# Mitigating Object Dependencies: Improving Point Cloud Self-Supervised Learning through Object Exchange

Yanhao Wu [1] Tong Zhang[2]✉ Wei Ke[1] Congpei Qiu[1] Sabine Süsstrunk[2] Mathieu Salzmann[2]
[1] School of Software Engineering, Xi'an Jiaotong University, China
[2] School of Computer and Communication Sciences, EPFL Switzerland

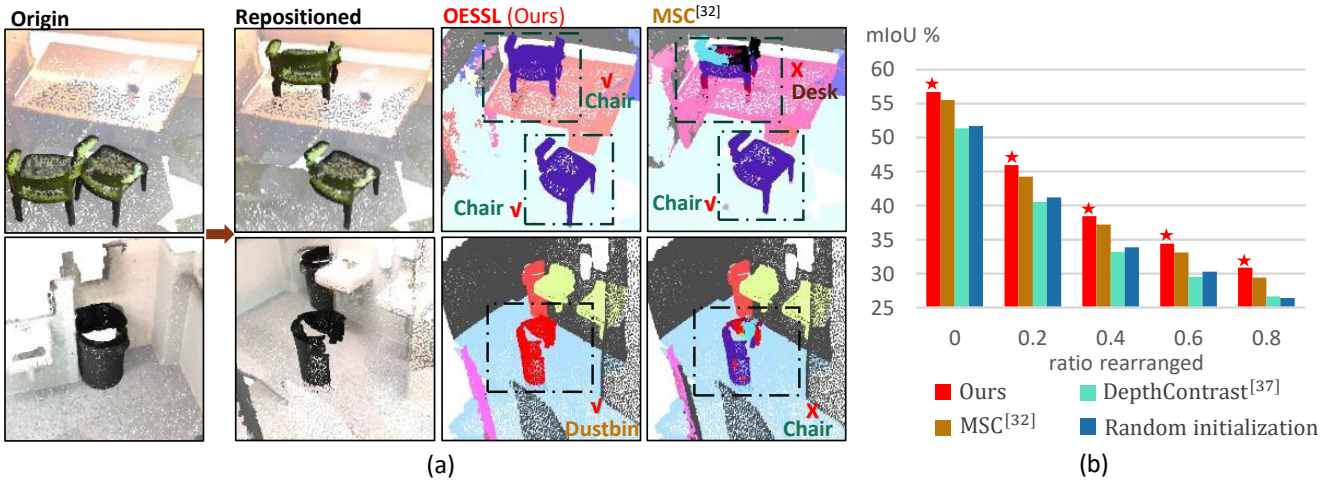


Figure 1. (a) Visualization of semantic segmentation for edited scenes. We relocate objects to places where they appear less frequently. Our pre-trained model segments the relocated object accurately, while the pre-trained model from MSC [32] labels the objects incorrectly. (b) Bar chart depicting the semantic segmentation performance on ScanNet [9] with varying ratios of rearranged objects. The X-axis indicates the ratios of rearranged objects for each scene, and the Y-axis shows the mean Intersection over Union (mIoU) scores. The models are pre-trained and fine-tuned on ScanNet with 10% labels. We compare OESSL (Ours) with MSC [32], DepthContrast [37], and training from scratch (weights are randomly initialized).

## Abstract

*In the realm of point cloud scene understanding, particularly in indoor scenes, objects are arranged following human habits, resulting in objects of certain semantics being closely positioned and displaying notable inter-object correlations. This can create a tendency for neural networks to exploit these strong dependencies, bypassing the individual object patterns. To address this challenge, we introduce a novel self-supervised learning (SSL) strategy. Our approach leverages both object patterns and contextual cues to produce robust features. It begins with the formulation of an object-exchanging strategy, where pairs of objects with comparable sizes are exchanged across different scenes, effectively disentangling the strong contextual dependencies. Subsequently, we introduce a context-aware feature learning strategy, which encodes object patterns without relying on their specific context by aggregating object features across various scenes. Our extensive experiments demonstrate the superiority of our method over existing SSL techniques, further showing its better robustness to environmental changes. Moreover, we showcase the applicability of our approach by transferring pre-trained models to diverse point cloud datasets.* [1]

## 1. Introduction

Understanding the semantic content of 3D point cloud data, particularly indoor scenes, is crucial in diverse fields, including applications such as indoor robotics [3, 5, 29, 35]. Recent advancements in deep learning [8, 31] have showcased remarkable results in this domain. While effective, these methods rely heavily on annotated training data and fail when faced with distribution shifts in the test data [38]. Consequently, the extraction of resilient object features from unlabeled data has become critical to advance the field.

Existing self-supervised learning (SSL) methods [2, 19, 20, 25, 33] concentrate on feature aggregation by creating positive pairs from the same object in different augmented

[1] Our code is available at https://github.com/YanhaoWu/OESSL.
✉: Corresponding author

views of the scene. This maintains the relative relationships between objects unchanged, thus failing to account for the object dependencies. Notably, in indoor point cloud scenes, object correlations are influenced by human habits, such as the association of tables with chairs, or toilets with sinks, resulting in strong inter-object entanglements. As demonstrated in Figure 1(a), the pre-trained models like [32] struggle to segment objects with unconventional correlations, such as chairs on desks or dustbins located away from walls. Although Mix3D [21] has been proposed to augment the data by randomly combining two scenes, it does not reason at the level of objects. Thus, the overlaps between objects introduced by this method can disrupt the coherent patterns formed by these objects. Without ground-truth labels, this disruption leads to less meaningful features, limiting the suitability of this approach in an SSL setting.

In this paper, our main focus is on developing an effective method for augmenting scene point clouds at the object level to mitigate the impact of human-induced biases in the context of self-supervised learning. Simultaneously, we aim to extract features that are more robust to varied inter-object correlations by better encoding both object patterns and contextual information. To this end, we introduce (i) an **Object Exchange Strategy**: This approach involves exchanging the positions of objects of comparable size in different scenes. By doing so, we effectively break the strong correlations between objects while alleviating issues related to object overlap. (ii) A **Context-Aware Object Feature Learning Strategy**: We first take the remaining objects, which share similar context in two randomly augmented views, as positive samples to encode the necessary contextual information and object patterns. To counter strong inter-object correlations, we minimize the feature distance between the exchanged objects in distinct contextual settings. Note that the contextual cues for a single object can vary significantly across scenes. Therefore, minimizing the feature distance between the exchanged objects enables the model to solely focus on out-of-context object patterns. These two components collectively provide a practical framework for learning robust features that encapsulate both object patterns and contextual information.

