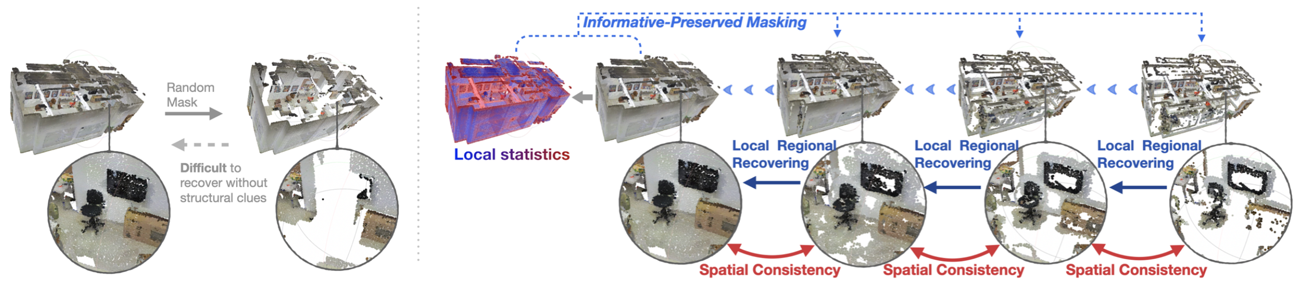
MM-3DScene: 3D Scene Understanding by Customizing Masked Modeling with Informative-Preserved Reconstruction and Self-Distilled Consistency

Mingye Xu[1,3,4*], Mutian Xu[2,5*], Tong He[4], Wanli Ouyang[4], Yali Wang[1,4†], Xiaoguang Han[2,5], Yu Qiao[1,4†]

[1] The Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology,
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[2] SSE, CUHKSZ [3]University of Chinese Academy of Sciences
[4]Shanghai Artificial Intelligence Laboratory [5] FNii, CUHKSZ

mingyexu.github.io/mm3dscene



(a) Random masked modeling for large-scale 3D scenes. (b) Our MM-3DScene masked modeling for large-scale 3D scenes.

Figure 1. ***How to apply masked modeling for large-scale 3D scenes?*** (a) Conventional random masked modeling on 3D scenes may cause a high risk of uncertainty.In this figure, a ***chair*** and a ***TV*** are totally masked, which are extremely difficult to be recovered without any context guidance. **(b)** Our **MM-3DScene** exploits local statistics to discover and *preserve* representative structured points, effectively simplifying the pretext task. At each learning step, our method focuses on restoring *regional* geometry, and enjoys less ambiguity. Moreover, since unmasked areas are underexplored during reconstruction, the model is encouraged to maintain the intrinsic *spatial consistency* on unmasked points between different masking ratios, which requires the consistent understanding of unmasked areas.

Abstract

Masked Modeling (MM) has demonstrated widespread success in various vision challenges, by reconstructing masked visual patches. Yet, applying MM for large-scale 3D scenes remains an open problem due to the data sparsity and scene complexity. The conventional random masking paradigm used in 2D images often causes a high risk of ambiguity when recovering the masked region of 3D scenes. To this end, we propose a novel informative-preserved reconstruction, which explores local statistics to discover and preserve the representative structured points, effectively enhancing the pretext masking task for 3D scene understanding. Integrated with a progressive reconstruction manner, our method can concentrate on modeling regional geometry and enjoy less ambiguity for masked reconstruction. Besides, such scenes with progressive masking ratios can also serve to self-distill their intrinsic spatial consistency, requiring to learn the consistent representations from unmasked areas. By elegantly combining informative-preserved reconstruction on ***masked*** *areas and consistency self-distillation from* ***unmasked*** *areas, a unified framework called* ***MM-3DScene*** *is yielded. We conduct comprehensive experiments on a host of downstream tasks. The consistent improvement* (e.g., *+6.1% mAP@0.5 on object detection and +2.2% mIoU on semantic segmentation) demonstrates the superiority of our approach.*

1. Introduction

3D scene understanding plays an essential role in various visual applications, such as virtual reality, robot navigation, and autonomous driving. Over the past few years, deep learning has dominated 3D scene parsing tasks [26, 40, 73]. However, traditional supervised learning methods require massive annotation of 3D scene data that are extremely la-

* Equal contribution.
† Corresponding authors.

borious to obtain [9], where millions of points or mesh vertices per scene need to be labeled.

To solve this, self-supervised learning (SSL) becomes a favorable choice since it can extract rich representations without any annotation [10, 16, 17]. Masked Modeling (MM) [16, 57], as one of the representative methods in SSL, recently draws significant attention in the vision community. Recently, It has been explored in 3D vision [31,39,51,67,69,70], where these 3D MM approaches randomly mask local regions of point clouds, and pre-train neural networks to reconstruct the masked areas. Nevertheless, such random masking paradigms are not feasible for large-scale 3D scenes, which often causes a high risk of reconstruction ambiguity. As illustrated in Fig. 1 (a), a chair and a TV are totally masked, which are extremely difficult to be recovered without any context guidance. Such ambiguity often makes MM difficult to learn informative representation for 3D scenes. Hence, we ask a natural question: *can we customize a better way of masked modeling for 3D scene understanding?*

To tackle this question, we propose a novel informative-preserved masked reconstruction scheme in this paper. Specifically, we leverage local statistics of each point (*i.e.*, the difference between each point and its neighboring points in terms of color and shape) as guidance to discover the representative structured points which are usually located at the boundary regions in the 3D scene. We denote these points as '**Informative Points**' since they provide highly useful information hints and rich semantic context for *assisting* masked reconstruction. To this end, our mask strategy is definite: to *preserve* Informative Points in a scene and mask other points. In this way, the basic geometric information of a scene is explicitly *retained*, which effectively simplifies the pretext task and reduces ambiguity.

Based on our mask strategy, a progressive masked reconstruction manner is integrated, to better model the *masked* areas. As illustrated in Fig. 1 (b), during each iteration, our method *concentrates* on reconstructing the local *regional* geometric patterns rather than rebuilding the original intact scene. In doing so, it enjoys less ambiguity and is able to restore accurate geometric information.

