstructs a given real-world 3D partial scan input. As opposed to ShapeInversion [68], our proposed method is designed to complete real-world point cloud scenes, which entails significant challenges outlined in Section III-C.

D. Semantic 3D Scene Completion

The goal of this task is to recover a complete 3D representation of volumetric occupancy/point clouds with semantic labels for the objects in the scene. SSCNet [13] is a pioneering work combining these two tasks in an end-to-end fashion. ESSCNet [73] introduced Spatial Group Convolution (SGC) to divide the input volume into different groups, prior to applying 3D sparse convolution. An efficient method for reducing computational cost and extracting features from multi-channel input is VVNet [74], which combines 2D and 3D CNNs with a differentiable projection layer. The multi-branch architecture of ForkNet [75] leverages the idea of generative models to sample new pairs of training data, further alleviating the problem of limited training samples of real-world scenes. A self-cascaded context aggregation network called CCPNet was proposed by Zhang et al. [76] to reduce semantic gaps in multi-scale 3D contexts and incorporate local geometric details. In [77], a geometric embedding mechanism based on depth information was proposed for predicting 3D sketches and guiding the reconstruction based on the embedded features. Chen et al. [77] proposed SketchSSC to learn the 3D boundary of all objects in the scene to quickly estimate the resolution of the invariant features. Using joint color and geometry feature learning, Hou et al. [24] propose RevealNet, a 3D neural network architecture that detects individual object instances and infers their full object geometry based on an

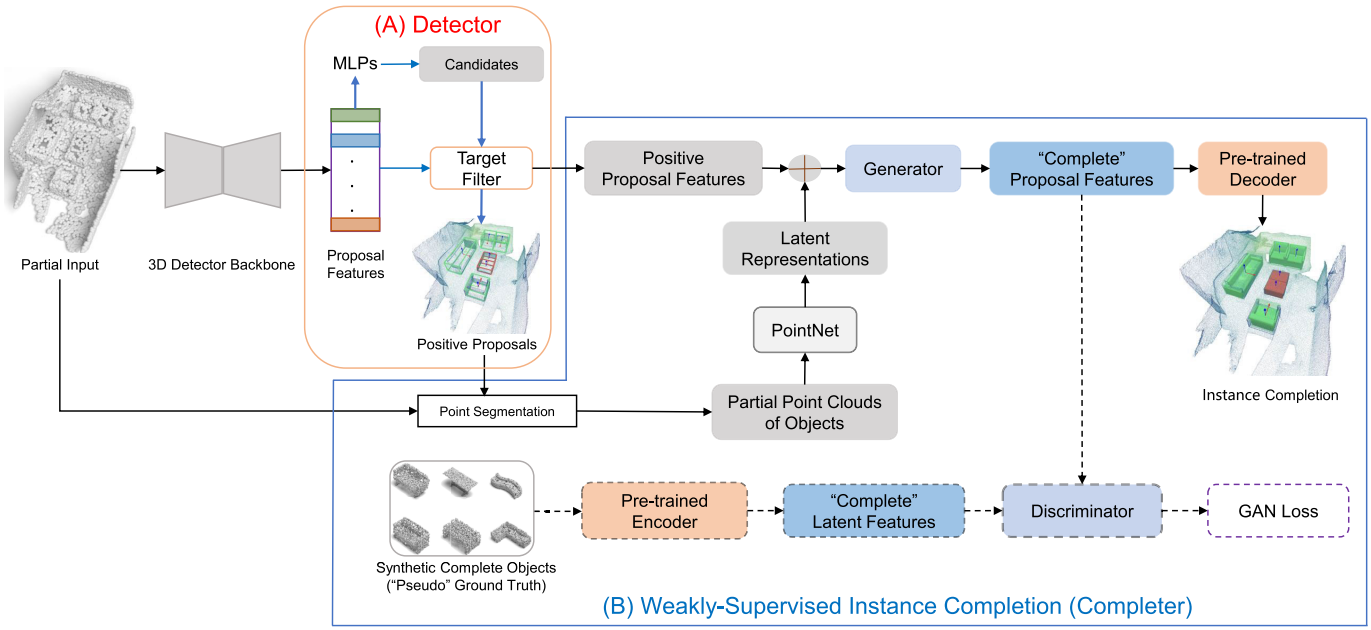


Fig. 2. The framework for weakly-supervised point cloud instance completion, composed of (A) a detector module for identifying and generating proposal features, followed by (B) a completion module that refines these features into complete instances. This method employs a pre-trained decoder, a discriminator for GAN loss, and integrates "pseudo" ground truth during training. It aims to replicate the performance of fully supervised 3D object detection and completion without extensive labeling.

incomplete RGB-D scan of the scene. In constrast to previous volumetric semantic scene completion, Nie et al. [11] proposed RfD-Net, which uses the VoteNet [37] to detect 3D objects and utilizes implicit function to achieve per-instance reconstruction from an incomplete point cloud of a scene.

The goal of this work is also to achieve semantic instance completion from raw point cloud scenes. As opposed to RfD-Net [11], our proposed method is **weakly-supervised** since it only requires bounding boxes to complete point cloud scenes.

III. Proposed WSSIC-Net Framework

A. Overview

We propose a weakly-supervised instance completion framework for raw point clouds (see Fig. 2) to alleviate the need for the ground truth of partial scans. Given a partial point cloud scan $\mathbf{P}_{in} \in \mathbb{R}^{N\times 3}$ as input, a 3D detector backbone is used to encode each partial point cloud scene into proposal features, which are then fed to a fully supervised 3D detection module (Detector) and a weakly-supervised instance completion module (Completor). **1) For the 3D detection**, the MLPs are used to regress the parameters of bounding boxes using a fully supervised approach (the details are presented in Section III-B). **2) The point cloud completion of objects** is accomplished using the proposed weakly-supervised approach. More specifically, an auto-encoder is first pre-trained on synthetically complete point clouds of objects. We propose to use the predicted objectness scores to achieve target filter to reduce the influence of negative proposal features. To reduce the influence of background points on positive proposal features, the positive proposal features are concatenated with the latent representations of segmented points from the positive proposals. The concatenated proposal features are then transformed into "complete" proposal features using a Generative Adversarial Network (GAN) under the supervision of the GAN loss and the unidirectional Hausdorff distance loss (HL) [12]. Finally, the generated complete proposal features are fed to a pre-trained decoder to generate complete point clouds of objects. **3) For the joint training of 3D detection and instance completion**, an Objectness-aware Chamfer Distance (OCD) loss is proposed to jointly optimize the 3D detection module and the instance completion module. The pre-trained encoder and discriminator modules are only used here during the training phase.