Furthermore, the exchanged objects may violate conventional human placement rules and appear incompatible with their environmental context. To effectively recognize such relocated objects, the model needs to comprehend both object patterns and context information. We therefore introduce an auxiliary task to enhance features related to both object and context. This task involves predicting which points belong to the objects that have been relocated. By engaging in this task, the model gains a more comprehensive understanding of both object patterns and contextual information.

Our contributions can be summarized as follows:

- We introduce a novel point cloud **O**bject **E**xchange **S**elf-**S**upervised **L**earning framework, named OESSL, for indoor point clouds that learn object-level feature representations by encapsulating both object patterns and contextual information.
- We propose a novel object-exchanging strategy that breaks the strong correlations between objects without incurring object overlap.
- We introduce an auxiliary task aimed at regularizing each object point feature to make it context-aware.

Our experiments on several datasets, including ScanNet [9], S3DIS [4], and Synthia4D [24], demonstrate the effectiveness of our method, especially in terms of robustness to the contextual noise, as shown in Fig. 1 (b).

## 2. Related Work

**Training with context data augmentation.** For image data, some researchers propose to add new instances to scenes to generate diverse training samples [11, 12, 14, 30, 36]. Conversely, [6, 10, 23, 26, 39] suggest removing contextual cues as data augmentation can also improve the model performance. However, the techniques designed for images cannot be directly applied to point clouds due to their distinct data nature.

In the 3D domain, 4dcontrast [7] augments scenes with moving synthetic objects and encourages feature similarity between corresponding objects. However, 4dcontrast needs synthetic datasets to obtain shapes, and moving a single object introduces limited contextual diversity. Nekrasov *et al.* [21] propose a data augmentation named Mix3D which involves directly combining two point clouds and training models using augmented scenes in a supervised manner. The merged scene becomes chaotic with occlusions and overlaps, hindering the extraction of object-level features in the self-supervised learning (SSL) setting. Additionally, this exchange lacks meaningful object interactions and disrupts contextual information. By contrast, our object exchange strategy integrates objects from different scenes, greatly increasing the diversity of contextual cues while alleviating object overlap.

**Self-supervised learning for 3D point clouds.** Self-supervised learning for point clouds has developed rapidly in recent years [1, 2, 16, 19, 25]. In indoor scenes, recent research [22, 32, 34] explores the nature of 3D point cloud data by aggregating features within the same point/object. For example, Pointcontrast [34] and MSC [32] aggregate spatial features by maximizing the similarity between corresponding point features; DepthContrast [37] and STRL [17] aggregate features in each region and pull features from different views together. Although effective, the correlation between indoor objects is strongly influenced by human bias, resulting in strong entanglements between objects. Therefore, aggregating features from indoor objects

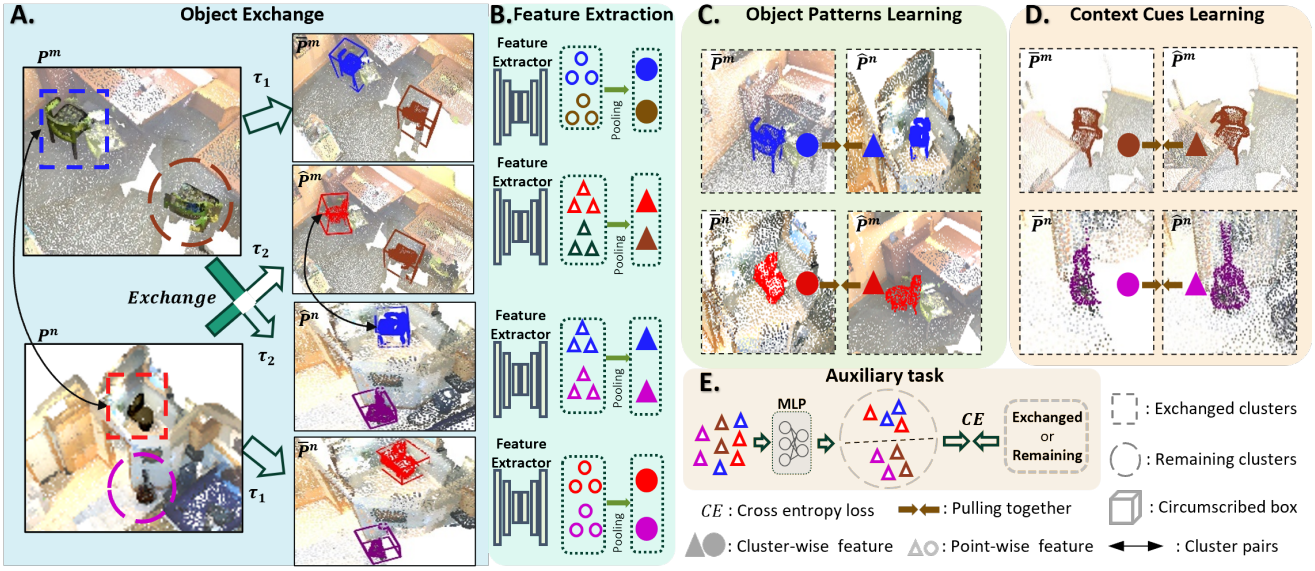


Figure 2. **Overview of our OESSL**. **A**. Given two randomly selected point clouds $P^m$ and $P^n$, we first perform clustering and generate minimum circumscribed boxes for every cluster. Clusters with similar circumscribed boxes are matched as cluster pairs. We exchange points of matched clusters and apply augmentation on $P^m$ and $P^n$ to generate novel views $\hat{P}^m$, $\hat{P}^n$, alongside two augmented views $\bar{P}^m$ and $\bar{P}^n$ without exchange. **B.** Every scene is passed through a feature extractor (Backbone) to obtain point-wise and cluster-wise features. **C.** We minimize the cluster feature distance obtained from the exchanged clusters in the different scenes (i.e.,$\bar{P}^m$ and $\hat{P}^n$,$\bar{P}^n$ and $\hat{P}^m$). **D.** We maximize the feature similarity between the remaining clusters in the augmented scenes (i.e.,$\bar{P}^m$ and $\hat{P}^m$, $\bar{P}^n$ and $\hat{P}^n$). **E.** The point-wise features are passed through a multilayer perceptron (MLP) to classify which points belong to the relocated objects. The cross-entropy loss is used for classification. $\tau_1$ and $\tau_2$ are data augmentations, such as random flipping and random clipping.

may lead to the model overfitting to inter-object correlations and ignoring object patterns.