Moreover, we realize the information of *unmasked* areas (*i.e.*, Informative Points) is underexplored. We find that there exists point correspondence in the unmasked areas under progressive masking ratios. Accordingly, we introduce a dual-branch encoding scheme for learning such intrinsic consistency, with the ultimate goal of unearthing the consistent (*i.e.*, masking-invariant) representations from unmasked areas. This leads to a more powerful SSL framework on 3D scenes, called MM-3DScene, which elegantly combines the masked modeling on the masked and unmasked areas in 3D scenes together, while *complements* each other. It achieves superior performance in Table 7 (v).

Datasets	Complexity	Task	Gain (from scratch)
S3DIS	Entire floor, office	segmentation	(+1.5%) mIoU
ScanNet	Large rooms	segmentation	(+2.2%) mIoU
		detection	(+4.4) mAP@0.25
		detection	(+6.1) mAP@0.5
SUN-RGBD	Cluttered rooms	detection	(+2.9) mAP@0.25
		detection	(+4.4) mAP@0.5

Table 1. **Summary of fine-tuning MM-3DScene** on various downstream tasks and datasets for 3D understanding. Our MM-3DScene conspicuously boosts the performance of the baseline model trained from scratch.

Our contributions are motivated and comprehensive:

- We raise the concept of Informative Points – the points providing significant information hints, and indicate that preserving them is critical for assisting masked modeling on 3D scenes (Table 8).
- For masked areas, we propose an informative-preserved reconstruction scheme to focus on restoring the regional geometry in a novel progressive manner, which explicitly simplifies the pretext task.
- For unmasked areas, we introduce a self-distillation branch, which is encouraged to learn spatial-consistent representations under progressive masking ratios.
- A unified self-supervised framework, called MM-3DScene, delivers performance improvements across on various downstream tasks and datasets (Table 1).

2. Related Work

3D Scene Understanding With the rapid development of deep learning methods in point cloud analysis and the emergence of large-scale 3D datasets [1,4,9,15,58], the research focus gradually migrate from synthetic, single object analysis [32,35,55,60,61,63] to complex large-scale scene understanding, especially scene segmentation [13, 18–20, 62, 72], 3D object detection [33, 37, 40]. PointNet [42] and its variants [29, 43, 52, 53] extract local features from neighbors through hierarchical grouping architecture to capture fine-grained representation. Thomas *et al*. [49] defines deformable kernel point convolution to capture the point cloud representation using a set of learnable kernel points. In PAConv [59], a continuous convolutional kernel is built by dynamically combining several weight banks, where the coefficients are learned from point positions. Zhao *et al*. [73] proposes to assign learnable attentional weights to local point features, and introduces a Transformer [50]-like architecture into point cloud analysis. Rather than focusing on developing deep architectural details, in this paper, we explore an effective self-supervised pre-training mechanism for scene understanding based on the basic version of Point Transformer.

Unsupervised Scene Pre-training Self-supervised learning (SSL) has recently achieved great success in 2D vision [2, 5, 6, 14, 16, 17, 57] and NLP tasks [3, 10, 11]. But there has been limited exploration of SSL for 3D vision. Most of the existing 3D-SSL methods aim to understand 3D point clouds, and can be divided into two types, masked modeling (MM)-based methods, and contrastive learning-based methods, respectively.

MM-based pre-training usually takes as input point clouds to recover itself as the pretext task. OcCo [51] first constructs an occlusion point cloud and applies an encoder-decoder mechanism to reconstruct the original object. Yu *et al*. [67] proposes the idea of restoring the masked proxies under the supervision of the pre-trained tokenizer. Borrowing the idea from MAE [16], Pang *et al*. [39] designs a masked auto-encoder to recover the masked parts of objects. However, most of these MM-based methods are focused on 3D shape-level pre-training, how to apply such a scheme on 3D scenes has not been fully investigated. Although some concurrent works [21, 36] voxelize the outdoor automative point clouds and randomly mask voxels, they are only manifested to especially benefit the LiDAR-based object detection task. In this paper, we creatively apply the self-supervised pre-training directly on 3D scenes by a scene-specific masked reconstruction.

On the other hand, using the idea of contrastive scene contexts for pre-training has been shown to be effective [7, 24, 56]. Xie *et al*. [56] and Hou *et al*. [24] perform scene pre-training using contrastive learning of point features between a pair of overlapping scans. 4DContrast [7] first composites the synthetic 3D object with real-world 3D scans to create 4D sequential data, and utilizes the contrastive loss to learn 4D invariance constraints. For our MM-based framework, we leverage the momentum contrast [17] to self-distill the consistency between the same scenes under different masking degrees, for unearthing the hidden information from unmasked areas.

3. Method

3.1. Overview

As shown in Fig. 3, we propose MM-3DScene for masked modeling on 3D scenes, which can effectively alleviate the uncertainty and unpredictability during masked reconstruction for enhancing the self-supervised pre-training on complex 3D scenes.

Firstly, we propose a Local-Statistics-Guided masking strategy in Sec. 3.2, which explores the local difference for discovering and preserving Informative Points during the scene masking. Next, our mask strategy is integrated with a progressive reconstruction manner, yielding an informative-preserved reconstruction scheme. Finally, we introduce a self-distillation branch for learning the spatial consistency in the unmasked areas under progressive masking ratios. By elegantly combining informative-preserved reconstruction and self-distilled consistency, MM-3DScene is yielded. The framework details are presented in Sec. 3.3.

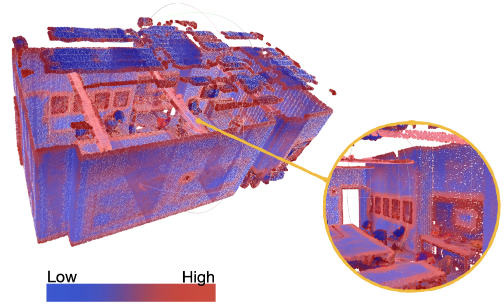


Figure 2. **The heat map of local statistics on 3D scenes**. Points with high statistics value also provide highly important semantic information for understanding or reconstructing 3D scenes.