B. Fully Supervised Object Detection in 3D Point Clouds

We first train a 3D detector to detect 3D objects from partial 3D point clouds $\mathbf{P} \in \mathbb{R}^{N\times 3}$. In particular, our framework uses the raw point cloud scans to generate N proposal features $\mathbf{F} \in \mathbb{R}^{N\times D}$. Next, the parameters of box candidates (center $\mathbf{c} \in \mathbb{R}^3$, scale $\mathbf{s} \in \mathbb{R}^3$, orientation angle $\theta \in \mathbb{R}$, class label l, and objectness score O_s) are estimated based on $\mathbf{F}$. Finally, the target filter is used to remove redundant candidates to obtain positive proposals in the training and inference processes, respectively. More specifically, the features $\mathbf{F} \in \mathbb{R}^{N\times D}$ are learned by aggregating the local geometry around the partial objects to capture the semantics and geometry information of the local region. Then, the detector head composed of MLPs is used to regress these parameters. Among them, the objectness score is used to identify whether there is an object in the predicted box according to the $\mathcal{L}_1$ distance between this box and any ground truth center. Note that, we directly leverage the ground-truth bounding boxes as the supervision signals for this module.

C. Proposed Weakly-Supervised Instance Completion

The proposal features (partial representations of partial objects) $\mathbf{F} \in \mathbb{R}^{N\times D}$ are fed to the proposed weakly-supervised coarse-to-fine instance completion module, which transforms them into "complete" proposal features, as shown in Fig. 2. To achieve this, an auto-encoder on the complete synthetic

Fig. 3. Illustration of the impact of the HL loss. With a GAN loss, only the general shape of a category can be predicted. The HL loss can adjust the general shape of a category to a specific 3D shape that is as similar as possible to the input partial object.

dataset is firstly pre-trained in a self-supervised manner. We specifically used AtlasNet [57] as the auto-encoder framework. This choice was made due to AtlasNet's ability to reconstruct full 3D mesh models without the prerequisite of water-tight 3D input. The pre-training of AtlasNet was carried out on the ShapeNetCore dataset [78]. The outputs of the encoder and decoder are "complete" latent features $\mathbf{C}_f \in \mathbb{R}^{M\times 1024}$ and full-shape point clouds of objects $\mathbf{P}_c \in \mathbb{R}^{2500\times 3}$, respectively. After that, the encoder and decoder are used at different stages of the pipeline.

1) General Shape Generation: In order to include more geometric information into the proposal features and reduce the influence of background points on the proposal features, we propose to segment point clouds of partial objects $\mathbf{P}_p \in \mathbb{R}^{2048\times 3}$ from the predicted positive proposals and encode those point clouds of partial objects into latent features $\mathbf{Obj}_f \in \mathbb{R}^{M\times 1024}$ using the PointNet [79]. Then, the latent features $\mathbf{Obj}_f \in \mathbb{R}^{M\times 1024}$ are concatenated with the positive proposal features in order to form the refined proposal features. Subsequently, the refined proposal features are fed to a generator, which generates the generally "complete" proposal features under the supervision of "pseudo" ground truth and the GAN loss. Finally, the generated generally"complete" proposal features $\mathbf{C}_g \in \mathbb{R}^{M\times 1024}$ are fed to a pre-trained decoder to obtain the complete point clouds of objects $\mathbf{P}_{re} \in \mathbb{R}^{M\times 2500\times 3}$ of corresponding categories.

2) Synthetic-to-Real-World Point Cloud Completion: As a result of the GAN loss, the generated complete point clouds of objects correspond to general 3D structures of each category (see Fig. 3 of Section IV-E). The 3D general structures are then adjusted to the unique shapes of the corresponding objects in the partial scene. This is achieved by applying the unidirectional Hausdorff distance loss (HL) [12] from the partial input to its completion. The GAN loss and the unidirectional Hausdorff distance loss are used to generate generally and specifically latent representations, respectively.

Although this framework can greatly reduce dependence on the ground truth of complete 3D objects, several challenges remain.

- **Geometrical alignment.** Normally, synthetic objects have fixed orientations and sizes, while real-world objects have highly variable orientations and scales, as seen in Fig. 2.
- **Negative Proposals.** Instance completion is based on the object detection output. When training a 3D object detector in an end-to-end manner, it is preferable to use a fixed and larger number of estimated objects from a partial point cloud scene. This, however, leads to negative proposals, which have a negative impact on the instance completion.
- **Synthetic-to-real Domain Alignment.** Synthetic point clouds of objects may differ greatly from real-world point clouds of objects in terms of data distribution.
- **Semantic Alignment.** Without the ground truth of complete point cloud scenes, how to build correspondences between partial objects of input scenes and synthetic objects is an important but challenging task.

In the following part of this section, we present the detailed strategies used to overcome the aforementioned challenges. We will demonstrate the effectiveness of these strategies in the experimental part (see Section IV).

- **Geometrical Alignment.** Two data augmentation strategies are proposed to train the auto-encoder when partial point clouds are not geometrically aligned with complete synthetic point clouds. **1)** A data augmentation system is proposed for each category based on the orientation and size of objects in partial scenes according to ground-truth bounding boxes. **2)** To better mimic the distribution of real-world 3D scans, we then randomly select synthetically complete point clouds of objects with the same number and category as the input of the 3D detection for the training of the auto-encoder. In this case, when a partial point cloud scene and its "pseudo" ground truth are fed to the network, the pre-trained encoder with data augmentation can better capture the orientation and size information of objects. Thus, the pre-trained decoder can decode complete objects with similar orientations and sizes to the partial objects. To further achieve geometrical alignment, the unidirectional Hausdorff distance Loss (HL) [12] (from the partial input to its completion) was adopted here to transform the general 3D structures in the bounding boxes into unique shapes of the corresponding objects in the partial scene.
- **Negative Proposals.** We propose to leverage both the training and inference processes to reduce the impact of negative proposals. During training, the negative proposal features $\mathbf{F} \in \mathbb{R}^{N\times D}$ are filtered using the prediction of objectness score. By integrating the predicted objectness score of each box proposal into the Chamfer Distance (CD) [80], an OCD loss function was proposed to globally regularize the distance between the predicted complete scene and the partial input (see Section III-E). At inference, the 3D NMS (Non-Maximum Suppression) [37] was used to remove the redundant box proposals and points in these redundant boxes.
- **Synthetic-to-real Domain Alignment.** In order to tackle the synthetic-to-real domain alignment challenge, we propose a GAN-based approach. Firstly, the latent representations $\mathbf{Obj}_f \in \mathbb{R}^{M\times 1024}$ of point clouds of objects segmented from positive proposals are learned using PointNet [79], which has the same structure as that of the pre-trained encoder. The latent representations $\mathbf{Obj}_f \in \mathbb{R}^{M\times 1024}$ of point clouds of objects are combined with the positive proposal features $\mathbf{F}_p \in \mathbb{R}^{M\times D}$ to form the refined proposal features $\mathbf{R}_p \in \mathbb{R}^{M\times(D+1024)}$. In this way, the object-level geometric information is explicitly integrated into the positive proposal features $\mathbf{F}_p \in \mathbb{R}^{M\times D}$.

