

Shape Anchor Guided Holistic Indoor Scene Understanding

Mingyue Dong¹, Linxi Huan¹, Hanjiang Xiong¹, Shuhan Shen², Xianwei Zheng^{1*}

¹The State Key Lab. LIESMARS, Wuhan University

²Institute of Automation, Chinese Academy of Sciences

{dmy25148, whu_hlx, xionghanjiang, zhengxw}@whu.edu.cn

shshen@nlpr.ia.ac.cn

Abstract

This paper proposes a shape anchor guided learning strategy (AncLearn) for robust holistic indoor scene understanding. We observe that the search space constructed by current methods for proposal feature grouping and instance point sampling often introduces massive noise to instance detection and mesh reconstruction. Accordingly, we develop AncLearn to generate anchors that dynamically fit instance surfaces to (i) unmix noise and target-related features for offering reliable proposals at the detection stage, and (ii) reduce outliers in object point sampling for directly providing well-structured geometry priors without segmentation during reconstruction. We embed AncLearn into a reconstruction-from-detection learning system (AncRec) to generate high-quality semantic scene models in a purely instance-oriented manner. Experiments conducted on the challenging ScanNetv2 dataset demonstrate that our shape anchor-based method consistently achieves state-of-the-art performance in terms of 3D object detection, layout estimation, and shape reconstruction. The code will be available at <https://github.com/Geo-Tell/AncRec>.

1. Introduction

Holistic indoor scene understanding from partial observations (e.g., single-view images or 3D scans) is a comprehensive task that provides 3D semantic scene models for indoor applications. Early works studied this task with a reconstruction-from-detection framework that recovers the geometries of room structures and objects from the corresponding detection in a separate way. Later, end-to-end learning methods were proposed to simultaneously perform layout estimation, object detection, and shape prediction in one forward pass for semantic scene reconstruction [27, 37, 23]. With the recent success of point-based 3D detection and instance reconstruction, surging interest has been witnessed in detecting and modeling objects directly

*Corresponding author.

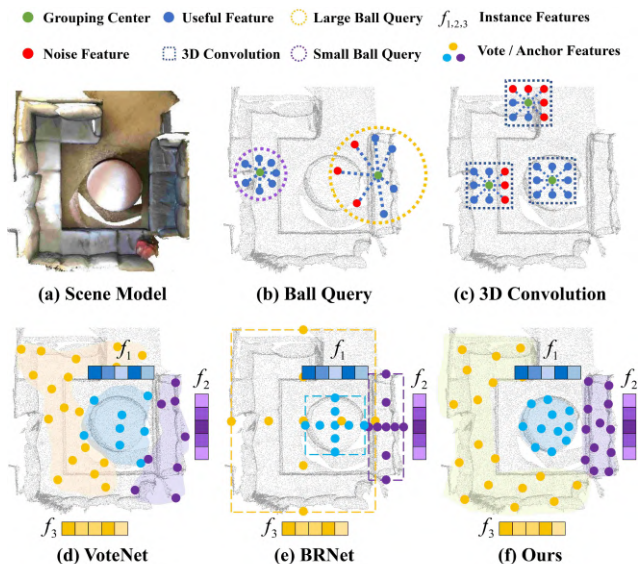


Figure 1. Comparison between different feature grouping strategies. (a) the original scene model. (b)-(e) The different feature grouping operations all suffer from the issue of confusing non-target noise with useful features. (f) The proposed shape anchor guided grouper directly generates anchors at the instance surface to merge instance-related features, which largely alleviates noise interference.

from sparse point clouds [28, 33].

Benefiting from the use of rich geometry information, current scan-based deep methods have improved the performance of semantic scene reconstruction. However, they still left two issues that bottleneck high-quality semantic reconstruction: (1) the noisy instance feature learning at the detection phase, and (2) the difficulty in retrieving instances from sparse point clouds for reconstruction.

At the detection phase, it is required to group features for instance representation learning. The ball query [30, 31] and 3D convolution [39, 20] are two basic operations for point feature grouping, but they often mix massive noise with informative features due to the fixed grouping range,

as shown in Fig. 1 (b) and (c). To compensate for the deficiency caused by the fixed range, VoteNet [14] and its variants [34, 29, 38, 35] adopt a voting strategy to cluster object features by moving surface points towards object centers. Albeit more flexible than the basic operations, the voting-based strategy often generates an unconstrained grouping area that brings quantities of outliers as illustrated in Fig. 1 (d). BRNet [10] hence restricts the grouping space by sampling around representative points given by coarse box proposals. Nevertheless, limited by box-like grouping area, the sampling points can still fall far beyond targets when the objects are irregularly shaped, as depicted in Fig. 1 (e).

At the reconstruction stage, the retrieval of outlier-free object points is a prerequisite for object recovery. Due to the noise introduced by feature grouping in the previous detection phase, the points grouped for localizing objects can hardly serve as ideal reconstruction priors as indicated by Fig. 1. Consequently, the current methods are forced to employ an additional foreground classifier [28] or replace the detector with a complex instance segmentation backbone to sample object points from the raw scans [33]. However, the existence of numerous non-target outliers in the search space challenges instance segmentation, resulting in increased risks of gluing different instances and misclassifying background points.

Based on the discussion above, the two noise interference issues during detection and reconstruction are actually highly coupled and can be resolved together as long as outliers are excluded during feature grouping. To this end, we are motivated to propose a shape anchor-guided learning strategy (*AncLearn*) that generates surface anchors to fit the feature grouping areas to object shape distributions, as displayed in Fig. 1 (f). With the geometry constraint provided by surface anchors, it is able to merge local target-focused features for predicting reliable object proposals and construct a shape-aware search space for robustly sampling instance points without segmentation during reconstruction. The proposed anchor-guided learning strategy can be easily embedded into an end-to-end learning system to accomplish object detection, layout estimation, and instance reconstruction for holistic scene understanding. The main contributions are summarized as follows:

- We present a shape anchor guided learning strategy to simultaneously address the issues of noisy feature learning in detection and instance point retrieval during reconstruction.
- We embed the proposed anchor-guided learning strategy into an end-to-end learning system to accomplish object detection, layout estimation, and instance reconstruction for holistic scene understanding in a purely instance-oriented way.
- Extensive experiments demonstrate that our AncRec achieves high-quality semantic scene reconstruction

with state-of-the-art performance in instance detection and mesh prediction on the challenging ScanNetv2 dataset [12] (some ground truths provided by Scan2CAD [1] and SceneCAD [3]).