By contrast, our method disrupts the correlations between objects to mitigate the model's dependence on contextual information. Additionally, we introduce a context-aware object feature learning strategy that leverages both object patterns and contextual information.

## 3. Method

The overall framework of our method is depicted in Fig. 2 and contains two parts: Object exchange and context-aware object feature learning. We discuss these components in detail below.

### 3.1. Object exchange

**Unsupervised clustering.** Let us be given a series of point clouds $P = \{P^1, P^2, ..., P^T\}$ depicting $T$ scenes, where $P^k = (X^k, C^k) = \{(x_1^k, c_1^k), (x_2^k, c_2^k)..., (x_{N_k}^k, c_{N_k}^k)\}$ represents the $k$-th point cloud with $N_k$ 3D points $x_i^k \in \mathbb{R}^3$ and corresponding RGB colors $c_i^k \in \mathbb{R}^3$. For each 3D point set $X^k$, we compute normals for each point following [9]. This process yields a set of $N_k$ point noramls $O^k = \{O_1^k, O_2^k, ..., O_{N_k}^k\}$, $O_i^k \in \mathbb{R}^3$. Then, these points are taken as vertices to construct a graph whose weight matrix is defined as:

$$D = 2 - (D_{nor} + \alpha * D_{feat}) , \quad (1)$$

where $D_{nor}$ represents the matrix of pairwise cosine similarity between the normals of two points, while $D_{feat}$ represents the matrix of pairwise cosine similarity based on point features. The parameter $\alpha \in [0, 1]$ serves as a weight, balancing the influence of the two matrices. We initialize $\alpha$ at 0 and iteratively update it during the feature learning process in Sec. 3.2.1. Note that when the positions of two points, $i$ and $j$, are not spatially adjacent, $D_{ij}$ is set to a large number. Subsequently, we employ the GraphCut [13] algorithm, a graph-based segmentation method, to cluster the points into $M_k$ clusters [9]. The center of each cluster is determined as the average of all points belonging to that cluster.

**Exchanging objects with comparable size.** To ensure meaningful object exchange without causing overlap with nearby objects, we adopt a systematic approach. We first apply [27] to all the clusters to generate $M_k$ minimum circumscribed boxes, denoted as $B^k = \{B_1^k, B_2^k, ..., B_{M_k}^k\}$, where $B_i^k$ represents the length, width, and height of the $i$-th box in scene $k$. The pairwise box similarity is defined as the Euclidean distance between the vectors composed of length, width, and height, such that smaller distances correspond to higher similarity. To enhance the diversity of exchanged objects, we employ a hybrid sampling strategy. For the $\beta M_k$ clusters in scene $k$, where $\beta$ is the preset exchange proportion of the clusters, we first select $\frac{\beta}{2} M_k$ clusters using the farthest point sampling (FPS) algorithm, ensuring a

representative spatial distribution. The remaining clusters are then chosen via random sampling, introducing an element of randomness in the selection process.

Next, we introduce a similarity degree matrix, $V \in \mathbb{R}^{\beta M_k \times M_h}$, where $V_{i,j}$ indicates the pairwise box similarity between cluster $i$ in scene $k$ and cluster $j$ in scene $h$. Following a greedy strategy, we match box pairs with the highest similarity in $V$. Subsequently, the points belonging to the corresponding matched clusters are exchanged between the two scenes. Leveraging $V$ helps to avoid object overlap, emphasizing the variability in contextual cues for a single object across different scenes. Further insights into the generation of robust features by exploiting such objects are discussed in Sec. 3.2.2.

## 3.2. Context-aware Object Feature Learning

Having defined our object exchange strategy, we provide more detail on how to extract the features and establish our feature learning framework.

### 3.2.1 Feature extraction

Given an input point cloud $P^n$ and a randomly selected point cloud $P^m$ from the dataset, we apply our object exchange strategy and data augmentation to create two novel views $\hat{P}^m$ and $\hat{P}^n$, alongside two augmented views $\bar{P}^m$ and $\bar{P}^n$ without exchanging. To capture both point-wise and cluster-wise information, we leverage MinkUnet [8] as our backbone encoder, denoted as $\phi$.

We initiate the feature extraction process by forwarding $\hat{P}^m$ through the backbone encoder, obtaining point-wise features $\hat{f}_i^m = \phi(\hat{P}^m)$ for each 3D point. Organizing these features according to clusters results in a set of point-wise features, $\hat{F}^m = \{\hat{F}_1^m, \hat{F}_2^m, ..., \hat{F}_{M_m}^m\}$, where $\hat{F}_i^m \in \mathbb{R}^{N_{m,i} \times d}$, with $N_{m,i}$ representing the number of points in cluster $i$ from point cloud $\hat{P}^m$, and $d$ is the feature dimension of each $\hat{f}_i^m$. Additionally, we employ max-pooling on point features based on the clusters obtained using GraphCut, generating cluster-wise features $\hat{C}^m = \{\hat{c}_1^m, \hat{c}_2^m, \cdots, \hat{c}_{M_m}^m\}$, where $\hat{c}_i^m \in \mathbb{R}^{1 \times d}$. The features in the other scenes can be obtained in the same way, as shown in Fig. 2.

### 3.2.2 Feature aggregation

Aiming at a balanced concurrent ratio among objects of different semantics, we operationalize our approach through two central strategies for aligning cluster features: Object Patterns Learning and Contextual Cues Learning, both detailed below. Furthermore, we introduce an auxiliary task dedicated to enhancing the encoder's awareness of whether an object's feature distribution is in an unconventional location. This design aims to mitigate the challenges associated with cluster-level feature alignment by having regularization on point-level distribution.

**Object patterns learning.** To encourage the model to learn object patterns, we minimize the feature distance between the clusters/points in the same cluster in different scenes. Note that the contextual cues for a single object can vary significantly between different scenes. Minimizing the feature distance between exchanged objects enables the model to solely focus on object patterns.