3.2. Local-Statistics-Guided Masking

As we analyzed in Sec. 1, the traditional random mask strategy is not feasible for complex large-scale 3D scenes, which often causes a high risk of ambiguity during masked reconstruction. In order to explore a more effective masked modeling mechanism for 3D scenes, we need to first design a better mask strategy, aiming to *reduce the ambiguity* during the masked reconstruction pretext task.

In 3D scenes, some representative structured points in provide highly important information hints and rich semantic context for assisting the scene understanding or reconstruction tasks. Accordingly, preserving them may help a lot to simplify the reconstruction. Thus, the first question is: *how to find such points?*

In this paper, we adopt local statistics as straightforward guidance to discover the representative points. For complex 3D scenes, we use the local difference of each point (*i.e*., the difference between each point and its neighboring points in terms of colors and coordinates) to calculate the local statistics of each point. Specifically, for each point p_i in the scene, we first use K-Nearest Neighborhood (KNN) search to obtain its K neighboring points p_{ik} in Euclidean space, and then calculate the local difference to denote point statistics:

$$D_q(p_i) = \sum_{k=1}^{K} \|p_{i,q} - p_{i_k,q}\|_2 \tag{1}$$

where $p_{i,q} \in \mathrm{R}^{1 \times C_q}$ is point statistic of p_i. Specifically, $p_{i,0}$ and $p_{i,1}$ are the point coordinates and colors, which are normalized and accumulated together as the local statistics. $D = \sum_{q=0}^{1}(\alpha_q \times norm(D_q))$. where $D \in \mathrm{R}^{N \times 1}$, N is the number of the points in the scene.

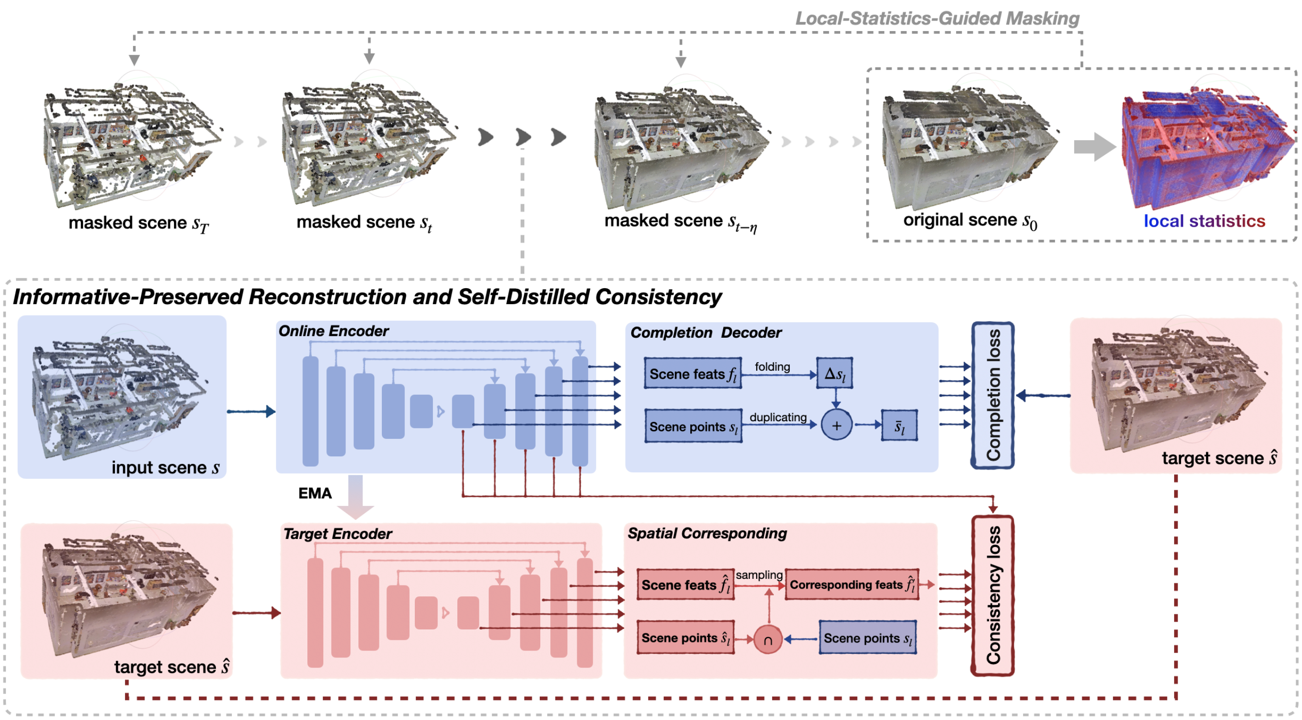


Figure 3. Overview of our customized masked modeling framework **MM-3DScene**. We first propose the Local-Statistics-Guided masking strategy to discover and preserve the representative structured points. This mask strategy yields an Informative-Preserved Reconstruction, where our method focuses on restoring the *regional* geometric patterns of **masked areas** at each learning step, enhancing the pretext masking task with less ambiguity. Moreover, since **unmasked areas** are underexplored during reconstruction, we introduce a self-distillation branch to maintain the intrinsic *spatial consistency* under progressive masking ratios, which enables MM-3DScene to learn consistent (*i.e.*, masking-invariant) representations of the unmasked areas.

For a clearer picture, we visualize the heat map of local statistics on the 3D scene in Fig. 2, it can be seen that points with high statistics value are concentrated in the foreground objects or contours of the scene (*red regions*), and these areas are relatively more informative and provide highly important information hints for understanding or reconstructing a 3D scene, denoted by **'Informative Points'** in this paper. Informative Points provide highly useful information hints and rich semantic context for *assisting* masked reconstruction. To this end, our mask strategy is definite: to *preserve* Informative Points in a scene and mask other points. Through this, the basic geometric information of a scene is explicitly *retained*, which effectively reduces ambiguity in the pretext masking task.