Next, the refined proposal features $\mathbf{R}_p \in \mathbb{R}^{M\times(D+1024)}$ are fed to the generator to obtain "complete" proposal features $\mathbf{C}_g \in \mathbb{R}^{M\times1024}$. Subsequently, $\mathbf{C}_g \in \mathbb{R}^{M\times1024}$ and the "complete" latent features $\mathbf{C}_f \in \mathbb{R}^{M\times1024}$ were fed to the discriminator to predict "false" and "true" "complete" features, respectively. In this way, the domain gap between the positive proposal features $\mathbf{F}_p \in \mathbb{R}^{M\times D}$ of partial point clouds and the "complete" latent features $\mathbf{C}_f$ of "pseudo" ground truth is narrowed.

- **Semantic Alignment.** The semantic alignment between the partial point clouds of objects and their ground truth of complete 3D objects are aligned. However, the semantic alignment between the partial point clouds of objects and the synthetic complete objects needs to be carefully designed. In this part, we build a "pseudo" ground truth from the synthetic dataset for each partial point cloud scan by randomly selecting complete point clouds of objects with the same number and category as the detection ground truth of each partial point cloud scan. In this way, the pre-trained decoder can decode complete objects, which have semantic correspondence to the partial objects.

D. Joint Task Learning

The dedicated 3D detector aims to locate the positions of partial objects while the instance completor aims to recover the full-shapes of the corresponding partial objects. A more complete partial object is more likely to show better performance in 3D object detection. Similarly, the estimated orientations and sizes of the bounding boxes that are generated by the 3D detection algorithm significantly affect the performance of the point cloud completion branch. In this paper, we propose an Objectness-aware Chamfer Distance (OCD) loss to jointly optimize 3D detection and point cloud completion. The details of the proposed OCD loss are given in Section III-E.

E. Loss Functions

In this section, we summarize the learning targets and corresponding loss functions.

1) Detection Loss: The 3D detector predicts the center $\mathbf{c} \in \mathbb{R}^3$, scale $\mathbf{s} \in \mathbb{R}^3$, orientation angle $\theta \in \mathbb{R}$, class label l, and objectness score O_s. Following RfD-Net [11], we supervise the objectness scores O_s for votes that are located either close to a ground truth object center (within 0.3 meters) or far from any center (by more than 0.6 meters). We consider proposals generated from those votes as positive and negative proposals, respectively. Objectness predictions for other proposals are not penalized. Objectness is supervised via a cross entropy loss normalized by the number of non-ignored proposals in the batch. In the case of positive proposals, we further supervise the bounding box estimation and the class prediction based on the closest ground truth bounding box. We supervise the box center loss $\mathcal{L}_c$ using Smooth-L1 loss, and follow the works in [37] and [81] to disentangle the scale loss $\mathcal{L}_s$ and the heading angle loss $\mathcal{L}_\theta$ into a hybrid of classification (cross entropy) and regression (Smooth-L1) losses. The standard cross entropy loss was used for semantic classification. We also require an extra vote loss for using VoteNet as the backbone [37]. For the weights of the 3D detection loss function, we set $\lambda_{cls} = 0.1$, which is the hybrid ratio combining the classification and regression losses, *i.e.*, $\lambda_{cls}\mathcal{L}_{cls} + \mathcal{L}_{reg}$. The box loss $\mathcal{L}_{det}$ can thus be denoted as:

$$\mathcal{L}_{det} = \mathcal{L}_v + \mathcal{L}_c + \mathcal{L}_s + \mathcal{L}_\theta + \lambda_l\mathcal{L}_l + \lambda_{obj}\mathcal{L}_{obj}. \quad (1)$$

where $\lambda_{obj} = 0.5$ and $\lambda_l = 0.1$. $\mathcal{L}_v$, $\mathcal{L}_c$, $\mathcal{L}_l$ and $\mathcal{L}_{obj}$ are the vote loss, the box center loss, the classification loss and the objectness loss, respectively.

2) Completion Loss: A GAN was used to train our proposed weakly-supervised instance completion model. Given the input of the GAN, including the "complete" latent features $\mathbf{C}_f \in \mathbb{R}^{M\times1024}$ and the generated "complete" proposal features $\mathbf{C}_g \in \mathbb{R}^{M\times1024}$, we seek to optimize the following completion loss $\mathcal{L}_{com}$ over the generator G_θ and a discriminator D_χ.

$$\mathcal{L}_{com} = \min_\theta \max_\chi \mathbb{E}[log(D_\chi(\mathbf{C}_f))] + \mathbb{E}[log(1 - D_\chi(G_\theta(\mathbf{C}_g)))]. \quad (2)$$

In our experiments, we designed a GAN based on the least square GAN [82], which is easier to train. We hence minimized both the discriminator and generator losses defined as:

$$\mathcal{L}_{D_\chi} = \mathbb{E}[D_\chi(\mathbf{C}_f) - 1]^2 + \mathbb{E}[D_\chi(G_\theta(\mathbf{C}_g))]^2. \quad (3)$$

$$\mathcal{L}_{G_\theta} = \mathbb{E}[D_\chi(G_\theta(\mathbf{C}_g)) - 1]^2. \quad (4)$$

$\mathcal{L}_{G_\theta}$ can only guarantee that the point cloud predicted from the pre-trained decoder is a complete object of a category. However, it cannot guarantee that the complete prediction will match the partial input in shape. Following [83], we used the unidirectional Hausdorff distance Loss (HL) $\mathcal{L}_{rec}^{HL}$ (from the partial input to its matching complete object) to adjust the complete prediction to be as close as possible to the input. Thus, the generator loss can be reshaped as:

$$\mathcal{L}_{G_\theta} = \alpha\mathbb{E}[D_\chi(G_\theta(\mathbf{C}_g)) - 1]^2 + \beta\mathcal{L}_{rec}^{HL}(S_{p_i}, D_{ec}(\mathbf{C}_{g_i})). \quad (5)$$

where S_{p_i} is the i-th partial point clouds of an object, $\mathbf{C}_{g_i}$ is the i-th refined latent feature and D_{ec} is the pre-trained decoder. In addition, we follow [83] to set the parameters α and β as 0.25 and 0.75, respectively.

3) Objectness-Aware Chamfer Distance Loss: All points are treated equally by the bidirectional distance and are assigned the same weight [80]. In contrast, according to the predicted objectness scores, we assigned different weights to points belonging to different box proposals, as shown in Eq. 6. Using this approach, we can optimize both object detection and instance completion at the same time.

$$\mathcal{L}_{OCD} = \mathbf{S}_{obj}\mathbf{CD}(\mathbf{P}_{re}, \mathbf{P}_{in_{obj}}). \quad (6)$$

where $\mathbf{S}_{obj}$, $\mathbf{P}_{re}$, $\mathbf{P}_{in_{obj}}$ and $\mathbf{CD}(\mathbf{P}_{re}, \mathbf{P}_{in_{obj}})$ are the objectness score of points, predicted point clouds of objects, input partial point clouds of objects and bidirectional CD loss. As a result of this loss, we can jointly optimize the network.