2. Related Work

Semantic scene understanding has been extensively studied over the past years. Many methods focused on acquiring the semantics of 3D scenes [22, 15, 8], while others recovered scene geometries by scene completion [18, 13, 36]. Recently, increasing interest has emerged in semantic scene reconstruction, which recovers both the semantics and geometric shapes of objects. By treating semantic reconstruction as a problem of holistic scene understanding, promising progress has been made based on a reconstruction-from-detection principle [28]. In the following, we review the research from the two core aspects of the reconstruction-from-detection pipeline, *i.e.*, 3D object detection and scene-aligned instance reconstruction.

2.1. 3D Object Detection

Over the past years, deep detectors have gained great success in 2D object detection [40, 25]. The progress of 2D deep detection has inspired the development of deep learning techniques for recognizing objects from scene point clouds. Compared to image data, point clouds provide rich surface geometry clues for locating objects in real scenes. Nevertheless, the sparse, irregular, and orderless characteristics of point clouds make them hard to be handled by the grid-based convolution model.

Works in the early stage leveraged the 2D proposals as 3D detection constraints or projected point clouds into regular 2D/3D grids. Although these 3D detectors are applicable, their performance is still limited by the 2D detectors and the missing geometric details brought by projection. To directly learn the rich geometry features from the raw points, PointNet [30] and PointNet++ [31] used a ball query operation for grouping point features. Later, PointRCNN [32] adopted the point-wise features extracted by PointNet++ for producing instance proposals with point clouds. Hindered by the fixed feature grouping range of grid-based convolutions or ball query operation, these methods often mixed massive noise with informative features, which impairs the reliability of proposals. Considering that the observed object surface points usually lie far away from the object centers, VoteNet [14] introduced a voting mechanism for proposal generation. Based on the voting mechanism, numerous variants further refined proposal features with context relationships [34], hierarchical clues [35], and hybrid geometric primitives [38]. However, the issue of outliers in votes still blocks the learning of representative box features. BRNet [10] derived proposal features based on the generated virtual points given by coarse box proposals,

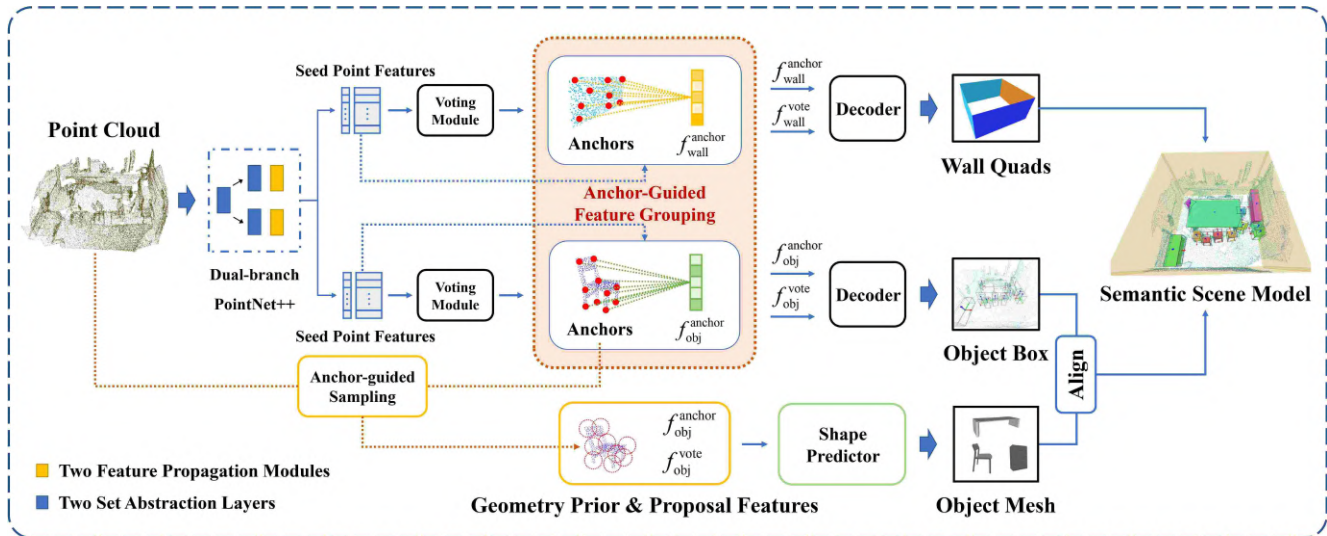


Figure 2. Overview of the AncRec framework. At the detection stage, seed point features for walls and objects are first learned with a modified dual-branch PointNet++. Following each branch, a voting module and an anchor-guided feature grouper respectively generate proposal features $f_{obj(wall)}^{vote}$ and $f_{obj(wall)}^{anchor}$, which are fed into the decoders to predict object bounding boxes and wall quads. The room layout is then constructed by processing the wall quads into connected corners. At the reconstruction stage, objects with high objectness scores are reconstructed under the guidance of f_{obj}^{vote} , f_{obj}^{anchor} , and the geometry priors sampled by shape anchors. Finally, object models are arranged in the scene with the predicted layout according to the spatial alignment provided by predicted bounding boxes.

whereas the virtual points can fall into non-target areas and bring noise. In this paper, we directly merge features anchored on target surfaces into robust proposal representations for indoor instance learning.