3.3. MM-3DScene

After introducing the mask strategy with local statistics guidance, we illustrate the details of our MM-3DScene – a self-supervised masked modeling framework customized for 3D scene understanding. Our pre-training architecture consists of two parts: informative-preserved reconstruction and self-distillation of spatial consistency, respectively aiming at better modeling of masked areas and unmasked areas in 3D scenes. These two parts are elegantly combined while *complementing* each other in our MM-3DScene.

Informative-preserved reconstruction. Based on the proposed mask strategy, we introduce an informative-preserved reconstruction manner.

To begin with, for the original scene s_0, we align all scene points in descending order according to the local statistics D. Then we use the *incremental* masking ratio $\theta = \{\theta_1, ..., \theta_t, ..., \theta_T\}$ to *progressively* mask the scene s_0 according to D from low to high, forming the scene sequence $\mathcal{S} = \{s_1, ...s_{t-1}, s_t, ..., s_T\}$ as shown in the Fig. 3. It can be seen that in the masked sequence $\mathcal{S}$, the masked regions are gradually shifted from the background surface to the foreground objects, and the representative structural points are preserved. It is also worth noting that scene s_t is the *subset* of scene s_{t-1}.

During each training iteration, we randomly select a masked scene s_t in sequence $\mathcal{S}$ as the input, and take a *relatively* more complete one $s_{t-\eta}$ as the target for masked reconstruction, where η indicates the masked gap to be recovered (to facilitate the subsequent description, we later define the input scene as $s = s_t$ and the target as $\hat{s} = s_{t-\eta}$). To encode point-wise representations from the input scene s, we utilize a well-established network Φ_{OE} as our backbone. Φ_{OE} is a point based feature extractor with hierarchical structure. The hierarchical encoded features $\mathcal{F}$ of scene s can be represented as $\mathcal{F} = \Phi_{OE}(s)$ where $\mathcal{F} = \{f_1, ..., f_l, ..., f_L\}$, and L is the layers of Φ_{OE}.

Then we utilize the features extracted from the online encoder to learn the coordinate variations for scene reconstruction. Note that the number of output points $\hat{s}$ is larger than that of input points s. Before generating coordinate variations via MLP, we follow SnowFlakeNet [54] to replicate the coordinates and features, producing more points for outputs. Specifically, for l-th layer, we have the displacement feature f_l and scene points s_l. Referring to FoldingNet [65], we incorporate a standard two-dimensional grid I to the displacement feature f_l, and then use two consecutive Multi-Layer Perceptrons (MLPs) [22] to generate coordinate variation Δs_l:

$$\Delta s_l = \Psi_2(f_l \oplus \Psi_1(f_l \oplus I)), \tag{2}$$

where Ψ_1 and Ψ_2 are two 3-layer MLPs, and $\oplus$ is the concatenation operator. Then, we add the coordinate variation Δs_l with the duplicated input mask scene s, which generates the predicted scene $\bar{s}_l$ in the l-th layer. Finally, in order to train the masked reconstruction task more efficiently, we choose the multi-scale symmetrical chamfer distance [48] as the reconstruction loss, with the following details:

$$\mathcal{L}_{PC} = \sum_{l}^{L}(\frac{1}{|\bar{s}_l|}\sum_{x\in\bar{s}_l}\min_{y\in\hat{s}}\|x-y\|^2 + \frac{1}{|\hat{s}|}\sum_{y\in\hat{s}}\min_{x\in\bar{s}_l}\|x-y\|^2). \tag{3}$$

The proposed reconstruction scheme encourages the model to *focus* on restoring the *regional* geometry in a novel progressive manner, which enjoys less ambiguity and is able to restore accurate geometric information, so to enhance the modeling on masked areas.

Consistency self-distillation. Moreover, we realize the information of unmasked areas (*i.e*., Informative Points) is underexplored. We find that there exists point correspondence in the unmasked areas under progressive masking ratios. Leveraging this unique behaviour of our method, and based on a generalized model of teacher-student mutual learning [12, 68], we introduce a self-distillation branch to maintain the intrinsic spatial consistency on unmasked areas during progressive reconstruction.

As Fig. 3 shows, we treat the online encoder as the student model and maintain a teacher model as the target encoder. It is worth mentioning that the target encoder has the same parameters and structure as the online encoder, but does not participate in the back-propagation of the gradient. The parameters of the target encoder are dynamically updated by the Exponential Moving Average (EMA) [12]:

$$\mathcal{W}_{\text{target}} \leftarrow [\beta\mathcal{W}_{\text{target}} + (1-\beta)\mathcal{W}_{online}], \tag{4}$$

where $\mathcal{W}_{\text{online}}$ and $\mathcal{W}_{\text{target}}$ are the parameters of online encoder and target encoder.

Next, we feed the target scene $\hat{s}$ to the target encoder and extract the target feature $\hat{\mathcal{F}} = \{\hat{f}_1, ..., \hat{f}_l, ..., \hat{f}_L\}$, where $\hat{f}_l \in \mathrm{R}^{\hat{N}_l\times C}$. Considering the input scene is the subset of target scene, where $s \subset \hat{s}$, we can select a subset of the target feature $\hat{f}_l$ that have natural spatial correspondence with the online feature f_l. We define such subset of target feature as $\hat{f}'_l \in \mathrm{R}^{N_l\times C}$, where $N_l < \hat{N}_l$. Finally, we use info-NCE loss [38] to model the spatial consistency of the feature representation between the input scene and the target scene, which can be formulated as:

$$\mathcal{L}_{CSD} = -\sum_{l}^{L}\sum_{(i,j)\in s_l}\log\frac{\exp\left(f_{i,l}\cdot\hat{f}'_{j,l}/\tau\right)}{\sum_{(\cdot,k)\in s_l}\exp\left(f_{i,l}\cdot\hat{f}'_{k,l}/\tau\right)}, \tag{5}$$

where the input scene s_l is the set of points that can find the spatial correspondences in the target scene $\hat{s}_l$. For online feature $f_{i,l}$ of point p_i, we take the corresponding point feature $\hat{f}'_{j,l}$ in the target scene as the positive sample, and use feature $\hat{f}'_{k,l}$ as the negative samples, where $\exists(\cdot,k) \in s_l$ and $k \neq j$.