4) Total Loss: Finally, the total loss $\mathcal{L}_{total}$ includes the detection loss $\mathcal{L}_{det}$, the completion loss $\mathcal{L}_{com}$ and the OCD loss $\mathcal{L}_{OCD}$:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \mathcal{L}_{com} + \mathcal{L}_{OCD}. \quad (7)$$

IV. Experiments

A. Datasets

1) The ScanNet v2 Dataset [84]: ScanNet v2 [84] consists of 1,513 real-world scans with point clouds labeled at the instance level. Following RevealNet [24] and RfD-Net [11], we also expect to achieve scene completion on the ScanNet v2 dataset. In contrast, in our experiments, only the ground-truth bounding boxes are used in the training procedure.

2) The ShapeNetCore Dataset [78]: The ShapeNetCore Dataset [78] was used as the synthetic dataset. It is a subset of the full ShapeNet dataset with single clean 3D models and manually verified category and alignment annotations. It covers 55 common object categories with about 51,300 unique 3D models. In our experiments, we selected 8 categories out of the 55 (display, bathtub, trashbin, sofa, chair, table, cabinet, and bookshelf) for the pre-training of the auto-encoder. These categories are the same as the ones in the ScanNet v2 dataset [84].

3) The Scan2CAD Dataset [19]: Scan2CAD [19] aligns the ShapeNet [85] models with the object instances in ScanNet. In our experiment, the Scan2CAD dataset is only used in the testing stage to evaluate the performance of the proposed method.

B. Evaluation Metrics

We evaluated our method on the proposed method, including 3D object detection and instance completion. Similar to RfD-Net [11] and RevealNet [24], when evaluating 3D detection and instance completion, the mean average precision at the 3D IoU threshold of 0.5 (mAP@0.5) was used.

C. Implementation Details

We implemented our approaches with PyTorch, and performed training with NVIDIA GeForce GTX 1080 Ti GPUs. In this paper, we employ AtlasNet [57] as the auto-encoder framework due to its ability to reconstruct complete 3D meshes without the necessity for water-tight 3D data. Initially, AtlasNet was pre-trained on the ShapeNetCore dataset [78] for a total of 200 epochs, with the optimization process being carried out by the Adam optimizer, with a learning rate of 0.001. The input to AtlasNet consisted of 2500 points. During the pre-training phase, we applied two types of data augmentation: one that varied the orientations and sizes of the objects, and another that involved randomly selecting synthetically complete point clouds that correspond to the same number and category of ground-truth bounding boxes within each batch. For instance, if a batch contained a scene with two partial chairs and three tables, we would randomly choose two complete chairs and three complete tables from the synthetic dataset to form a "pseudo" ground truth. Following the pre-training of the auto-encoder, the encoder and decoder parts of AtlasNet, with their weights now fixed, were utilized in the weakly-supervised instance completion process. For the segmentation of the point clouds from the positive proposals, we grouped points located within a radius r of these box centers using a group layer as in [86].

For the training of the 3D detection and instance completion, we randomly sampled 80K points from a partial scanned scene of the ScanNet v2 dataset as input. Then, we followed four steps to train the proposed WSSIC-Net. **First**, to make fair comparison with RfD-Net [11], we adopted the same 3D object detection structure and training strategy to train the 3D detector as RfD-Net [11]. Further, the pre-trained 3D detection can provide object positions, which can be leveraged by our proposed weakly-supervised instance completion model. **Second**, we froze the weights of the detection module and used the pre-trained encoder and decoder of AtlasNet to train the generator (3-layer perceptrons) and the discriminator (3-layer perceptrons). **Third**, we jointly trained the entire structure *without* the proposed OCD loss. **Forth**, we jointly trained the entire structure *with* the proposed OCD loss to further improve its performance.

TABLE I
3D Detection Comparisons. The mAP Scores Are Measured With the IoU Threshold at 0.5

Models	Input	mAP
3D-SIS [87]	Geo+Image	25.70
MLCVNet [88]	Geo Only	33.40
RevealNet [24]	Geo Only	29.29
RfD-Net [11]	Geo Only	38.71
Ours (w/o joint)	Geo Only	34.78
Ours (w/ joint)	Geo Only	**39.44**

D. Comparison to the State-of-the-Art

1) Quantitative Evaluation of 3D Detection: This section compares our 3D object detection with 3D-SIS [87], MLCVNet [88], RevealNet [24], and RfD-Net [11] (as shown in Table I). 3D-SIS [87] fuses the multi-view RGB images into the TSDF grids. MLCVNet [88] extends VoteNet [37] by considering contextual information between objects. RevealNet [24] discretizes the input scans into a volumetric grid that is preprocessed as a TSDF. In contrast to RevealNet [24], we exploited the sparsity of the point clouds, as in RfD-Net [14], to ignore empty regions, reducing computation load and making reconstruction more efficient. The proposed method outperforms RfD-Net [11] because the proposed Objectness-aware Chamfer Distance loss introduces more geometric constraints from the completion module to the detection module. Note that, the 3D detection performance of the RfD-Net [11] was obtained by directly testing the pretrained module available at: https://github.com/yinyunie/RfDNet.

2) Quantitative Evaluation of the Semantic Instance Completion: To evaluate our semantic instance completion, we measured how much the predicted object meshes overlay the ground-truths (from the Scan2CAD dataset [19]) and compared it to RevealNet [24] and RfD-Net [11]. In order to assess its performance for the task of instance-level scene completion, we followed the experimental protocols of RevealNet [24] by sequentially combining the prior arts of instance segmentation [87] with shape completion [78], or combining scan completion [16] with instance segmentation [87]. Our results show that our joint method largely outperforms the decoupled method and is comparable or even better than RfD-Net [11] and RevealNet [24]. The proposed method achieves

TABLE II

SEMANTIC INSTANCE COMPLETION COMPARISON. THE RESULTS OF (INST SEG [87] + SHAPE COMP [78]) AND (SCAN COMP [16] + INST SEG [87]) ARE PROVIDED BY [24]. THE MAP SCORES ARE MEASURED WITH THE MESH IOU THRESHOLD AT 0.5

Models	display	bathtub	trashbin	sofa	chair	table	cabinet	bookshelf	mAP
Inst Seg [87]+Shape Comp [78]	2.27	1.14	1.68	14.86	9.93	3.90	7.11	3.03	5.49
Scan Comp [16]+Inst Seg [87]	1.65	4.55	11.25	9.09	9.09	0.64	0.18	5.45	5.24
RevealNet [24]	13.16	13.64	18.19	24.79	15.87	11.28	8.60	10.60	14.52
RfD-Net [11]	26.67	27.57	23.34	15.17	12.23	1.92	14.48	13.39	16.90
Our WSSIC-Net	22.33	26.54	22.48	16.27	12.44	1.98	13.52	12.75	16.56

TABLE III

ABLATION STUDY ON SEMANTIC INSTANCE COMPLETION. THE MAP SCORES ARE MEASURED WITH THE MESH IOU THRESHOLD AT 0.5. NOTE THAT, "W/O", "AUG", AND "OCD" REPRESENT "WITHOUT", "DATA AUGMENTATION", AND "OBJECTNESS-AWARE CHAMFER DISTANCE", RESPECTIVELY

Models	display	bathtub	trashbin	sofa	chair	table	cabinet	bookshelf	mAP
w/o HL	21.03	21.70	17.66	13.18	9.16	1.82	11.33	10.08	12.25
w/o Aug1	21.66	22.68	18.18	13.98	9.82	1.86	11.86	10.88	13.18
com-only	21.78	22.86	18.35	14.26	10.14	2.48	11.78	11.75	13.66
w/o Aug2	22.08	25.87	21.96	15.58	11.28	1.86	12.02	11.48	14.81
w/o OCD	22.12	26.16	22.10	15.72	11.36	1.94	12.25	12.24	15.68
w/o Obj	22.68	26.34	22.17	15.88	11.48	1.93	12.68	12.56	16.06
Our WSSIC-Net	**23.33**	**26.54**	**22.48**	**16.27**	**12.44**	**1.98**	**13.52**	**12.75**	**16.56**

slightly lower performance compared to RfD-Net [11]. The reason is that RfD-Net [11] can learn more specific latent representations than ours in a fully supervised manner.