2.2. Scene-aligned Instance Reconstruction

Scene-aligned instance reconstruction requires not only modeling 3D object shapes but also correctly arranging the shapes in the 3D scene space. The semantic model of an indoor scene was early constructed with retrieval techniques, which search for CAD shapes or geometric primitives from an offline database and align the approximate object models to input scene data [1, 2, 17, 24]. Albeit able to present delicate scene models, the retrieval-based methods lack generalization ability to various scenes due to the inefficient inference and the limited database scale [37].

The promising advances in deep learnable shape representations motivated scene-aligned instance reconstruction without a finite model pooling [16, 11]. Many prior arts in literature generate 3D objects with explicit or implicit representations with learned features derived from 2D recognition results [19, 27, 37, 23]. Unlike previous single-view modeling methods, RevealNet [21] and the following works [5, 4] predicted the occupancy grids of semantic instances with 3D features extracted from voxelized RGB-D scans. For high scene resolution that is hard to acquire by volumetric representation, RfD-Net [28] and DIMR [33] extracted instance points from scans with segmentation techniques for subsequent single object reconstruction with deep implicit

functions [26, 9]. Because of the existence of non-target points, these two state-of-the-art methods have to use an extra foreground classifier or a complex instance segmentation backbone to sample object points for reconstruction. In this paper, we utilized the previously generated surface anchors to localize instance points in scans. Hereby, we introduced an easy but effective anchor-guided sampling strategy to offer geometry priors for the following instance reconstruction without an extra segmentation stage.

3. Method

We illustrate the proposed AncRec in Fig. 2. AncRec achieves semantic scene reconstruction with the proposed anchor-guided learning strategy in an instance-oriented way. It first localizes objects and walls via a dual-branch detector that is equipped with shape anchors learned for grouping point features at the detection stage. The object shape anchors are subsequently leveraged to sample instance geometry priors for predicting object meshes. Finally, the complete semantic scene model is reconstructed by arranging object meshes in the post-processed room layout with alignment to the parsed bounding boxes and poses. In the following, we elaborate on how the shape anchors are learned and leveraged to address the noise interference during instance detection and mesh prediction for high-quality semantic scene reconstruction.

3.1. Shape Anchor-guided Instance Detection

We modify the VoteNet into a dual-branch instance detector to simultaneously localize objects and walls in one

forward pass (Fig. 2). Although the voting mechanism of VoteNet is workable for generating proposals, the reliability of voting-based predictions is often impaired by the outliers in votes. We hence design an anchor-guided feature grouper to learn target-focused features for robust instance detection.

3.1.1 Anchor-guided Feature Grouper

We illustrate the mechanism of the anchor-guided feature grouper in Fig. 3. In the following, by taking the object detection branch as an example, we describe how it works with the voting module in three steps to predict the oriented 3D bounding boxes.

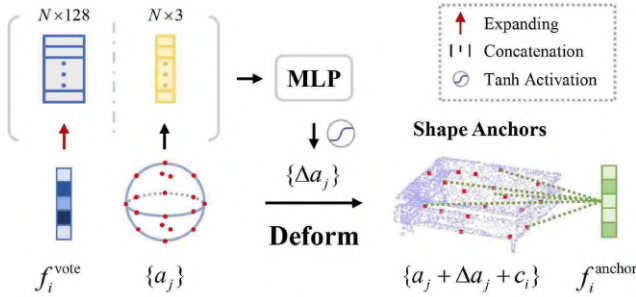


Figure 3. The anchor-guided feature grouper. The anchor-guided proposal feature f_i^{anchor} is the average of features at shape anchors. The shape anchors are deformed from initial anchors $\{a_j\}$ with offsets derived from the concatenation of the voting-based proposal feature f_i^{vote} and $\{a_j\}$. The anchors are then translated to the cluster center c_i .

Step 1: Anchor Generation. Given the i^{th} voting-based proposal feature $f_i^{\text{vote}} \in \mathbb{R}^{128}$, the anchor-guided grouper first generates shape anchors that depict the geometry of the i^{th} object candidate via template deformation. With N initial anchors $A = \{a_j \in \mathbb{R}^3 | j = 1, \dots, N\}$ uniformly selected from a unit ball surface, a deformation layer is applied to translate the initial anchors to the i^{th} candidate object surface by

$$\hat{a}_j = a_j + \underbrace{\text{Tanh}(\text{MLP}(\{a_j, f_i^{\text{vote}}\}))}_{\text{deformation offset } \Delta a_j} + c_i. \quad (1)$$

As indicated by Eq. (1), the deformation layer moves a_j with offsets Δa_j inferred by a multi-layer perceptron (MLP) and a Tanh activation function and then places the deformed anchors at the candidate center c_i clustered by the voting module. The training of the deformation layer is supervised by the Chamfer distance loss defined as

$$\begin{aligned} \mathcal{L}_{\text{anchor}} = & \frac{1}{|P_{\text{gt}}|} \sum_{p \in P_{\text{gt}}} \min_{\hat{a} \in \hat{A}} \|p - \hat{a}\|_2^2 \\ & + \frac{1}{|\hat{A}|} \sum_{\hat{a} \in \hat{A}} \min_{p \in P_{\text{gt}}} \|\hat{a} - p\|_2^2, \end{aligned} \quad (2)$$

where \hat{A} and P_{gt} denote the sets of deformed anchors and surface points sampled from the corresponding scene-aligned 3D meshes. Thereby, the anchors are learned to fit the i^{th} object surface, and the clustering center c_i is also refined with shape constraint provided by Eq. (2). Moreover, due to the supervision with complete surface points, the shape anchors can assist in exploiting context for recovering unobserved structures, *e.g.*, the missing object bottom that is supported by the observed floor areas.