Through self-distillation, our method is able to unearth the spatial-consistent (*i.e*., masking-invariant) representations from unmasked areas.

MM-3DScene. Ultimately, by elegantly combining informative- preserved reconstruction on masked areas and consistency self-distillation from unmasked areas, a unified framework called MM-3DScene is yielded. The final pre-training loss can be denoted by $\mathcal{L} = \zeta_1\times\mathcal{L}_{PC} + \zeta_2\times\mathcal{L}_{CSD}$, We find that setting both ζ_1 and ζ_2 to 1 yields the best result, which means that the masked reconstruction and the consistency distillation share *balanced importance* for our framework.

4. Experiments

We pre-train our MM-3DScene and demonstrate its effectiveness on a variety of generic downstream tasks for 3D scene understanding, including 3D semantic segmentation, 3D object detection, as well as data-efficient setup and linear probing evaluation.

4.1. Experimental Setup

Datasets. Following the state-of-the-art 3D self-supervised pre-training methods [7, 24, 56], we pre-train the proposed MM-3DScene on ScanNetv2 [9] dataset with 1201 train scans. As for 3D object detection, we fine-tune the pre-trained weights on ScanNetv2 training set (inner-domain) and evaluate on the corresponding validation set with 312 scenes. We also fine-tune our model on SUN RGB-D [46] training set for cross-domain evaluation. SUN RGB-D covers fully 3D object bounding box annotations for 10 categories, with 5,285 train frames and 5,050 test

Method	mAP@0.25	mAP@0.5
DSS [47]	15.2	6.8
F-PointNet [41]	19.8	10.8
GSPN [66]	30.6	17.7
3D-SIS [23]	40.2	22.5
VoteNet [40] (scratch)	58.7	35.4
RandomRooms [44] + VoteNet	61.3 (+2.6)	36.2 (+0.8)
PointContrast [56] + VoteNet	58.5 (-0.2)	38.0 (+2.6)
CSC [24] + VoteNet	-	39.3 (+3.9)
4DContrast [7] + VoteNet	-	40.0 (+4.6)
MM-3DScene (w/o L_{CSD}) + VoteNet	**61.9 (+3.2)**	**41.3 (+5.9)**
MM-3DScene + VoteNet	**63.1 (+4.4)**	**41.5 (+6.1)**

Table 2. 3D object detection results on ScanNetv2.

Method	mAP@0.25	mAP@0.5
DSS [47]	42.1	-
COG [45]	47.6	-
2D-driven [28]	45.1	-
F-PointNet [41]	54.0	-
VoteNet [40] (scratch)	57.7	32.9
PointContrast [56] + VoteNet	57.5 (-0.2)	34.8 (+1.9)
RandomRooms [44] + VoteNet	59.2 (+1.5)	35.4 (+2.5)
CSC [24] + VoteNet	-	36.4 (+3.5)
4DContrast [7] + VoteNet	-	38.2 (+5.3)
MM-3DScene + VoteNet	**60.6 (+2.9)**	**37.3 (+4.4)**

Table 3. 3D object detection results on SUN RGB-D.

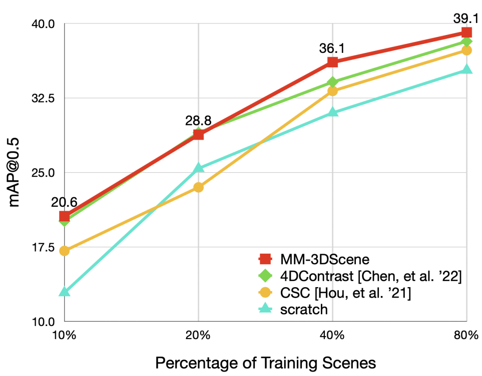


Figure 4. Data-efficient 3D object detection results on ScanNetv2.

frames. As for downstream semantic segmentation, in addition to conducting the experiments on the ScanNetv2, we also evaluate our pre-trained models on S3DIS [1], which contains 271 rooms from 3 different buildings.

Backbone networks. In 3D object detection, we follow [7, 24, 56] to use VoteNet [40] as the backbone. We *directly* perform MM-3DScene pre-training on the original PointNet++ layers, for minimizing the backbone effect on performance gains and better verifying the stand-alone effectiveness of pre-trained representations. Besides, we follow the data augmentations in the original backbone networks. For the scene semantic segmentation, we advocate using Point Transformer [73] as our backbone, considering

Method	mAcc	mIoU
PointNet++ [43]	-	53.5
PointConv [53]	-	61.0
PointASNL [64]	-	63.5
FPConv [30]	-	64.4
KPConv [49]	-	69.2
SR-UNet [8]	78.1	70.0
CSC [24] + SR-UNet	78.8 (+0.7)	70.7 (+0.7)
PointContrast [56] + SR-UNet	79.3 (+1.2)	71.3 (+1.3)
4DContrast [7] + SR-UNet	80.8 (+2.7)	72.3 (+2.3)
PointTrans [73] (scratch)	79.6	70.6
PointMAE [39] * + PointTrans	79.6 (+0.0)	70.6 (+0.0)
PointContrast [56] * + PointTrans	80.0 (+0.4)	70.9 (+0.3)
MM-3DScene (w/o L_{CSD}) + PointTrans	**81.2 (+1.6)**	**72.1 (+1.5)**
MM-3DScene + PointTrans	**82.0 (+2.4)**	**72.8 (+2.2)**

Table 4. 3D semantic segmentation results on ScanNetv2.

Method	mAcc	mIoU
PointNet [42]	49.0	41.1
PointCNN [29]	63.9	57.3
PointWeb [72]	66.6	60.3
PAConv [59]	73.0	66.6
KPConv [49]	72.8	67.1
PointTrans [73] (scratch)	76.5	70.4
PointMAE [39] * + PointTrans	76.4 (-0.1)	70.4 (+0.0)
PointContrast [56] * + PointTrans	76.9 (+0.4)	70.7 (+0.3)
MM-3DScene + PointTrans	**78.0 (+1.5)**	**71.9 (+1.5)**

Table 5. 3D semantic segmentation results on S3DIS Area-5.

its lightweight and effective attributes for 3D scene understanding (Table 6). We retain its original network architectures, and discard the output head in the pre-training process in order to obtain the fine-grained feature representation. The output head will be added to the network structure for the downstream training.