3) Qualitative Comparisons: We compared our semantic instance completion method with the state-of-the-art fully supervised RfD-Net [11]. Based on the results in Fig 4, the proposed method achieves a higher detection accuracy than RfD-Net [14] in terms of box numbers and scales. Furthermore, the proposed method achieves comparable or even better instance completion performance than RfD-Net, as shown in Fig. 4. The reason for this is that we used several effective constraints to train our weakly supervised model, and we can use more synthetic training data generated by the GAN. As can be seen from the first, third, and fifth rows of Fig 4, the proposed method produces more detailed tables and chairs than RfD-Net [11]. This is due to the fact that our proposed method reconstructs meshes from complete point clouds rather than incomplete point sampling as in RfD-Net [11].

E. Ablation Study

ScanNet v2 [84] dataset was used in our ablation study. We separately tested data augmentation of varying orientation and size (w/o Aug1), data augmentation of mimicking the distribution of real-world 3D scans (w/o Aug2), PointNet module (w/o $\mathbf{Obj}_f$), HL term (w/o HL), and Objectness-aware Chamfer Distance (w/o OCD) from our entire unpaired structure in order to test the impact of geometrical alignment, domain alignment, semantic alignment and negative proposals on the semantic instance completion evaluation. We conducted another ablation study to examine the relationship between 3D detection and instance completion, which included detection-only (det-only) model, detection with completion model (det-com-joint) and detection with completion model with OCD loss (det-com-OCD).

1) Semantic Instance Completion Ablation: Table III shows how the performance of our method drops when certain pipeline parts are removed. **1)** The HL loss has the most impact with a drop in performance of about 26%. This is because without the HL term in the generator loss, our method can only produce general 3D structures of each category, leading to mis-alignment of geometry with the Scan2CAD dataset, see Fig. 3. This is because the GAN loss can only generate a general complete proposal feature for each category under the supervision of semantically-consistent synthetic objects. It is possible to adjust the generated "complete" proposal feature to some degree with unidirectional Hausdorff distance loss (HL) to produce complete objects as similar to their corresponding partial objects as possible. **2)** The data augmentation of varying orientation and size of the synthetic objects (*w/o Aug1*) also has a significant impact with a drop in performance of about 20%. This is because the data augmentation has a significant impact on the HL loss. More specifically, HL loss requires the network to predict the correct orientation and size of objects. **3)** Comparing the *com-only* module and the *w/o OCD* module demonstrates the benefits of a joint training process for the detection and completion tasks. **4)** The comparison between the *w/o OCD* module and our WSSIC-Net demonstrates the effectiveness of the proposed OCD loss. This is because this loss can add geometric constraints to the 3D detection to predict more accurate orientations and scales, resulting in fewer negative proposals. Having a large number of accurate box proposals improves point cloud matching performance, resulting in better semantic instance completion performance. **5)** Using data augmentation of distribution mimicking of 3D detection for auto-encoder pretraining (*w/o Aug2*) is beneficial to the proposed method. This strategy can narrow the domain gap between synthetic and real-world datasets. **6)** Although the combination of the latent representations $\mathbf{Obj}_f \in \mathbb{R}^{M\times 1024}$ of point clouds of objects and the filtered proposal features has the least impact on the proposed method, it still provides more geometric information to the filtered proposal features, leading to better performance.