Step 2: Feature Grouping. The anchor-guided grouper propagates the seed point features (extracted by the dual-branch PointNet++ backbone) to each anchor via interpolation followed by an MLP. As the deformed anchors are mainly located on the object surface, the seed point features for propagation can be reliably selected from target-related areas. The i^{th} noise-reduced proposal representation $f_i^{\text{anchor}} \in \mathbb{R}^{128}$ is obtained by averaging the anchor features.

Step 3: Prediction Fusion. We employ two decoders to predict the object category and oriented box parameters from f_i^{vote} and f_i^{anchor} respectively for the i^{th} candidate. This is based on the consideration that the voting-based proposal feature f_i^{vote} contains more contextual information while the anchor-guided proposal feature f_i^{anchor} is more target-focused. The estimated parameters, denoted as Θ_i^{vote} and Θ_i^{anchor} , are averaged with learnable weights to obtain the final object parameters:

$$\Theta_i = w_1 \cdot \Theta_i^{\text{vote}} + w_2 \cdot \Theta_i^{\text{anchor}}. \quad (3)$$

From the three steps above, the attributes of objects are parsed with robustness to noise. The anchor-guided detection of wall instances works in a similar way. Considering that walls are generally connected to each other, we additionally deploy the attention operation used in [35] to enhance the anchor-guided wall proposal features with the strong relationship between walls for precise layout estimation.

3.1.2 The Detection Training Loss

The total loss function of the anchor-guided instance detector is defined as

$$\mathcal{L} = \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{wall}} + \mathcal{L}_{\text{obj}}^{\text{anchor}} + \mathcal{L}_{\text{wall}}^{\text{anchor}}. \quad (4)$$

In Eq. (4), \mathcal{L}_{obj} is the object detection loss given by [14] while $\mathcal{L}_{\text{wall}}$ is the wall quad loss used in [7]. $\mathcal{L}_{\text{obj}}^{\text{anchor}}$ and $\mathcal{L}_{\text{wall}}^{\text{anchor}}$ are the chamfer distance losses that supervise the anchor deformation learning in terms of positive object and wall candidates.

3.2. Anchor-guided Object Reconstruction

Instance points from the input scan are desirable geometry priors for reconstruction. Previous works utilized segmentation operations for instance point sampling. Despite

being applicable, accurate instance segmentation is difficult due to the existence of massive non-target noise in the search space. In contrast, we take advantage of the shape anchors to construct a shape-aware search space, in which object points can be directly localized with little noise interference. Fig. 4 illustrates the workflow of our anchor-guided instance point sampling.

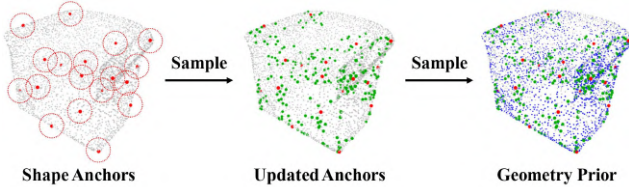


Figure 4. The anchor-guided instance point sampling. Object points are iteratively sampled from the input scan under the guidance of shape anchors. The sampled object points serve as the geometry prior for shape reconstruction.

We first add the shape anchors of positive proposals to the original scene scan to enhance indoor objects’ structural information. From the enhanced scan, we select points that lie within a given radius of each anchor and update the anchor set with these selected points to expand the searching space. With the searching space compactly fitting the object shape, we can further sample more instance points with high coverage. In this way, our anchor-guided sampling efficiently generates object geometry prior without reliance on an extra segmentation module. In the experiments, the anchor-guided sampling process iterates twice with the sampling radius set to the minimum distance between the anchors.

The geometry priors are next concatenated with the corresponding proposal features f_{obj}^{vote} and f_{obj}^{anchor} and encoded into shape embeddings f_{obj}^{shape} by a ResPointNet [28, 30]. Based on f_{obj}^{shape} , the decoder of BSP-Net [9] is adopted to predict the signed distances of query points at the canonical coordinates. Following the BSP-Net, we apply the Constructive Solid Geometry (CSG) technique to extract the shape surfaces.

3.3. Semantic Scene Reconstruction

We now build the semantic scene model with the anchor-guided results of the instance detector and shape predictor. The room structure is first constructed by transforming the detected wall quads into orderly connected corners with the merging technique [7]. Next, the models of objects with high objectness are chosen to be arranged in the scene with alignment to the predicted 3D bounding boxes [28]. In the end, the semantic scene model is built as a combination of the reconstructed layout and the scene-aligned object shapes.

4. Experiments

4.1. Experiment Settings

4.1.1 Dataset

We test our method on ScanNetv2 [12] with ground truths from Scan2CAD [1] and SceneCAD datasets [3] for holistic indoor scene understanding. (1) ScanNetv2 is a benchmark for indoor scene analysis with 1,513 scanned room point clouds; (2) Scan2CAD is an alignment dataset that matches ShapeNet models to their counterpart object instances in ScanNet with oriented 3D bounding boxes; and (3) SceneCAD provides 3D layout annotations for the scans in ScanNetv2. We follow [7] and [9] to pre-process the layout polygons and object meshes for network training supervision. The train/test split is kept in line with the previous works with eight challenging object categories considered in the experiments.

4.1.2 Evaluation Metrics

The quality of semantic scene reconstruction is evaluated with the performance of object detection, layout estimation, and scene-aligned shape reconstruction. In line with previous works [28, 7], we use mean average precision across all classes (mAP) with 0.5 IoU threshold for object detection and F1-score for layout estimation. To evaluate the reconstruction quality, we use the chamfer distance (CD) based mAP with thresholds 0.1 and 0.047, light field distance (LFD) based mAP with thresholds 5000 and 2500, as well as the mAP at 3D IoU thresholds 0.25 and 0.5.