Let $M_m^{ex}$ denote the number of exchanged clusters in $\hat{P}^n$ that are originally located in $P^m$. We define a loss function

$$L_{op}^m = \frac{1}{M_m^{ex}} \times \sum_{i=1}^{M_m^{ex}} \Big( \Big\| \frac{\hat{c}_i^n}{\|\hat{c}_i^n\|_2} - \frac{\bar{c}_i^m}{\|\bar{c}_i^m\|_2} \Big\|_2^2 + \frac{1}{N_{m,i}} \times \sum_{j=1}^{N_{m,i}} \Big\| \frac{\hat{f}_{i,j}^n}{\|\hat{f}_{i,j}^n\|_2} - \frac{\bar{c}_i^m}{\|\bar{c}_i^m\|_2} \Big\|_2^2 \Big), \tag{2}$$

where $\bar{c}_i^m$ and $\hat{c}_i^n$ are the cluster-level feature vectors of the same exchanged clusters in $\bar{P}^m$ and $\hat{P}^n$, and $\hat{f}_{i,j}^n$ represents the features of point $j$ belonging cluster $i$ in $\hat{P}^n$. The loss function $L_{op}^n$ for the point cloud $P^n$ can be obtained in the same way. We then employ the symmetrized loss

$$L_{op} = L_{op}^m + L_{op}^n. \tag{3}$$

**Contextual cues learning.** To learn contextual cues, we minimize the feature distance between the remaining clusters, which share similar contexts in two randomly augmented views. To constrain the feature of each point, we also minimize the distance between the point and the corresponding cluster features [33].

Let $M_m^{re}$ denote the number of remaining clusters that have not been exchanged in $P^m$. We write a loss

$$L_{context}^m = \frac{1}{M_m^{re}} \times \sum_{i=1}^{M_m^{re}} \Big( \Big\| \frac{\hat{c}_i^m}{\|\hat{c}_i^m\|_2} - \frac{\bar{c}_i^m}{\|\bar{c}_i^m\|_2} \Big\|_2^2 + \frac{1}{N_{m,i}} \times \sum_{j=1}^{N_{m,i}} \Big\| \frac{\hat{f}_{i,j}^m}{\|\hat{f}_{i,j}^m\|_2} - \frac{\bar{c}_i^m}{\|\bar{c}_i^m\|_2} \Big\|_2^2 \Big), \tag{4}$$

where $\bar{c}_i^m$, $\hat{c}_i^m$ are the cluster feature vectors of the same remaining cluster in $\bar{P}^m$ and $\hat{P}^m$, and $\hat{f}_{i,j}^m$ represents the feature of point $j$ belonging cluster $i$ in $\hat{P}^m$. The loss function $L_{context}^n$ for the point cloud $P^n$ can be obtained in the same way. We then define the symmetrized loss

$$L_{context} = L_{context}^m + L_{context}^n. \tag{5}$$

**Auxiliary task.** The auxiliary task aims to enable the model to gain a more comprehensive understanding of both object patterns and contextual information. For the point cloud $\hat{P}^m$, we define a vector $\hat{Y}^m = \left\{\hat{y}_1^m, \hat{y}_2^m, ..., \hat{y}_{\hat{N}_m}^m\right\}$, where $\hat{y}_i^m \in [0, 1]$ represents whether point $i$ belongs to an exchanged cluster and $\hat{N}_m$ represents the number of

points in $\hat{P}^m$. We forward the point features $\hat{F}^m$ to a multilayer perceptrons (MLP) to obtain point-wise prediction $\hat{Z}^m = \left\{\hat{z}_1^m, \hat{z}_2^m, ..., \hat{z}_{\hat{N}_m}^m\right\}$, where $\hat{z}_i^m \in \{0, 1\}$. For the point cloud $\hat{P}^n$, we obtain $\hat{Y}^n$ and $\hat{Z}^n$ in a same way. We then define a loss $L_{aux}$ encoding the standard cross entropy loss between $\hat{Y}^m$ and $\hat{Z}^m$, and $\hat{Y}^n$ and $\hat{Z}^n$.

Hence, our complete loss is written as

$$L_{total} = L_{context} + \lambda L_{op} + \gamma L_{aux}, \tag{6}$$

where $\lambda$ and $\gamma$ are weights balancing the three loss terms. We set $\lambda$ to 1 and $\gamma$ to 2 in our experiments.

# 4. Experiments

In this section, we first introduce our experimental settings, including the datasets, object exchange details, and implementation details. Then, we evaluate our pre-trained models on downstream tasks and analyze our framework.

## 4.1. Experimental Settings

**Datasets**. ScanNet [9] consists of 3D reconstructions of real rooms and comprises 1513 indoor scenes. We follow the setting in [8] and use a training and validation set, including 1201 and 312 scenes, respectively. The training set is used for pre-training and fine-tuning. Our framework utilizes scene-level point clouds for pre-training. The Standford Large-Scale 3D Indoor Space (S3DIS) [4] dataset contains 6 large-scale indoor areas [8]. We use area5 as validation data and the remaining areas as training data. Synthia4D is a large dataset that contains 3D scans of 6 sequences of driving scenes. Following [8], we split the Synthia4D dataset into train/val/test sets including 19888/815/1886 scenes.

**Object exchange details**. To obtain better segmentations for object exchange and feature extraction, we update the point features with our learned features to create the affinity matrix. We set the initial relative weight $\alpha$ to 0 in Eq. (1) and update the clusters twice during the training process: first at one third and then at two thirds, by setting $\alpha$ to 0.5. We set the similarity threshold in GraphCut [13] to 1.5 and merge the clusters with fewer than 300 points. In each scene, the clusters with any side length of the corresponding box exceeding 3 meters or less than 0.2 meters are not used for exchange.

**Implementation details**. We use MinkUnet [8] as the backbone feature extractor and build our framework on the basis of BYOL [15]. DepthContrast [37], MSC [32], STRL [17], and training from scratch are reproduced with the same backbone as ours to have fair comparisons. We pre-train the backbone on ScanNet for 200 epochs. The learning rate is initially set to 0.036 with a cosine annealing scheme with a minimum learning rate equal to $0.036 \times 10^{-4}$. We use SGD with a momentum of 0.9 and a weight decay of 0.0004 following STSSL [33]. We use 8 $\times$ GTX3090 GPUs for pre-training and the batch size for each GPU is 12, which leads to a total batch size of 96.