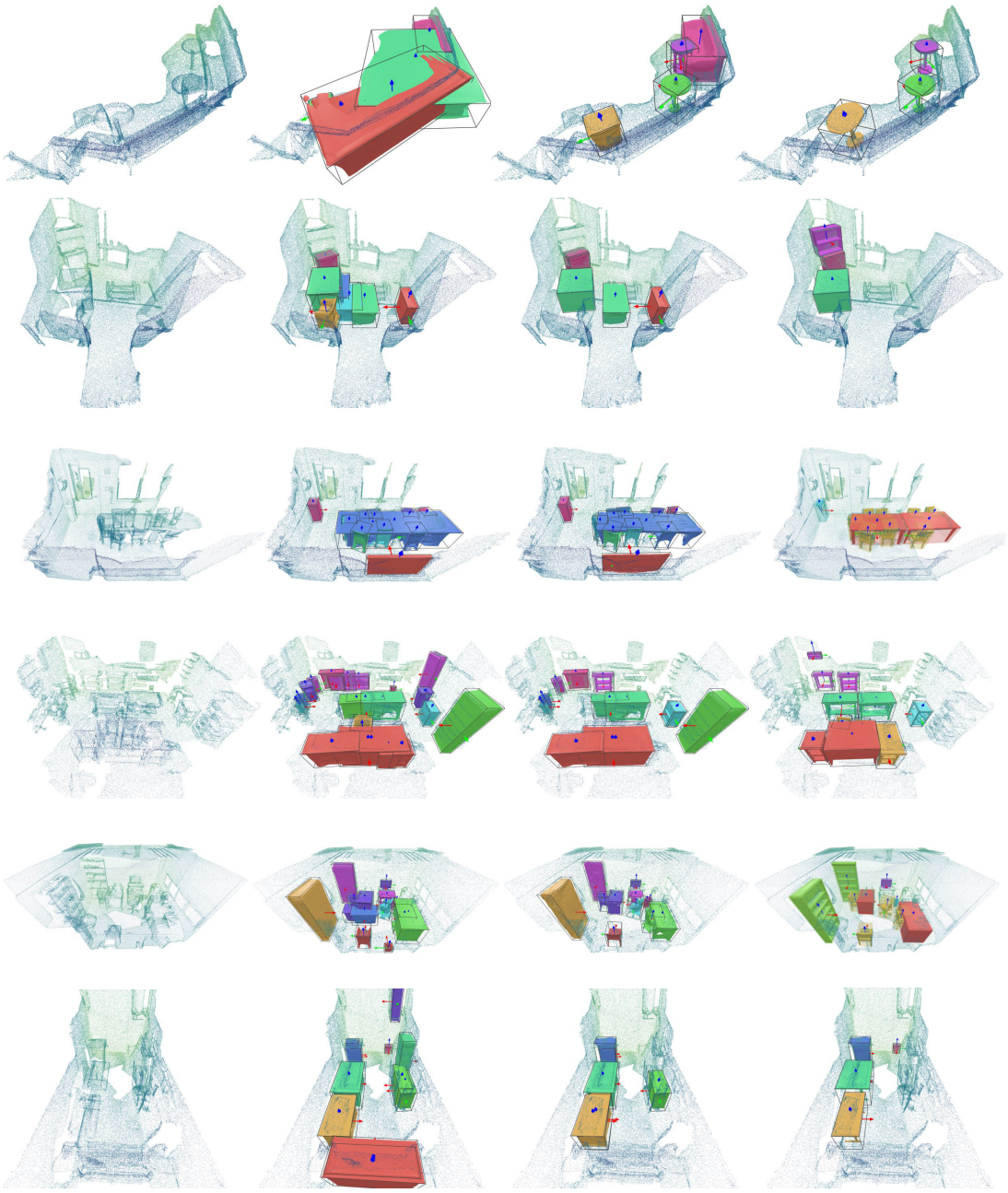


Fig. 4. Qualitative comparison to RfD-Net in 3D detection and instance completion. Different colors represent different categories.

TABLE IV

ABLATION STUDY ON 3D DETECTION. THE MAP SCORES ARE MEASURED WITH THE MESH IOU THRESHOLD AT 0.5

Models	display	bathtub	trashbin	sofa	chair	table	cabinet	bookshelf	mAP
det-only	10.83	27.87	16.48	49.72	75.99	48.51	32.95	15.86	34.78
det-com-joint	18.04	26.98	22.51	54.49	76.28	50.89	33.29	26.37	38.58
det-com-OCD	18.46	28.28	32.93	56.22	74.78	42.24	33.66	27.43	**39.32**

2) Detection Ablation: We also examined the effect of combining 3D detection and scene completion in Table IV. More specifically, the *det-only* model represents the 3D detection model trained on the ScanNet v2 dataset [84], whereas the *det-com-joint* and *det-com-OCD* models represent the entire pipeline of 3D detection and instance completion jointly trained on the ScanNet v2 dataset without/with the proposed joint loss. The *det-only* is supervised by ground-truth 3D

bounding boxes on the ScanNet v2 dataset, while the *det-com-joint* and *det-com-OCD* are supervised by ground-truth 3D bounding boxes on the ScanNet v2 dataset and unpaired complete objects on the ShapeNetCore dataset. Compared to the *det-only* model, we can conclude that high-performance scene completion can improve 3D detection performance. Additionally, the proposed OCD loss can further enhance the performance of 3D detection and instance completion. This ablation study demonstrates that the 3D detection and instance completion tasks can benefit from each other even with unpaired complete objects of point clouds. A comparison of the performance of RfD-Net [11] in Table I with that of *det-com-joint* module in Table IV illustrates how ground truth of complete objects can better benefit 3D detection than synthetic complete point clouds of objects. This is because, the ground truth of complete objects can provide more accurate 3D geometric information than synthetic complete objects.

V. Conclusion

A Weakly-Supervised Semantic Instance Completion Network (WSSIC-Net) for point clouds is presented in this paper. WSSIC-Net first maps each partial point cloud object into a proposal feature using a dedicated 3D detection model. Positive proposal features are then filtered from the proposal features before being concatenated with the latent features of partial objects segmented from the positive proposals. Next, the concatenated positive proposal features are fed to a GAN to learn from "complete" latent features represented by a pre-trained encoder using "pseudo" ground truth. A pre-trained decoder then extracts complete point clouds of the objects from the generated "complete" proposal features. Moreover, two data augmentations, "pseudo" ground truth, HL loss and the OCD loss are introduced to handle the challenges of geometrical alignment, semantic alignment, domain alignment and negative proposals. Experiments on the ScanNet v2 dataset demonstrate that the proposed weakly-supervised method achieves comparable or better results than the state-of-the-art **supervised** methods. Moreover, reported experimental results indicate that the 3D detection task can benefit from the instance completion task even when supervised by unpaired complete synthetic point clouds of objects. One limitation of our proposed approach is the requirement for the "pseudo" ground truth dataset to include the categories of every object that exists in real-world scenes. In future research, we plan to investigate the effects of incorporating additional modalities like normals and colors into this approach.

References

[1] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3D point clouds: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4338–4364, Dec. 2020.

[2] Q. Hu et al., "Learning semantic segmentation of large-scale point clouds with random sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8338–8354, Jul. 2021.

[3] S. Cheng, X. Chen, X. He, Z. Liu, and X. Bai, "PRA-Net: Point relation-aware network for 3D point cloud analysis," *IEEE Trans. Image Process.*, vol. 30, pp. 4436–4448, 2021.

[4] Y. Zheng, X. Xu, J. Zhou, and J. Lu, "PointRas: Uncertainty-aware multi-resolution learning for point cloud segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 6002–6016, 2022.

[5] J. Yang et al., "Learning to reconstruct and understand indoor scenes from sparse views," *IEEE Trans. Image Process.*, vol. 29, pp. 5753–5766, 2020.

[6] X.-F. Han, H. Laga, and M. Bennamoun, "Image-based 3D object reconstruction: State-of-the-Art and trends in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1578–1604, May 2021.

[7] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1738–1764, Apr. 2022.

[8] Q. Hu, B. Yang, S. Khalid, W. Xiao, N. Trigoni, and A. Markham, "SensatUrban: Learning semantics from urban-scale photogrammetric point clouds," *Int. J. Comput. Vis.*, vol. 130, no. 2, pp. 316–343, Feb. 2022.

[9] X. Liu, X. Liu, Y.-S. Liu, and Z. Han, "SPU-net: Self-supervised point cloud upsampling by coarse-to-fine reconstruction with self-projection optimization," *IEEE Trans. Image Process.*, vol. 31, pp. 4213–4226, 2022.

[10] A. Tao, Y. Duan, Y. Wei, J. Lu, and J. Zhou, "SegGroup: Seg-level supervision for 3D instance and semantic segmentation," *IEEE Trans. Image Process.*, vol. 31, pp. 4952–4965, 2022.

[11] Y. Nie, J. Hou, X. Han, and M. Nießner, "RfD-net: Point scene understanding by semantic instance reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4606–4616.

[12] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, Sep. 1993.

[13] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 190–198.

[14] S. Liu et al., "See and think: Disentangling semantic scene completion," in *Proc. NeurIPS*, vol. 31, Jan. 2018, pp. 261–272.

[15] J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan, "Anisotropic convolutional networks for 3D semantic scene completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3348–3356.

[16] A. Dai, D. Ritchie, M. Bokeloh, S. Reed, J. Sturm, and M. Nießner, "ScanComplete: Large-scale scene completion and semantic segmentation for 3D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4578–4587.

[17] X. Han et al., "Deep reinforcement learning of volume-guided progressive view inpainting for 3D point scene completion from a single depth image," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 234–243.

[18] W. Hu, Z. Fu, and Z. Guo, "Local frequency interpretation and non-local self-similarity on graph for point cloud inpainting," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4087–4100, Aug. 2019.