4.1.3 Implementation Details

The training of our method is conducted on a Titan GPU with two stages, and all parameters are updated by the Adam optimizer. In the first stage, we train the dual-branch instance detector of AncRec for simultaneous optimization of object detection and layout estimation modules. We set the batch size to 8, initialize the learning rate to $1e-3$, and adopt the *ReducerLrOnPlateau* learning scheduler in PyTorch. In the second stage, we use the object proposals predicted by the instance detector for mesh reconstruction. We train the shape predictor with the BSPNet decoder pre-trained on the 8-category ShapeNet data [6] for training stability and efficiency. The training process lasts 100 epochs with batch size set to 32, learning rate initialized as $1e-4$, and the same training schedule. We found the number of shape anchors N is an insensitive hyper-parameter and thus set it to 18 for computational efficiency.

4.2. Comparison and Analysis

Scene Reconstruction. We compare our method with the state-of-the-art works, *i.e.*, RfD-Net [28] and DIMR [33]. RfD-Net predicts object shapes with the instance points segmented from proposals given by VoteNet, while

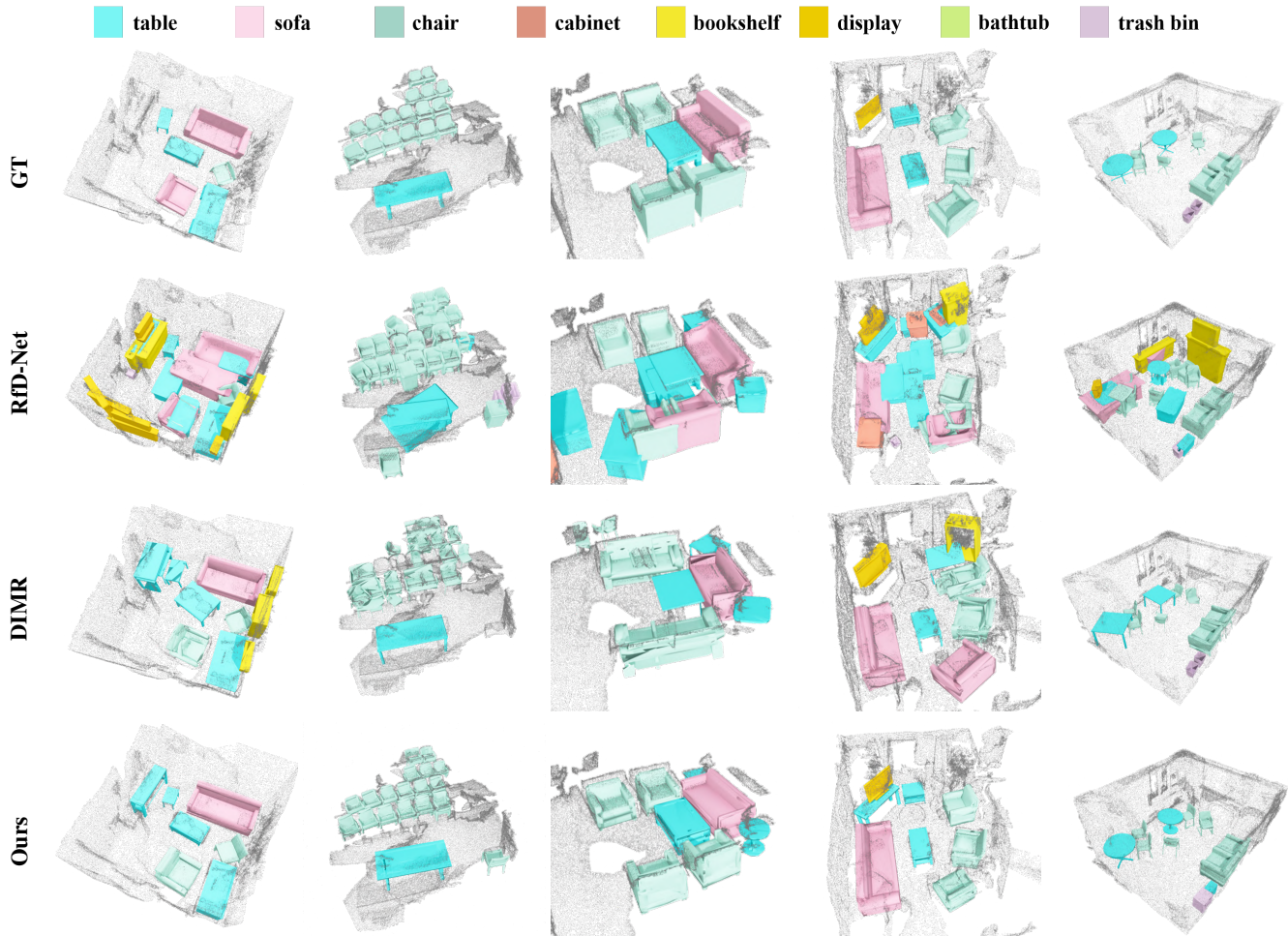


Figure 5. Visualization comparison of scene semantic reconstruction on ScanNetv2.

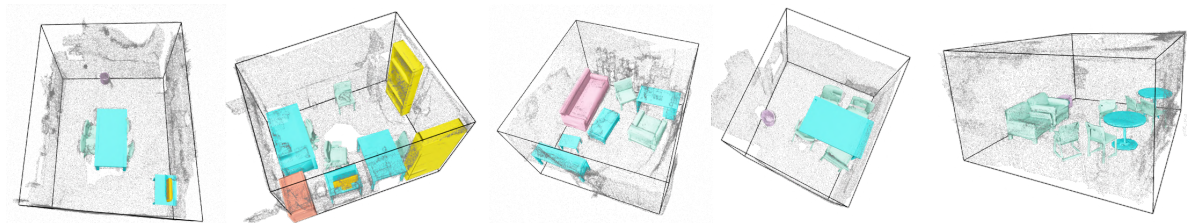


Figure 6. Holistic scene understanding by our proposed AnchorRec. (The comparison methods didn't conduct layout estimation.)