**Evaluation metrics**. We use the mean intersection over union (mIoU) and the overall point classification accuracy (Acc) to evaluate point cloud semantic segmentation, and average precision (mAP, AP@50%, AP@25%) for instance segmentation.

## 4.2. Scene Understanding

To evaluate the pre-training methods, we employ different numbers of labels to fine-tune the models. In line with previous methods [33, 37], we partition ScanNet, S3DIS, and Synthia4D into distinct regimes, each corresponding to different percentages of labeled data. Specifically, we downsample the training data to levels of 10%, 20%, 50%, and 100% for ScanNet and S3DIS, and 0.1%, 1%, 10%, and 100% for Synthia4D. To mitigate randomness, we downsample three different regimes for every percentage, fine-tune the models separately using each regime, and report the average performance. The number of training epochs for every label regime can be found in the supplementary.

**Indoor scene understanding**. To evaluate the improvement of our OESSL on indoor scene understanding, we fine-tune the pre-trained model on ScanNet.

| | 10% | 20% | 50% | 100% |
|---|---|---|---|---|
| From Scratch | 48.99 | 57.58 | 61.70 | 71.11 |
| DepthContrast [37] | 50.30 | 57.08 | 61.47 | 70.92 |
| STRL [17] | 46.94 | 58.94 | 61.85 | 71.03 |
| MSC [32] | 53.85 | 60.47 | 63.98 | 71.00 |
| **OESSL (ours)** | **54.37** | **61.27** | **64.56** | **71.28** |

Table 1. Pre-training on **ScanNet** and evaluating the fine-tuned models in different label regimes on **ScanNet** for semantic segmentation. We report the mIoU.

In Table 1, we show the semantic segmentation results obtained by fine-tuning with different percentages of training data. Our method achieves better mIou for all label regimes than MSC. Specifically, our method outperforms training from scratch by 5.38% at a level of 10% and MSC [32] by 0.8% at a level of 20%. In Table 2, we report the instance segmentation results driven by *PointGroup* [18]. When using 10% of the labels for fine-tuning, our method improves performance by 4.7% in AP@50% compared to the network without pre-training. This evidences that our pre-training framework is also beneficial for discriminating instances.

**Indoor scene transferability**. The contextual information significantly differs across datasets, making it difficult to transfer contextual features between different datasets, especially between indoor and outdoor scenes. By contrast, the object patterns, such as color and shape, are commonly shared between objects. Our method generates more trans-

| mAP / AP@50 / AP@25 | 10% | 20% | 50% |
|---|---|---|---|
| From Scratch | 12.63 / 25.90 / 43.33 | 23.63 / 41.42 / 60.73 | 30.91 / 51.25 / 68.38 |
| MSC [32] | 13.42 / 27.30 / 44.82 | 23.90 / 42.28 / **61.48** | 29.16 / 51.18 / 68.71 |
| **OESSL (ours)** | **15.30 / 30.60 / 49.94** | **24.67 / 43.28** / 60.86 | **31.73 / 52.06 / 69.80** |

Table 2. Pre-training on **ScanNet** and evaluating the fine-tuned models in different label regimes on **ScanNet** for instance segmentation [18]. We report the mAP, AP@50, AP@25.

| | 10% | 20% | 50% | 100% |
|---|---|---|---|---|
| From Scratch | 40.48 | 45.94 | 53.25 | 66.16 |
| DepthContrast[37] | 46.57 | 47.67 | 53.85 | 63.42 |
| STRL [17] | 36.99 | 46.13 | 55.11 | 64.71 |
| MSC [32] | 44.85 | 50.12 | 57.16 | 65.40 |
| **OESSL (ours)** | **49.22** | **52.67** | **61.79** | **66.90** |

Table 3. Pre-training on **ScanNet** and evaluating the fine-tuned models in different label regimes on **S3DIS** for semantic segmentation. We report the mIoU.

ferable features by encoding object patterns without relying on their specific context.

To demonstrate the transferability of the features learned via our method, we pre-train models on ScanNet and fine-tune them for semantic segmentation on S3DIS [4]. As shown in Table 3, our pre-trained model performs better than the other methods. Specifically, our method outperforms MSC [32] by 4.37% in mIoU with 10% of the labels. These results strongly confirm the effectiveness of our approach at extracting object features that remain robust to changes in the environment.

**Outdoor scene transferability**. We further fine-tune the models pre-trained on ScanNet for semantic segmentation using Synthia4D [24], a self-driving dataset with different contexts than indoor scenes. In Table 4 and Table 5, we report the mIoU obtained by fine-tuning the models using Synthia4D. Our method outperforms the other methods consistently across all label regimes. Specifically, our OESSL outperforms MSC [32] by 2.33% with 1% of the labels in the test set. When utilizing only 0.1% of the training data, all pre-trained models exhibit a substantial improvement compared to training from scratch. Notably, our method achieves the most significant improvement, resulting in an mIoU of 49.32% when evaluated on the validation set. The improvements on S3DIS [4] and Synthia4D [24] show that the features learned by our method generalize better than those learned by other methods.

| | 0.1% | 1% | 10% | 100% |
|---|---|---|---|---|
| From Scratch | 19.84 | 63.37 | 70.45 | 77.00 |
| DepthContrast [37] | 46.11 | 66.25 | 70.49 | 75.21 |
| STRL [17] | 39.64 | 65.59 | 69.45 | 77.33 |
| MSC [32] | 47.11 | 66.42 | 73.15 | 77.25 |
| **OESSL (ours)** | **49.44** | **68.75** | **73.42** | **77.48** |

Table 4. Pre-training on **ScanNet** and evaluating the fine-tuned models on **Synthia4D** for semantic segmentation. The models are evaluated on the **test set**. We report the mIoU.

| | 0.1% | 1% | 10% | 100% |
|---|---|---|---|---|
| From Scratch | 20.17 | 67.87 | 74.35 | 80.50 |
| DepthContrast [37] | 46.23 | 71.66 | 74.00 | 78.56 |
| STRL [17] | 38.27 | 70.49 | 73.80 | 80.95 |
| MSC [32] | 46.42 | 71.58 | 75.53 | 81.05 |
| **OESSL (ours)** | **49.32** | **74.17** | **77.04** | **81.31** |

Table 5. Pre-training on **ScanNet** and evaluating the fine-tuned models under different label regimes on **Synthia4D** for semantic segmentation. The models are evaluated on the **validation set**.