[19] A. Avetisyan, M. Dahnert, A. Dai, M. Savva, A. X. Chang, and M. Nießner, "Scan2CAD: Learning CAD model alignment in RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2614–2623.

[20] A. Avetisyan, A. Dai, and M. Niessner, "End-to-end CAD model retrieval and 9DoF alignment in 3D scans," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2551–2560.

[21] A. Avetisyan, T. Khanova, C. Choy, D. Dash, A. Dai, and M. Nießner, "SceneCAD: Predicting object alignments and layouts in RGB-D scans," in *Proc. IEEE Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 596–612.

[22] V. Ishimtsev et al., "CAD-deform: Deformable fitting of CAD models to 3D scans," in *Proc. ECCV*. Cham, Switzerland: Springer, Jan. 2020, pp. 599–628.

[23] M. Rouhani, A. D. Sappa, and E. Boyer, "Implicit B-spline surface reconstruction," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 22–32, Jan. 2015.

[24] J. Hou, A. Dai, and M. Nießner, "RevealNet: Seeing behind objects in RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2095–2104.

[25] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional occupancy networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Aug. 2020, pp. 523–540.

[26] J. Li, B. M. Chen, and G. H. Lee, "SO-Net: Self-organizing network for point cloud analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9397–9406.

[27] C. Wang, B. Samari, and K. Siddiqi, "Local spectral graph convolution for point set feature learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 52–66.

[28] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1907–1915.

[29] X. Ye, J. Li, H. Huang, L. Du, and X. Zhang, "3D recurrent neural networks with context fusion for point cloud semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 403–417.

[30] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[31] D. Xu, D. Anguelov, and A. Jain, "PointFusion: Deep sensor fusion for 3D bounding box estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 244–253.

[32] X. Ding, W. Lin, Z. Chen, and X. Zhang, "Point cloud saliency detection by local and global feature fusion," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5379–5393, Nov. 2019.

[33] Y. Xu, T. Fan, M. Xu, L. Zeng, and Y. Qiao, "SpiderCNN: Deep learning on point sets with parameterized convolutional filters," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 87–102.

[34] B. Wu, A. Wan, X. Yue, and K. Keutzer, "SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 1887–1893.

[35] D. Rethage, J. Wald, J. Sturm, N. Navab, and F. Tombari, "Fully-convolutional point networks for large-scale point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 596–611.

[36] H. Su et al., "SPLATNet: Sparse lattice networks for point cloud processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2530–2539.

[37] C. R. Qi, O. Litany, K. He, and L. Guibas, "Deep Hough voting for 3D object detection in point clouds," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9277–9286.

[38] S. Shi et al., "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10529–10538.

[39] Y. Zhang, Q. Hu, G. Xu, Y. Ma, J. Wan, and Y. Guo, "Not all points are equal: Learning highly efficient point-based detectors for 3D LiDAR point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 18953–18962.

[40] B. Yang et al., "Learning object bounding boxes for 3D instance segmentation on point clouds," in *Proc. NeurIPS*, vol. 32, Jan. 2019, pp. 1–20.

[41] M. Feng, S. Z. Gilani, Y. Wang, L. Zhang, and A. Mian, "Relation graph network for 3D object detection in point clouds," *IEEE Trans. Image Process.*, vol. 30, pp. 92–107, 2021.

[42] I. Misra, R. Girdhar, and A. Joulin, "An end-to-end transformer model for 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2906–2917.

[43] Z. Liu, Z. Zhang, Y. Cao, H. Hu, and X. Tong, "Group-free 3D object detection via transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2949–2958.

[44] J. Mao et al., "Voxel transformer for 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3164–3173.

[45] Q. Cai, Y. Pan, T. Yao, and T. Mei, "3D cascade RCNN: High quality object detection in point clouds," 2022, *arXiv:2211.08248*.

[46] Y. Wang et al., "Bridged transformer for vision and point cloud 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12114–12123.

[47] Q. Meng, W. Wang, T. Zhou, J. Shen, Y. Jia, and L. Van Gool, "Towards a weakly supervised framework for 3D point cloud object detection and annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4454–4468, Mar. 2021.

[48] J. Yin et al., "ProposalContrast: Unsupervised pre-training for LiDAR-based 3D object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 17–33.

[49] J. Yin et al., "Semi-supervised 3D object detection with proficient teachers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 727–743.

[50] T. Feng, W. Wang, X. Wang, Y. Yang, and Q. Zheng, "Clustering based point cloud representation learning for 3D analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8283–8294.

[51] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert, "PCN: Point completion network," in *Proc. Int. Conf. 3D Vis. (3DV)*, Sep. 2018, pp. 728–737.

[52] L. P. Tchapmi, V. Kosaraju, H. Rezatofighi, I. Reid, and S. Savarese, "TopNet: Structural point cloud decoder," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 383–392.

[53] M. Liu, L. Sheng, S. Yang, J. Shao, and S. Hu, "Morphing and sampling network for dense point cloud completion," in *Proc. AAAI*, Apr. 2020, vol. 34, no. 7, pp. 11596–11603.

[54] Y. Nie et al., "Skeleton-bridged point completion: From global inference to local adjustment," in *Proc. NeurIPS*, vol. 33, Jan. 2020, pp. 16119–16130.

[55] Z. Huang, Y. Yu, J. Xu, F. Ni, and X. Le, "PF-Net: Point fractal network for 3D point cloud completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7662–7670.

[56] X. Wang, M. H. Ang, and G. H. Lee, "Cascaded refinement network for point cloud completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 787–796.

[57] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry, "A papier-mâché approach to learning 3D surface generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 216–224.

[58] Y. Xia, Y. Xia, W. Li, R. Song, K. Cao, and U. Stilla, "ASFM-Net: Asymmetrical Siamese feature matching network for point completion," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1938–1947.

[59] L. Pan et al., "Variational relational point completion network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8524–8533.

[60] H. Xie, H. Yao, S. Zhou, J. Mao, S. Zhang, and W. Sun, "GRNet: Gridding residual network for dense point cloud completion," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 365–381.

[61] X. Wen et al., "PMP-net: Point cloud completion by learning multi-step point moving paths," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7439–7448.

[62] X. Yu, Y. Rao, Z. Wang, Z. Liu, J. Lu, and J. Zhou, "PoinTr: Diverse point cloud completion with geometry-aware transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12498–12507.

[63] Y. Wang, D. J. Tan, N. Navab, and F. Tombari, "Learning local displacements for point cloud completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1558–1567.

[64] B. Yang, H. Wen, S. Wang, R. Clark, A. Markham, and N. Trigoni, "3D object reconstruction from a single depth view with adversarial learning," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 679–688.

[65] W. Wang, Q. Huang, S. You, C. Yang, and U. Neumann, "Shape inpainting using 3D generative adversarial network and recurrent convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2317–2325.

[66] S. Gurumurthy and S. Agrawal, "High fidelity semantic shape completion for point clouds using latent optimization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1099–1108.