DIMR infers object models from the instance points provided by a complex instance segmentation backbone. As shown in Fig. 5, our AncRec can robustly recognize and reconstruct objects from noisy point clouds with accurate object localization and high-fidelity shape details. With the target-focused proposal feature learning offered by AncLearn in terms of object detection, our method generates fewer false positive models than the compared approaches. With more accurate detection results, our AncRec can leverage outlier-reduced instance priors sampled by AncLearn

to infer more detailed object structures, *e.g.*, the folding chairs are recovered in the second column of Fig. 5. Tab. 1 presents the quantitative comparison of different methods. Our AncRec outperforms other methods under all metrics. As the comparison methods didn't conduct layout estimation, we illustrate our holistic scene understanding results in Fig. 6. Considering that semantic scene reconstruction is a comprehensive problem, we further analyze the performance of our method on the key sub-tasks, *i.e.*, scene parsing and object reconstruction, as follows.

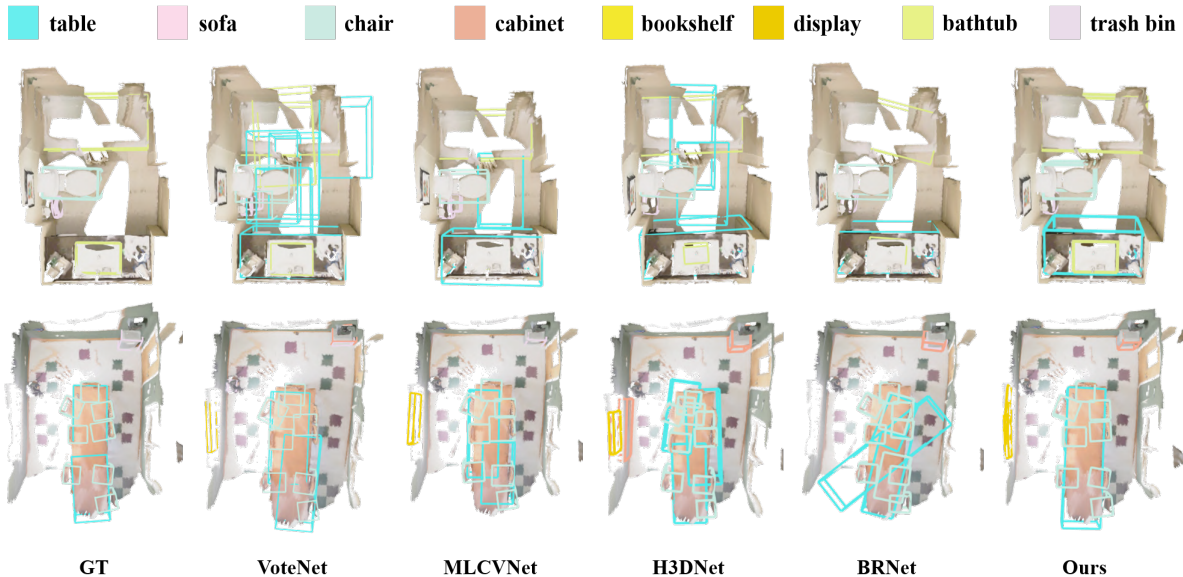


Figure 7. Visualization comparison results of object detection on ScanNetv2

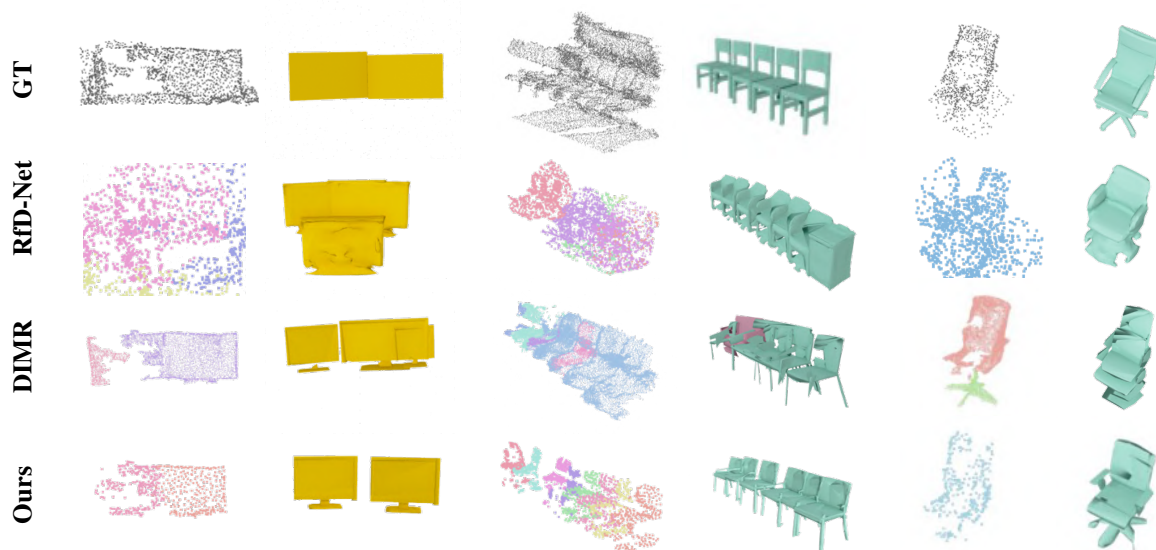


Figure 8. The influence of different instance point sampling strategies on object reconstruction. Points from the same instance are painted in the same color. Without the reliance on segmentation like RfD-Net and DIMR, our method can correctly separate adjacent objects and predict shape models.