### 4.3. Ablation study

In this section, we dissect our OESSL and analyze each component. Unless explicitly stated otherwise, the model is pre-trained and fine-tuned on ScanNet.

**Breaking entanglements between objects.** Due to inherent human biases, strong correlations exist among indoor objects, indicating that certain classes of objects are highly likely to co-occur. This co-occurrence introduces the risk of the model overfitting to inter-object relations.

In Fig. 5, we illustrate the frequency of any two classes of objects appearing together. In the original training dataset (on the top of 5), certain classes exhibit a high frequency of appearing together. For instance, the shower curtain and door consistently appear simultaneously, and the co-occurrence frequency between the counter and cabinet is 0.9. However, by exchanging objects between scenes, our approach alleviates the high co-occurrence frequencies between objects, as shown in the bottom of Fig.5.

**Performance under varied contexts.** Our method avoids overemphasizing contextual cues and is therefore less affected by context changes compared to other SSL techniques. To validate this, we evaluate the model's performance in scenes with varied contexts. Specifically, we create a new dataset, ScanNet-C, by replacing a proportion $\delta$ of the objects in ScanNet with randomly selected objects from the entire dataset. We report the ratio of the model's performance on ScanNet-C to its performance on ScanNet. A higher ratio indicates a lower impact from contextual changes. In the experiment, we vary $\delta$ and repeat the experiment three times, reporting the average to reduce randomness. As shown in Table 6, our pre-trained model consistently achieves higher mIoU values for all $\delta$ values, confirming that our method is indeed more robust to contextual changes than other methods.

In Fig. 3, we visualize the semantic segmentation for scenes generated by relocating objects in a reasonable but

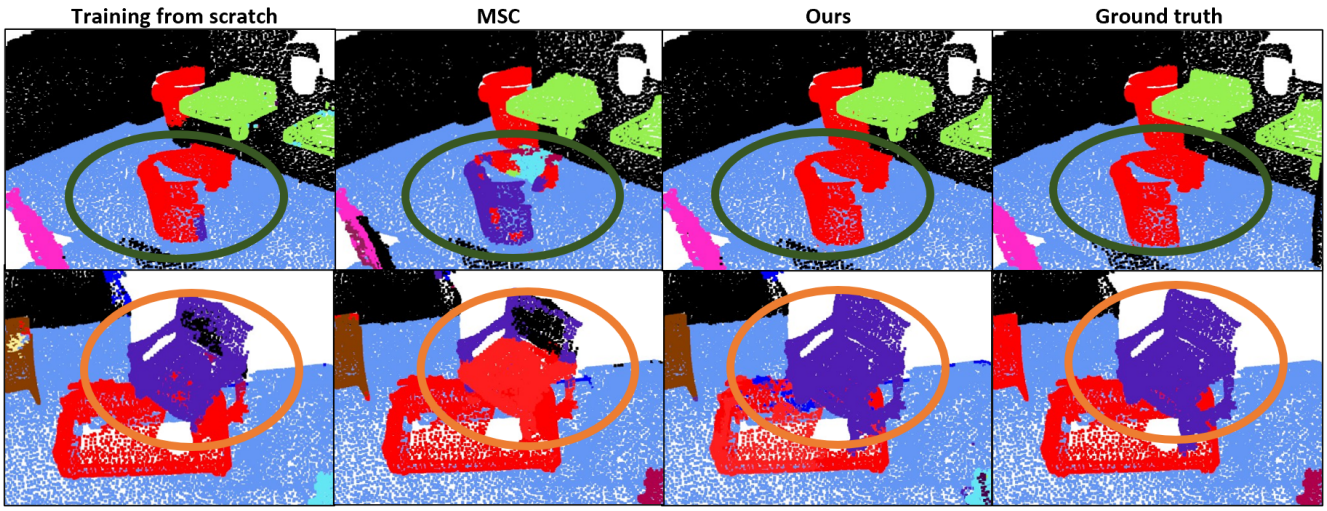


Figure 3. Segmentation results in scenes with objects relocated in unusual locations to eliminate contextual cues. We compare MSC [32], OESSL (Ours), and training from scratch (without pre-training). The model pre-trained with our method better distinguishes the relocated objects, as shown in the highlighted area (colored circles).

| Method \ $\delta$ | 0.2 | 0.4 | 0.6 | 0.8 |
|---|---|---|---|---|
| From Scratch | 79.60 | 65.50 | 58.56 | 51.01 |
| DepthContrast [37] | 78.90 | 64.55 | 57.48 | 51.88 |
| MSC [32] | 79.70 | 67.05 | 59.63 | 52.92 |
| OESSL (ours) | **80.99** | **67.75** | **60.67** | **54.43** |

Table 6. **Comparison of robustness to contextual changes.** We evaluate models on ScanNet-C with different proportions $\delta$ of replaced objects. We report the ratio(%) of the model's performance on ScanNet-C to its performance on ScanNet.

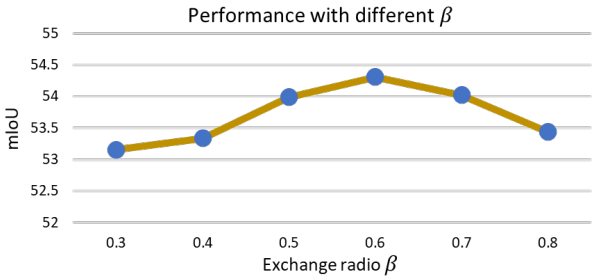


Figure 4. Comparison of mIoU on ScanNet, after fine-tuning the models pre-trained with different $\beta$.

unusual location. Specifically, a dustbin is placed far from the walls and a sofa is placed on the desk. For such objects with unreliable contextual cues, MSC [32] and the model without pre-training fail to segment the point clouds. By contrast, our OESSL accurately segments the objects, benefiting from object patterns learning. For additional visualizations and detailed information about ScanNet-C, please refer to the supplementary material.