[67] X. Wen, Z. Han, Y.-P. Cao, P. Wan, W. Zheng, and Y.-S. Liu, "Cycle4Completion: Unpaired point cloud completion using cycle transformation with missing region coding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13075–13084.

[68] J. Zhang et al., "Unsupervised 3D shape completion through GAN inversion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1768–1777.

[69] Y. Li and G. Baciu, "HSGAN: Hierarchical graph learning for point cloud generation," *IEEE Trans. Image Process.*, vol. 30, pp. 4540–4554, 2021.

[70] R. Li, X. Li, C.-W. Fu, D. Cohen-Or, and P.-A. Heng, "PU-GAN: A point cloud upsampling adversarial network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7203–7212.

[71] C.-L. Li, M. Zaheer, Y. Zhang, B. Poczos, and R. Salakhutdinov, "Point cloud GAN," 2018, *arXiv:1810.05795*.

[72] M. Cheng, G. Li, Y. Chen, J. Chen, C. Wang, and J. Li, "Dense point cloud completion based on generative adversarial network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5701310.

[73] J. Zhang, H. Zhao, A. Yao, Y. Chen, L. Zhang, and H. Liao, "Efficient semantic scene completion network with spatial group convolution," in *Proc. IEEE Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 733–749.

[74] Y.-X. Guo and X. Tong, "View-volume network for semantic scene completion from a single depth image," 2018, *arXiv:1806.05361*.

[75] Y. Wang, D. J. Tan, N. Navab, and F. Tombari, "ForkNet: Multi-branch volumetric semantic completion from a single depth image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8607–8616.

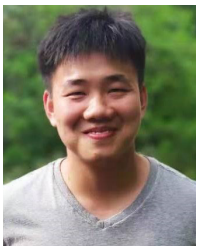
[76] P. Zhang, W. Liu, Y. Lei, H. Lu, and X. Yang, "Cascaded context pyramid for full-resolution 3D semantic scene completion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7800–7809.

[77] X. Chen, K.-Y. Lin, C. Qian, G. Zeng, and H. Li, "3D sketch-aware semantic scene completion via semi-supervised structure prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4192–4201.
[78] A. Dai, C. R. Qi, and M. Nießner, "Shape completion using 3D-encoder-predictor CNNs and shape synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5868–5877.
[79] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.
[80] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3D object reconstruction from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 605–613.
[81] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. NeurIPS*, vol. 28, Dec. 2015, pp. 91–99.
[82] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2794–2802.
[83] X. Chen, B. Chen, and N. J. Mitra, "Unpaired point cloud completion on real scans using adversarial training," 2019, *arXiv:1904.00069*.
[84] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5828–5839.
[85] A. X. Chang et al., "ShapeNet: An information-rich 3D model repository," 2015, *arXiv:1512.03012*.
[86] C. R. Qi, Y. Li, H. Su, and L. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. NeurIPS*, vol. 30, Jan. 2017, pp. 1–16.
[87] J. Hou, A. Dai, and M. Nießner, "3D-SIS: 3D semantic instance segmentation of RGB-D scans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4421–4430.
[88] Q. Xie et al., "MLCVNet: Multi-level context VoteNet for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10447–10456.

Zhiheng Fu received the B.E. degree in electric engineering from Northeastern University (NEU), Shenyang, China, in 2015, and the M.E. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Software Engineering (CSSE), UWA. He has published several papers in major journals and conferences, including IEEE TRANSACTIONS ON IMAGE PROCESSING, ICCV, ECCV, *Pattern Recognition*, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His current research interests include 3D vision and deep learning.

Yulan Guo (Senior Member, IEEE) is currently a Professor with Sun Yat-sen University. He has authored over 200 articles at highly referred journals and conferences. His research interests include 3D vision, low-level vision, and robotics. He is a Senior Member of ACM. He served as an Area Chair for CVPR 2025/2023/2021, ICCV 2025/2021, ECCV 2024, NeurIPS 2024, and ACM Multimedia 2021. He organized over ten workshops, challenges and tutorials in prestigious conferences, such as CVPR, ICCV, ECCV, and 3DV. He served as a Senior Area Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and an Associate Editor for *The Visual Computer* and *Computers and Graphics*.

Minglin Chen (Graduate Student Member, IEEE) received the B.E. degree in communication engineering from South China Normal University (SCNU), Guangzhou, China, in 2017, and the M.E. degree in software engineering from the University of Chinese Academy of Science (UCAS), Beijing, China, in 2020. He is currently pursuing the Ph.D. degree with Sun Yat-sen University (SYSU). His research interests include 3D vision, reconstruction, and generation.



Qingyong Hu received the M.Eng. degree in information and communication engineering from the National University of Defense Technology (NUDT) in 2018 and the D.Phil. degree in computer science from the University of Oxford. He is currently an Associate Professor and an Independent Researcher. He has published several papers in major journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *International Journal of Computer Vision*, CVPR, and NeurIPS. His research interests include 3D computer vision, particularly the semantic understanding of large-scale 3D point clouds, instance segmentation, and registration. More recently, he has expanded his focus to include multimodal large language models. He was a recipient of Huawei UK AI Fellowship and the World Artificial Intelligence Conference Youth Outstanding Paper Award.

Hamid Laga is currently a Professor with Murdoch University, Australia. His research interests include machine learning, computer vision, computer graphics, and pattern recognition, with a special focus on the 3D reconstruction, modeling and analysis of static and deformable 3D objects, and on machine learning for agriculture and health. He was a recipient of the best paper awards at SGP2017, DICTA2012, and SMI2006.

Farid Boussaid received the M.S. and Ph.D. degrees in microelectronics from the National Institute of Applied Science, Toulouse, France, in 1996 and 1999, respectively. He joined Edith Cowan University, Perth, Australia, as a Post-Doctoral Research Fellow; and the Visual Information Processing Research Group, as a member, in 2000. He joined The University of Western Australia, Crawley, Australia, in 2005, where he is currently a Professor. His current research interests include smart CMOS sensors, computer vision, and machine learning.

Mohammed Bennamoun (Senior Member, IEEE) is currently a Winthrop Professor with the Department of Computer Science and Software Engineering, UWA, and a Researcher of computer vision, machine/deep learning, robotics, and signal/speech processing. He has published four books (available on Amazon), one edited book, one Encyclopedia article, 14 book chapters, more than 160 journal articles, more than 250 conference publications, and 16 invited and keynote publications. His H-index is 77 and his number of citations is more than 32 000 (Google Scholar). He was awarded more than 90 competitive research grants, from Australian Research Council, and numerous other Government, UWA, and industry Research Grants. He successfully supervised more than 40 Ph.D. students to completion. He won the Best Supervisor of the Year Award at QUT (1998), and received award for research supervision at UWA (2008 and 2016) and the Vice-Chancellor Award for Mentorship (2016). He delivered conference tutorials at major conferences, including: IEEE Computer Vision and Pattern Recognition (CVPR 2016), Interspeech 2014, IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), and European Conference on Computer Vision (ECCV). He was also invited to give a Tutorial at an International Summer School on Deep Learning (DeepLearn 2017).