Table 1. Quantitative results on the ScanNetv2 dataset. We evaluate the reconstruction quality with mAP at different thresholds. (higher is better)

	IoU@0.25	IoU@0.5	CD@0.1	CD@0.047	LFD@5000	LFD@2500
RfD-Net	42.5	16.9	45.7	19.1	28.6	7.8
DIMR	46.3	12.5	51.9	25.7	29.5	8.6
AnchorRec	52.9	18.9	56.8	29.4	30.3	9.9

Scene Parsing. We compare AncRec with the state-of-the-art methods regarding 8-category object detection and layout estimation. The quantitative results in Tab. 2 show that our AnchorRec edges out the comparison methods by

significant performance gains. Especially for the categories of the display, bathtub, and trash bin, AncRec exceeds the second best by 6.6%, 5.8%, and 10.7% in precision measurement. The results indicate that, with adaptive grouping feature areas, AncLearn works effectively for detecting objects shaped irregularly, *e.g.*, diverse-scaled bathtubs, thin displays, and tiny trash bins. In Tab. 2, we also compare our method with PQ-Transformer [7], the current state-of-the-art scene parsing approach, in terms of layout estimation. The results demonstrate that our AncLearn is also applicable to localizing walls that are particularly large and thin.

Table 2. Quantitative results of object detection and layout estimation. The AP scores are measured with an IoU threshold of 0.5. Higher values mean better results.

Method	Object									Layout
	table	sofa	chair	cabinet	bookshelf	display	bathtub	trash bin	mAP	F1-scores
VoteNet [14]	44.5	47.5	76.3	28.8	20.1	8.0	26.0	14.6	33.2	-
MLCVNet [35]	50.5	46.6	81.8	25.5	19.2	7.8	25.3	19.2	34.5	-
H3DNet [38]	53.6	52.4	76.0	25.6	27.2	9.4	20.7	10.0	34.4	-
BRNet [10]	56.4	43.1	80.9	35.4	26.1	10.4	28.2	11.0	36.4	-
PQ-Transformer [7]	52.7	52.5	81.1	32.1	36.4	6.7	20.6	20.8	37.8	67.0
AnchorRec (Ours)	57.3	56.1	84.5	36.9	29.8	17.0	34.0	31.5	43.4	70.5

Table 3. Numeric results on object reconstruction. The AP scores are measured with CD at a threshold of 0.1. (Higher is better.)

Method	table	chair	bookshelf	sofa	trash bin	cabinet	display	bathtub	mAP
RfD-Net [28]	25.51	82.11	32.53	44.21	44.74	28.37	65.51	42.57	45.70
DIMR [33]	39.44	81.04	38.24	44.09	62.60	23.57	75.12	50.93	51.88
AnchorRec (Ours)	53.60	86.23	36.82	50.92	59.59	39.56	73.68	54.32	56.84

Visualized object detection result provided in Fig. 7 further manifests that our AnchorRec can reliably delineate objects with compact bounding box prediction.

Object Mesh Reconstruction. In Tab. 3, we also evaluate the performance of different methods with respect to class-wise object reconstruction. The evaluation is based on how matchable the predicted object meshes are with the scene-aligned ground truths. The numeric results in Tab. 3 show that AncRec achieves the best performance on 5 categories and the overall chamfer distance-based mAP score.

4.3. Ablation Studies

AncLearn for Scene Parsing. We study different settings of AncLearn for instance detection and report the results in Tab. 4. Compared to the baseline without AncLearn, the proposed AncLearn brings consistent improvement to object detection and layout estimation. Especially, when cooperating with the self-attention layer for layout estimation, AncLearn enables the dual-branch instance detector of AncRec to obtain the best performance in both parsing tasks.

Table 4. Ablation study of the proposed AncLearn. IoU@0.5 and F1 score respectively reflect the performance of object detection and layout estimation. Higher values mean better results.

OD anchor	layout anchor	layout SA	IoU@0.5	F1 score
			41.12	60.62
✓			43.33(+2.21)	60.59(- 0.03)
	✓		40.21(-0.91)	62.61(+ 1.99)
✓	✓		41.24(+0.12)	64.00(+3.38)
		✓	41.16(+0.04)	65.34(+4.72)
✓	✓	✓	43.38(+2.26)	70.45(+9.83)

AncLearn for Object Reconstruction. In the shape predictor of AncRec, AncLearn works to sample instance points as geometry priors for reconstruction. With the in-

stance detector of AncRec, we compare AncLearn with the segmentation-based strategy used in RfD-Net and a box cropping sampling method. The quantitative comparison in Tab. 5 shows that AncLearn outperforms the other two approaches by providing instructive geometry priors with shape anchors in terms of all metrics.

Table 5. Comparison between different instance point sampling methods. Higher values mean better results.

	IoU@0.5	IoU@0.25	CD@0.1
Segmentation [28]	17.4	51.4	55.8
Box Cropping	17.7	51.9	55.5
AncLearn (Ours)	18.9	52.9	56.8
	CD@0.047	LFD@5000	LFD@2500
Segmentation [28]	29.1	29.1	7.8
Box Cropping	28.8	29.9	9.8
AncLearn (Ours)	29.4	30.3	9.9

4.4. Contributions of the Vote and Anchor Features

As mentioned in Sec. 3.1.1, we perform instance detection by taking both vote and anchor features into account. To investigate the difference in their contributions to the detection parameters, we visualize the two weight vectors in Eq. (3) for perceptual comparison. As shown in Fig. 9, vote features have a larger impact on scoring objectness, indicating that the context encoded in vote features is useful in differentiating between object and non-object regions. Compared to vote features, anchor features contribute significantly more to angle prediction, which demonstrates the superiority of anchor-guided strategy in providing shape clues for learning shape-sensitive parameters. Considering the gap between the bounding box and the shape of objects, the vote features contribute more to predicting the bounding

box center and size residuals.



Figure 9. The learned weights for combining vote and anchor features.

4.5. The Characteristics of the Learned Anchors

The supervision with surface points derived from complete 3D models enables the anchors to depict object shapes, even with unobserved structures, which assists in object detection and reconstruction when some parts of objects are missing. There also exists a difference between anchor distribution characteristics for objects and non-object areas. As shown in Fig. 10, the anchors and sampled points depict object shapes, while those for non-object areas scatter irregularly. This characteristic may play a role in differentiating objects from non-object regions.