Implementation details. For pre-training, we use AdamW [34] optimizer with a weight decay of 0.0005. The initial learning rate is set to 0.001 and decayed at the 60% and 80% epochs. We pre-train the network for 300 epochs with a batch size of 8. The reconstruction gap η is set to 0.1. We set the scale parameter τ as 1.

For downstream semantic segmentation, we follow the setting of Point Transformer [73], and utilize SGD optimizer with a momentum of 0.9 and weight decay of 0.0001. In addition to the standard data augmentation, we also use random cropping of the scene for more effective training. For S3DIS, we fine-tune our model for 150 epochs with a data loop of 30, where the voxel size is 4cm. The data loop for ScanNet is set to 6 and the scene resolution is 2cm.

For downstream object detection, the fine-tuning settings all follow 4DContrast [7], where the networks are trained for 500 epochs, the initial learning rate is set to 0.001 and decayed by a factor of 0.5 at epoch 250, 350, 450. The batch size is 6 for ScanNet and 16 for SUN RGB-D.

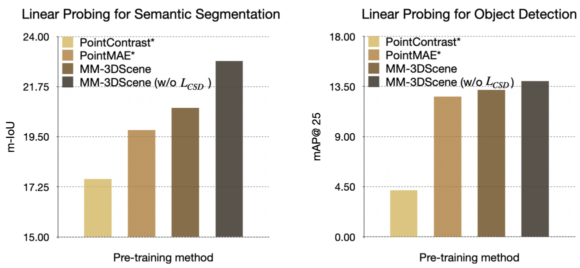


Figure 5. Linear probing on semantic segmentation and object detection.

4.2. 3D Object Detection

ScanNetv2 object detection. 3D object detection is a widely used downstream task for scene understanding. We report the fine-tuning results of object detection for ScanNetv2 in the Table 2, our MM-3DScene achieves the state-of-the-art performance among the SSL methods. With of MM-3DScene pre-training, it achieves a significant improvement of **4.4** on mAP@0.25, and **6.1** on mAP@0.5 compared to training from scratch. Meanwhile, even if we train without the consistency distillation, the result also surpasses the state-of-the-art method 4DContrast [7], which strongly demonstrates the effectiveness of our customized masked modeling method for object detection. Note that, *without* RGB, our method still gets **41.5** mAP@0.5 on ScanNetv2 detection, which is much higher than 4DC (40.0) and PC (38.0). Moreover, we conduct experiments with the updated H3DNet [71] as the backbone and achieve significant improvements, which can be found in appendix.

SUN RGB-D object detection. We also conduct the cross-domain experiments on SUN RGB-D dataset for object detection, which is shown in Table 3. We pre-train our method on ScanNetv2 and fine-tune on SUN RGB-D, for which our method substantially exceeds the training-from-scratch baseline and surpasses 3D-based pre-training of CSC [24], PointContrast [56] and RandomRooms [44]. *Notably*, pretraining 4DContrast [7] requires a *prerequisite* non-trivial generation of spatio-temporal correspondence, while our method produces the masked scenes *on the fly*.

Data-efficient evaluation. As shown in Fig. 4, We evaluate our method for fine-tuning with limited training data, where the results of CSC and 4DContrast are officially borrowed from [7]. Our MM-3DScene surpasses both of them in almost all data-efficient settings. It is worth noting that our method achieves **39.1** mAP@0.5 when finetuning with only **80**% training data, which even exceeds many methods trained on 100% of the training data in Table 2.

4.3. 3D Semantic Segmentation

ScanNetv2 semantic segmentation. Semantic segmentation is a common but challenging task. It requires models

Method	Model Size	mIoU
SR-UNet [8] (scratch)	37.85M	70.0
4DContrast (3D) [7] + SR-UNet	70.85M (+33M)	71.7 (+1.7)
4DContrast (4D) [7] + SR-UNet	75.85M (+38M)	72.3 (+2.3)
PointTrans [73] (scratch)	7.76M	70.6
MM3DScene + PointTrans	8.63M (+0.87M)	72.8 (+2.2)

Table 6. Comparisons of model parameters for different pre-training methods.

to predict the semantic classes of each point, which is involved in the fine-grained understanding of 3D scenes. Table 4 lists the downstream semantic segmentation results on ScanNetv2. For SR-UNet based methods, we directly quote the results from [7], which do not use the RGB information as input. However, for Point Transformer backbone, the RGB information is indispensable and omitting it will cause a deteriorated performance drop. Thus, we follow its original input setting (points + rgb) and apply various pre-training methods on it. We denote the methods with * when they are adapted to Point Transformer backbone, e.g., "PointMAE [39]* + PointTrans". The table shows that the random masking strategy of PointMAE does not help the downstream segmentation task. We also observe that PointContrast* [56] has a marginal improvement over the scratch model, which may indicate that contrastive learning is good at learning discriminative features but not robust to occlusions (as suggested by [25, 27]). In contrast, our MM-3DScene achieves superior results over these conventional pre-training methods. Furthermore, our MM-3DScene also outperforms 4DContrast with SRUNet backbone, while having a much lighter model size (Table 6).

S3DIS semantic segmentation. To further validate the effectiveness of our method, we conduct experiments on the S3DIS dataset [1], the results are shown in Table 5. We achieve a competitive result of 71.9% mIoU on S3DIS Area-5, which gets a relative improvement of 1.5% from scratch. The qualitative results are illustrated in appendix.

4.4. Linear Probing

Linear probing is *underexplored* in 3D scene understanding, we remedy this defect by locking the pre-trained backbone network, and only fine-tuning the segmentation head and detection head. The experiments are conducted on S3DIS Area-5 and the ScanNetv2 validation set. Fig. 5 shows results following different pre-training methods. Compared to the conventional random masking method (*i.e.*, PointMAE [39]) or contrastive-based method (*i.e.*, PointContrast [56]), our MM-3DScene performs best on the linear probing task. It is worth noting that our framework without learning spatial consistency performs better on linear probing, we guess the reason is that *fewer constraints* can enhance the generalizability.