**Effect of the exchanged object proportion.** In this study, we aim to clarify the impact of the exchange ratio on the learning process. The hyperparameter $\beta$ represents the proportion of exchanged clusters in the object exchange strategy. In our approach, when the number of available clusters in the scene exceeds 20, we set $\beta$ to 0.5; otherwise, we set it to 1. We keep $\beta$ fixed during pretraining to evaluate its impact on the model. The experiments are repeated three times to mitigate randomness. As depicted in Fig. 4, the performance initially increases and then decreases as $\beta$ increases. We hypothesize that this is because a higher $\beta$ has the potential to increase the risk of object overlap, thereby completely disrupting existing contextual information. As shown in the bottom of Fig. 6, when $\beta$ is set to 0.5, the desk is replaced by a bed, breaking the correlation between desk and sofa. However, a chair exchanges positions with the pillow on the sofa, disrupting the object patterns when $\beta$ equals 0.7. The best-performing model corresponds to setting $\beta$ to 0.6, which balances the number of exchanged objects and non-overlapping objects.

| | mIoU(%) | Acc(%) |
|---|---|---|
| From Scratch | 48.99 | 78.88 |
| MSC [32] | 53.85 | 80.49 |
| Baseline+Mix3D | 52.62 | 80.19 |
| **OESSL** | **54.37** | **81.15** |

Table 7. Ablation study on the loss function with 10% of the labels on ScanNet. We report mIoU/Acc.

**Comparison with Mix3D.** Mix3D [21] is an augmentation that directly combines two point clouds to generate novel scenes and is effective for supervised semantic segmentation training. Different from supervised training, self-supervised pre-training aims to generate structured embeddings. Specifically, the objects of the same class should be close in feature space and far from the objects from other classes. The overlap between objects incurred by Mix3D makes it difficult to distinguish the patterns between different object classes, resulting in an irregular feature space. Unlike Mix3D, our proposed object-exchanging strategy

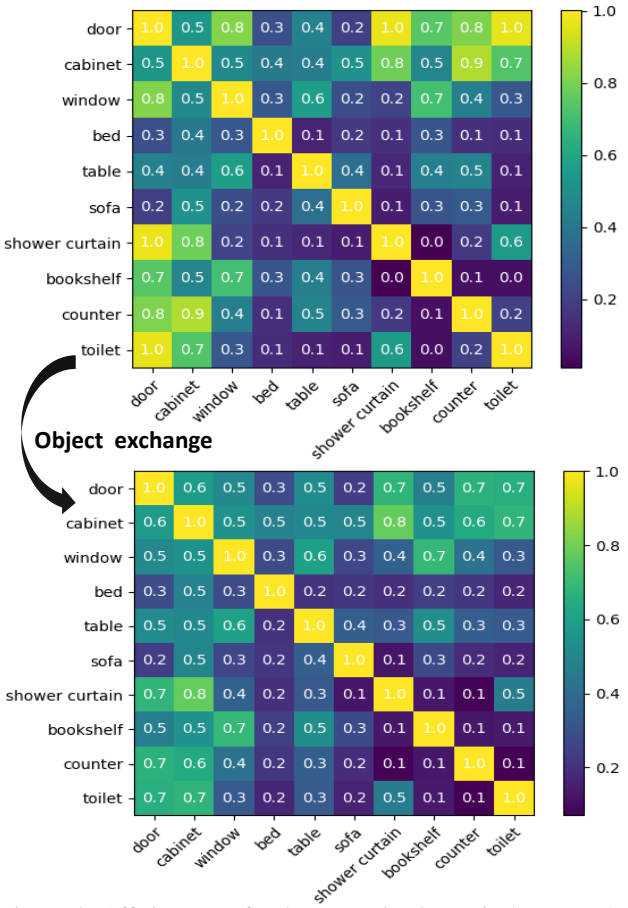


Figure 5. Affinity maps for the semantic classes in ScanNet [9]. Top: affinity map for the training set. Bottom: affinity map for the training set after object exchange.

| Method | Context | OP | Aux | mIoU |
|---|---|---|---|---|
| Baseline | ✓ | | | 53.12 |
| Baseline + $L_{OP}$ | ✓ | ✓ | | 53.90 |
| OESSL | ✓ | ✓ | ✓ | **54.37** |

Table 8. Ablation study on the **loss functions** with 10% of the labels on ScanNet. **Context**: Context cues learning, **OP**: Object pattern feature learning, **Aux**: Auxiliary task.

mitigates object overlaps, as shown in the top of Fig. 6.

To further highlight the effectiveness of our proposed object-exchanging strategy, we replace it with the Mix3D method and minimize feature distance between corresponding points/clusters in the newly generated scenes. This setting, referred to as Baseline+Mix3D in Table 7, yields an mIoU of 52.62%, lower than MSC and OESSL. It implies that Mix3D is not suitable for self-supervised learning.

**Loss functions.** We ablate the three loss functions in Eq. 6 to validate their effectiveness. Initially, we set $\beta$ to 0, ensuring that only the remaining clusters contribute, and only the loss function of Eq.5 is applied. We refer to this configuration as the baseline. Subsequently, by adjusting $\beta$, we activate the loss function in Eq.3, specifically designed for object pattern learning. This setting is denoted as Baseline+$L_{OP}$. The results in Table 8 show that Baseline+$L_{OP}$ outperforms the baseline, achieving an mIoU of 53.90%. Our OESSL extends this by incorporating an auxiliary task, resulting in a remarkable mIoU of 54.37%, demonstrating superior performance.

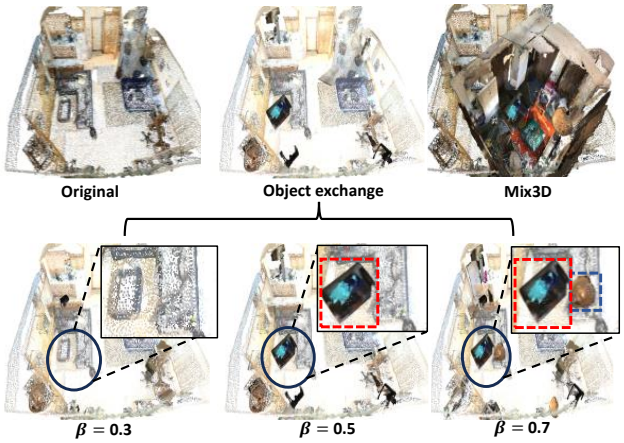


Figure 6. Top: Visual comparison of scenes generated by Mix3D and our strategy. Bottom: Scenes generated by different $\beta$ using the object exchange strategy. When $\beta$ is set to 0.5, the desk is replaced by a bed (highlighted in the red box), but a chair is exchanged with the pillow (highlighted in the blue box) when $\beta$ increases to 0.7. For better visualization, we enhance the color contrast between objects from different scenes.