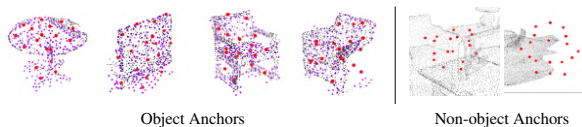


Figure 10. The characteristics of the learned anchors. The anchors and sampled points are shown in red and purple respectively.

5. Conclusions

In this paper, we introduce a shape anchor guided learning strategy (AncLearn) that is embedded into a reconstruction-from-detection learning system (AncRec) to handle the issue of noise interference in point-based holistic scene understanding. Extensive experiments demonstrate that AncRec achieves high-quality indoor semantic scene reconstruction. The quantitative and qualitative results show that AncRec outperforms current methods in terms of object detection, layout estimation, and shape modeling. The ablation studies convincingly verify that AncLearn can largely exclude noise from search space for reliable feature grouping and robust instance point sampling. In the future, it is promising to study the application of shape anchor guided learning strategy in other point-based 3D vision tasks.

Acknowledgment

This research was supported by the Fundamental Research Funds for the Central Universities of China under Grant 2042022dx0001, NSFC-projects under Grant

42071370, Wuhan University-Huawei Geoinformatics Innovation Laboratory, and the Open fund of Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources under Grant KF202106084.

References

- [1] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X Chang, and Matthias Nießner. Scan2cad: Learning cad model alignment in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2614–2623, 2019. 2, 3, 5
- [2] Armen Avetisyan, Angela Dai, and Matthias Nießner. End-to-end cad model retrieval and 9dof alignment in 3d scans. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pages 2551–2560, 2019. 3
- [3] Armen Avetisyan, Tatiana Khanova, Christopher Choy, Denver Dash, Angela Dai, and Matthias Nießner. Scenecad: Predicting object alignments and layouts in rgb-d scans. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 596–612. Springer, 2020. 2, 5
- [4] Aleksei Bokhovkin and Angela Dai. Neural part priors: Learning to optimize part-based object completion in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9032–9042, 2023. 3
- [5] Alexey Bokhovkin, Vladislav Ishimtsev, Emil Bogomolov, Denis Zorin, Alexey Artemov, Evgeny Burnaev, and Angela Dai. Towards part-based understanding of rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7484–7494, 2021. 3
- [6] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5
- [7] Xiaoxue Chen, Hao Zhao, Guyue Zhou, and Ya-Qin Zhang. Pq-transformer: Jointly parsing 3d objects and layouts from point clouds. *IEEE Robotics and Automation Letters*, 7(2):2519–2526, 2022. 4, 5, 7, 8
- [8] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In *European Conference on Computer Vision*, pages 543–560. Springer, 2022. 2
- [9] Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. Bsp-net: Generating compact meshes via binary space partitioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 45–54, 2020. 3, 5
- [10] Bowen Cheng, Lu Sheng, Shaoshuai Shi, Ming Yang, and Dong Xu. Back-tracing representative points for voting-based 3d object detection in point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8963–8972, 2021. 2, 8
- [11] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *Proceedings of the IEEE/CVF con-*

- ference on computer vision and pattern recognition, pages 6970–6981, 2020. 3
- [12] 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
- [13] 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
- [14] Zhipeng Ding, Xu Han, and Marc Niethammer. Votenet: A deep learning label fusion method for multi-atlas segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 202–210. Springer, 2019. 2, 4, 8
- [15] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9031–9040, 2020. 2
- [16] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4857–4866, 2020. 3
- [17] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Aligning 3d models to rgb-d images of cluttered scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4731–4740, 2015. 3
- [18] Xiaoguang Han, Zhaoxuan Zhang, Dong Du, Mingdai Yang, Jingming Yu, Pan Pan, Xin Yang, Ligang Liu, Zixiang Xiong, and Shuguang Cui. Deep reinforcement learning of volume-guided progressive view inpainting for 3d point scene completion from a single depth image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243, 2019. 2
- [19] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3
- [20] 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. 1
- [21] Ji Hou, Angela Dai, and Matthias Nießner. Revealnet: Seeing behind objects in rgb-d scans. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2095–2104, 2019. 3
- [22] 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
- [23] Linxi Huan, Xianwei Zheng, and Jianya Gong. Georec: Geometry-enhanced semantic 3d reconstruction of rgb-d indoor scenes. *ISPRS Journal of Photogrammetry and Remote Sensing*, 186:301–314, 2022. 1, 3
- [24] Hamid Izadinia, Qi Shan, and Steven M Seitz. Im2cad. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5134–5143, 2017. 3
- [25] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128:261–318, 2020. 2
- [26] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4460–4470, 2019. 3
- [27] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1, 3
- [28] Yinyu Nie, Ji Hou, Xiaoguang Han, and Matthias Nießner. Rfd-net: Point scene understanding by semantic instance reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4608–4618, 2021. 1, 2, 3, 5, 8
- [29] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Invotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4404–4413, 2020. 2
- [30] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1, 2, 5
- [31] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1, 2
- [32] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019. 2
- [33] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Point scene understanding via disentangled instance mesh reconstruction. *arXiv preprint arXiv:2203.16832*, 2022. 1, 2, 3, 5, 8
- [34] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Dening Lu, Mingqiang Wei, and Jun Wang. Venet: Voting enhancement network for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3712–3721, 2021. 2
- [35] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. Mlcvnet: Multi-level context votenet for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10447–10456, 2020. 2, 4, 8

- [36] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Simon Chen, Yifan Liu, and Chunhua Shen. Towards accurate reconstruction of 3d scene shape from a single monocular image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):6480–6494, 2022. [2](#)
- [37] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8833–8842, 2021. [1](#), [3](#)
- [38] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020. [2](#), [8](#)
- [39] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. [1](#)
- [40] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. [2](#)