4.5. Ablation Studies

Why MM-3DScene is a better way for masked modeling on large-scale scenes? Table 7 shows the ablation study of our MM-3DScene framework on S3DIS semantic segmentation. When we only use the informative masked modeling, (ii) achieves an improvement of 0.7% mIoU over the train-from-scratch method. On the other hand, focusing on the unmasked areas (setting iii), we only learn the spatial consistency of the unmasked informative points during each iteration, which still makes performance gains for downstream segmentation. Finally, setting iv integrates informative-preserved modeling of the masked areas and spatial consistency learning on unmasked areas together, which further boosts the result to 71.9% mIoU.

How MM-3DScene simplifies the pretext task? Table 8 shows the ablation study on mask strategies. Comparing the settings of (a) and (d) in Table 8, we found that our informative-preserved masked modeling can significantly improve the downstream performance, while random mask modeling does not. The possible reason is that such masking will randomly drop Informative Points, thus bringing great uncertainty to the pretext task. With the same consistency loss, PointMAE gets 71.0%, which is 0.9% lower than ours. While our approach can make the masked areas perceptible via informative-preserved masking. Instead, when we use the *informative-abandoned* masking setting (c) by masking off the informative structured points, the downstream performance is significantly reduced, demonstrating that masking the Informative Points may cause large unpredictability.

Moreover, different from (d), setting (e) utilizes the progressive manner by reconstructing a masked scene into a more complete one, and shows a significant improvement. We also apply the progressive reconstruction manner based on the random masking (setting (b)), and the results are improved accordingly.

Memory cost. Current 3D pre-training methods [7, 24, 56] basically adopts SR-UNet [8] as the backbone with 37.85M parameters, We advocate the much more lightweight Point Transformer [73], with just 7.76M parameters. In Table 6, we compare our model size with 4DContrast [7] which uses SR-UNet. Our MM-3DScene *only adds* 0.87M parameters to the backbone. Besides, our approach does not require the time-consuming pre-generation of contrastive point cloud pairs for pretraining used in PointContrast [56] and CSC [24]. Our masked scenes are generated *on the fly* during network training.

	Pre-training method	Pre-training loss	mIoU	mAcc
i	Scratch	-	70.4	76.5
ii	MM-3DScene (MM only)	$\mathcal{L}_{PC}$	71.1 (+0.7)	77.2 (+0.7)
iii	MM-3DScene (consistency only)	$\mathcal{L}_{CSD}$	70.9 (+0.5)	77.1 (+0.6)
iv	MM-3DScene	$\mathcal{L}_{PC}, \mathcal{L}_{CSD}$	71.9 (+1.5)	78.0 (+1.5)

Table 7. Ablation study on MM-3DScene framework on S3DIS semantic segmentation.

	Masked Modeling	Mask Strategy	Progressive	mIoU	mAcc
	Scratch	-	-	70.4	76.5
(a)	PointMAE* [39]	Random	×	70.4	76.4
(b)	PointMAE* [39]	Random	✓	70.6	76.5
(c)	MM-3DScene (MM only)	Informative-abandoned	×	70.2	76.1
(d)	MM-3DScene (MM only)	Informative-preserved	×	70.9	76.9
(e)	MM-3DScene (MM only)	Informative-preserved	✓	71.1	77.2
(f)	MM-3DScene	Informative-preserved	✓	71.9	78.0

Table 8. Ablation study on mask strategies on S3DIS semantic segmentation.

5. Conclusion

We have presented MM-3DScene, a customized masked modeling framework for 3D scene understanding. It explicitly preserves the representative structured points, which provides highly useful information clues to simplify the pretext task of masked reconstruction. At each learning step, a masked scene is reconstructed in a progressive manner, so that to focus on restoring regional geometry and enjoy less ambiguity. Moreover, a self-distillation branch is integrated for maintaining the intrinsic spatial consistency on unmasked areas under the progressive masking ratios. Extensive experiments on various downstream tasks verified that our MM-3DScene significantly boosts the performance of baseline models trained from scratch.

Limitations. In this paper, we mainly focus on indoor scenes, following recent self-supervised methods [7, 24] for 3D scene understanding. We believe that the generic design insight of our masked modeling may inspire more researchers to solve 3D outdoor perception, 3D shape understanding, and 2D image recognition.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (NO. 2022ZD0160505), and in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (No. 2020355). It was also partially supported by NSFC62172348, Outstanding Youth Fund of Guangdong Province with No. 2023B1515020055 and Shenzhen General Project with No. JCYJ20220530143604010, the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Projects No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), and by Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055). Besides, thanks to Ji Hou for helpful suggestions on experiments.

References

[1] I. Armeni, O. Sener, A. R. Zamir, H. Jiang, I. Brilakis, M. Fischer, and S. Savarese. 3d semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 2, 6, 7

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEit: BERT pre-training of image transformers. In *ICLR*, 2022. 3

[3] et.al Brown, Tom. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *NeurIPS*, 2020. 3

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017. 2

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021. 3

[6] Xinlei Chen*, Saining Xie*, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, 2021. 3

[7] Yujin Chen, Matthias Nießner, and Angela Dai. 4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding. In *ECCV*, 2022. 3, 5, 6, 7, 8

[8] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 6, 7, 8

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 2, 5

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, 2021. 3

[12] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *IJCV*, 2021. 5

[13] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 2

[14] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *NeurIPS*, 2020. 3

[15] Ankur Handa, Viorica Patraucean, Vijay Badrinarayanan, Simon Stent, and Roberto Cipolla. Scenenet: Understanding real world indoor scenes with synthetic data. In *CVPR*, 2016. 2

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2, 3

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 2, 3

[18] Tong He, Dong Gong, Zhi Tian, and Chunhua Shen. Learning and memorizing representative prototypes for 3d point cloud semantic and instance segmentation. In *ECCV*, 2020. 2