**Different backbones.** We conduct experiments using SPVCNN [28] as the backbone. The results, presented in Table 9, demonstrate the effectiveness of our method with SPVCNN [28].

| Method | mIoU(%) | Acc(%) |
|---|---|---|
| From Scratch | 45.59 | 77.38 |
| Baseline | 47.38 | 78.68 |
| OESSL (ours) | **49.02** | **79.25** |

Table 9. Ablation study on backbones. The models are pre-trained on ScanNet and tested with 10% labels.

## 5. Conclusion

In this paper, we have introduced a SSL framework for point clouds, aiming to capture object features that are robust to noise and contextual variations. It starts by exchanging objects with comparable sizes between different scenes, breaking strong inter-object entanglements, and then learning both object patterns and contextual cues by leveraging exchanged and remaining objects. Altogether, our approach provides practical tools to learn robust context-aware representation features for indoor scenes. Our experiments evidence that our method outperforms the previous SSL methods for indoor point clouds.

**Acknowledgement.** This work was supported in part by the National Natural Science Foundation of China under Grant No. 62376209 and the Swiss National Science Foundation via the Sinergia grant CRSII5-180359.

## References


[1] Idan Achituve, Haggai Maron, and Gal Chechik. Self-supervised learning for domain adaptation on point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 123–133, 2021. 2

[2] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International Conference on Machine Learning*, pages 40–49. PMLR, 2018. 1, 2

[3] Ziyad Alenzi, Emad Alenzi, Mohammad Alqasir, Majed Alruwaili, Tareq Alhmiedat, and Osama Moh'd Alia. A semantic classification approach for indoor robot navigation. *Electronics*, 11(13):2063, 2022. 1

[4] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1534–1543, 2016. 2, 5, 6

[5] Hermann Blum, Francesco Milano, René Zurbrügg, Roland Siegwart, Cesar Cadena, and Abel Gawel. Self-improving semantic perception for indoor localisation. In *Conference on Robot Learning*, pages 1211–1222. PMLR, 2022. 1

[6] Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. Gridmask data augmentation. *arXiv preprint arXiv:2001.04086*, 2020. 2

[7] 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

[8] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 1, 4, 5

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 2, 3, 5, 8

[10] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2

[11] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 364–380, 2018. 2

[12] Debidatta Dwibedi, Ishan Misra, and Martial Hebert. Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1301–1310, 2017. 2

[13] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 3, 5

[14] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928, 2021. 2

[15] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020. 5

[16] 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

[17] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6535–6545, 2021. 2, 5, 6

[18] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020. 5, 6

[19] Longlong Jing, Yucheng Chen, Ling Zhang, Mingyi He, and Yingli Tian. Self-supervised modal and view invariant feature learning. *arXiv preprint arXiv:2005.14169*, 2020. 1, 2

[20] Hao Li, Dingwen Zhang, Nian Liu, Lechao Cheng, Yalun Dai, Chao Zhang, Xinggang Wang, and Junwei Han. Boosting low-data instance segmentation by unsupervised pre-training with saliency prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15485–15494, 2023. 1

[21] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3d: Out-of-context data augmentation for 3d scenes. In *2021 International Conference on 3D Vision (3DV)*, pages 116–125. IEEE, 2021. 2, 7

[22] Lucas Nunes, Rodrigo Marcuzzi, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Segcontrast: 3d point cloud feature representation learning through self-supervised segment discrimination. *IEEE Robotics Autom. Lett.*, 7(2): 2116–2123, 2022. 2

[23] Congpei Qiu, Tong Zhang, Yanhao Wu, Wei Ke, Mathieu Salzmann, and Sabine Süsstrunk. Mind your augmentation: The key to decoupling dense self-supervised learning. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[24] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016. 2, 6

[25] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[26] Krishna Kumar Singh, Hao Yu, Aron Sarmasi, Gautam Pradeep, and Yong Jae Lee. Hide-and-seek: A data aug-

mentation technique for weakly-supervised localization and beyond. *arXiv preprint arXiv:1811.02545*, 2018. 2

[27] Jack Sklansky. Finding the convex hull of a simple polygon. *Pattern Recognition Letters*, 1(2):79–83, 1982. 3

[28] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*, pages 685–702. Springer, 2020. 8

[29] Hugues Thomas, Matthieu Gallet de Saint Aurin, Jian Zhang, and Timothy D Barfoot. Learning spatiotemporal occupancy grid maps for lifelong navigation in dynamic scenes. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 484–490. IEEE, 2022. 1

[30] Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M Rehg, and Visesh Chari. Learning to generate synthetic data via compositing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 461–470, 2019. 2

[31] Peng-Shuai Wang. Octformer: Octree-based transformers for 3d point clouds. *ACM Transactions on Graphics (SIGGRAPH)*, 42(4), 2023. 1

[32] Xiaoyang Wu, Xin Wen, Xihui Liu, and Hengshuang Zhao. Masked scene contrast: A scalable framework for unsupervised 3d representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9415–9424, 2023. 1, 2, 5, 6, 7

[33] Yanhao Wu, Tong Zhang, Wei Ke, Sabine Süsstrunk, and Mathieu Salzmann. Spatiotemporal self-supervised learning for point clouds in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5251–5260, 2023. 1, 4, 5

[34] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pretraining for 3d point cloud understanding. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020. 2

[35] Wenhao Xue, Yang Yang, Lei Li, Zhongling Huang, Xinggang Wang, Junwei Han, and Dingwen Zhang. Weakly supervised point cloud segmentation via deep morphological semantic information embedding. *CAAI Transactions on Intelligence Technology*, 2023. 1

[36] Lingzhi Zhang, Tarmily Wen, Jie Min, Jiancong Wang, David Han, and Jianbo Shi. Learning object placement by inpainting for compositional data augmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 566–581. Springer, 2020. 2

[37] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 1, 2, 5, 6, 7

[38] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021. 1

[39] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020. 2