[19] Tong He, Chunhua Shen, and Anton van den Hengel. DyCo3d: Robust instance segmentation of 3d point clouds through dynamic convolution. In *CVPR*, 2021. 2

[20] Tong He, Wei Yin, Chunhua Shen, and Anton van den Hengel. Pointinst3d: Segmenting 3d instances by points. In *ECCV*, 2022. 2

[21] Georg Hess, Johan Jaxing, Elias Svensson, David Hagerman, Christoffer Petersson, and Lennart Svensson. Masked autoencoders for self-supervised learning on automotive point clouds. *arXiv preprint arXiv:2207.00531*, 2022. 3

[22] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Netw.*, 1991. 5

[23] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *CVPR*, 2019. 6

[24] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *CVPR*, 2021. 3, 5, 6, 7, 8

[25] Zhicheng Huang, Xiaojie Jin, Chengze Lu, Qibin Hou, Ming-Ming Cheng, Dongmei Fu, Xiaohui Shen, and Jiashi Feng. Contrastive masked autoencoders are stronger vision learners. *arXiv preprint arXiv:2207.13532*, 2022. 7

[26] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. *CVPR*, 2020. 1

[27] Li Jing, Jiachen Zhu, and Yann LeCun. Masked siamese convnets. *arXiv preprint arXiv:2206.07700*, 2022. 7

[28] Jean Lahoud and Bernard Ghanem. 2d-driven 3d object detection in rgb-d images. In *ICCV*, 2017. 6

[29] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *NeurIPS*, 2018. 2, 6

[30] Yiqun Lin, Zizheng Yan, Haibin Huang, Dong Du, Ligang Liu, Shuguang Cui, and Xiaoguang Han. Fpconv: Learning local flattening for point convolution. In *CVPR*, 2020. 6

[31] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *ECCV*, 2022. 2

[32] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *CVPR*, 2019. 2

[33] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *ICCV*, 2021. 2

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. 6

[35] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual MLP framework. In *ICLR*, 2022. 2

[36] Chen Min, Xinli Xu, Dawei Zhao, Liang Xiao, Yiming Nie, and Bin Dai. Voxel-mae: Masked autoencoders for pre-training large-scale point clouds. *arXiv preprint arXiv:2206.09900*, 2022. 3

[37] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. In *ICCV*, 2021. 2

[38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[39] Yatian Pang, Wenxiao Wang, Francis E. H. Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *ECCV*, 2022. 2, 3, 6, 7, 8

[40] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, 2019. 1, 2, 6

[41] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*, 2018. 6

[42] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 2, 6

[43] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 2, 6

[44] Yongming Rao, Benlin Liu, Yi Wei, Jiwen Lu, Cho-Jui Hsieh, and Jie Zhou. Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection. In *ICCV*, 2021. 6, 7

[45] Zhile Ren and Erik B. Sudderth. Three-dimensional object detection and layout prediction using clouds of oriented gradients. In *CVPR*, 2016. 6

[46] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 5

[47] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *CVPR*, 2016. 6

[48] Lyne P Tchapmi, Vineet Kosaraju, Hamid Rezatofighi, Ian Reid, and Silvio Savarese. Topnet: Structural point cloud decoder. In *CVPR*, 2019. 5

[49] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 2019. 2, 6

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[51] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matthew J. Kusner. Unsupervised point cloud pre-training via occlusion completion. In *ICCV*, 2021. 2, 3

[52] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.*, 2019. 2

[53] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *CVPR*, 2019. 2, 6

[54] Peng Xiang, Xin Wen, Yu-Shen Liu, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Zhizhong Han. Snowflakenet: Point cloud completion by snowflake point deconvolution with skip-transformer. In *ICCV*, 2021. 5

[55] Tiange Xiang, Chaoyi Zhang, Yang Song, Jianhui Yu, and Weidong Cai. Walk in the cloud: Learning curves for point clouds shape analysis. In *ICCV*, 2021. 2

[56] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*, 2020. 3, 5, 6, 7, 8

[57] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *CVPR*, 2022. 2, 3

[58] Mutian Xu, Pei Chen, Haolin Liu, and Xiaoguang Han. To-scene: A large-scale dataset for understanding 3d tabletop scenes. In *ECCV*, 2022. 2

[59] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *CVPR*, 2021. 2, 6

[60] Mutian Xu, Junhao Zhang, Zhipeng Zhou, Mingye Xu, Xiaojuan Qi, and Yu Qiao. Learning geometry-disentangled representation for complementary understanding of 3d object point cloud. In *AAAI*, 2021. 2

[61] Mingye Xu, Zhipeng Zhou, and Yu Qiao. Geometry sharing network for 3d point cloud classification and segmentation. In *AAAI*, 2020. 2

[62] Mingye Xu, Zhipeng Zhou, Junhao Zhang, and Yu Qiao. Investigate indistinguishable points in semantic segmentation of 3d point cloud. In *AAAI*, 2021. 2

[63] Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *ECCV*, 2018. 2

[64] Xu Yan, Chaoda Zheng, Zhen Li, Sheng Wang, and Shuguang Cui. Pointasnl: Robust point clouds processing using nonlocal neural networks with adaptive sampling. In *CVPR*, 2020. 6

[65] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *CVPR*, 2018. 5

[66] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *CVPR*, 2019. 6

[67] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, 2022. 2, 3

[68] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *CVPR*, 2019. 5

[69] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: Multi-scale masked autoencoders for hierarchical point cloud pre-training. In *NeurIPS*, 2022. 2

[70] Yabin Zhang, Jiehong Lin, Chenhang He, Yongwei Chen, Kui Jia, and Lei Zhang. Masked surfel prediction for self-supervised point cloud learning. *arXiv preprint arXiv:2207.03111*, 2022. 2

[71] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *ECCV*, 2020. 7

[72] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. PointWeb: Enhancing local neighborhood features for point cloud processing. In *CVPR*, 2019. 2, 6

[73] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. In *ICCV*, 2021. 1, 2, 6, 7, 8