Synthetic-to-Real Domain Generalized Semantic Segmentation for 3D Indoor Point Clouds

Yuyang Zhao[1]
yuyang.zhao@u.nus.edu
Na Zhao[2]
na_zhao@sutd.edu.sg
Gim Hee Lee[1]
gimhee.lee@nus.edu.sg

[1] Department of Computer Science
National University of Singapore
Singapore, Singapore

[2] ISTD Pillar
Singapore University of Technology and Design
Singapore, Singapore

Abstract

Recent advancements in semantic segmentation for 3D indoor scenes have yielded impressive results using large-scale annotated data. However, existing methods operate under the assumption that training and testing data share the same distribution, resulting in performance degradation when evaluated on out-of-distribution scenes. To address the high annotation cost and performance degradation, we introduce a synthetic-to-real domain generalization setting for this task, which trains a robust model on synthetic domains and evaluates its performance on unseen real-world target domains. The domain shift between synthetic and real-world point cloud data mainly lies in the different layouts and point patterns. To address these problems, we first propose a clustering instance mix (CINMix) augmentation technique to diversify the layouts of the source data. In addition, we augment the point patterns of the source data and introduce non-parametric multi-prototypes to ameliorate the intra-class variance enlarged by the augmented point patterns. The multi-prototypes can model the intra-class variance and rectify the global classifier in both training and inference stages. Experiments on the synthetic-to-real benchmark demonstrate that both CINMix and multi-prototypes can narrow the distribution gap and thus improve the generalization ability on real-world datasets.

1 Introduction

Semantic segmentation of 3D indoor point clouds is a fundamental task for scene understanding, with significant potential for real-world applications such as robotics [1] and building information management [30]. Recent advances in deep neural networks and abundant annotated data have led to highly successful 3D semantic segmentation models [23, 24, 29] under fully-supervised settings , achieving remarkable performance on several benchmark datasets [2, 7]. However, these models typically suffer from performance degradation when evaluated on data with different distributions from the training data. The distribution shift between training and testing data mainly lies in the variations in indoor scene layouts and

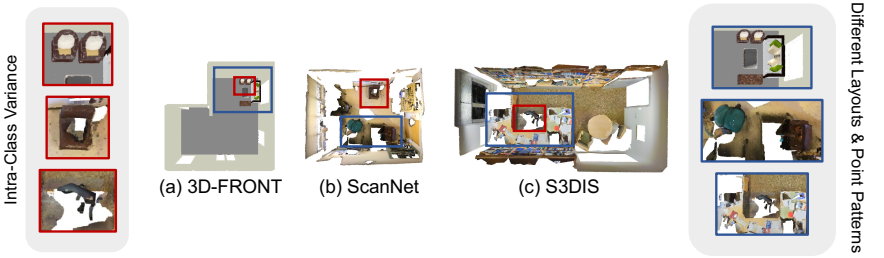

Figure 1: Distribution gap between the synthetic source ((a) 3D-FRONT) and real-world target ((b) ScanNet and (c) S3DIS) data. The distribution gap mainly lies in different layouts and point patterns, which exacerbate the intra-class variance.

point cloud patterns caused by differences in 3D data capture devices and procedures. As a result, the performance of modern 3D semantic segmentation methods is not guaranteed with the existence of domain shift, which is inevitable in real-world applications. Moreover, acquiring diverse annotated data for training these models is costly, hindering the development of highly generalizable models.

In view of the performance degradation and heavy annotation cost of 3D indoor point clouds, unsupervised domain adaptation (UDA) is investigated in 3D indoor scenes [8], which requires the labeled synthetic source data and unlabeled real-world target data. However, such setting still focuses on the testing data in the same distribution as the target real-world data, and the collection of real-world data is also difficult. Thus, we first introduce the synthetic-to-real domain generalization for 3D indoor point cloud semantic segmentation (DG-Indoor-Seg), which only leverages one synthetic dataset to train a robust model that can generalize well to various testing environments in the real-world. While domain generalized semantic segmentation has been studied in 2D driving scenes [38, 40] in recent years, where domain shift is mainly caused by varying weather, time, and conditions in different streets, our work focuses on the 3D indoor scenarios with different domain shift. As shown in Fig. 1, in contrast to the clean and ordered layout of synthetic 3D scenes from 3D-FRONT dataset [10], real-world 3D scenes from ScanNet [7] and S3DIS [2] datasets are characterized by cluttered layouts and complex object-furniture relationships. Additionally, the point clouds of real scenes are relatively sparse and noisy, with severe occlusions compared to synthetic point clouds.

Drawing on our observations of the two types of domain shift in 3D indoor point cloud semantic segmentation, we address the challenge through a two-pronged approach: diversifying the synthetic layouts and investigating the shift of feature representations caused by point pattern augmentation. While 2D domain generalization methods [21, 27], *e.g.*, patch [33] or class [20] mixing, have proven effective in improving generalization ability in driving scenes, these techniques may not work as well for 3D point clouds, which are spatially irregular with varying sizes. Recent works [8, 19] have explored 3D mixing techniques using entire scenes or cuboids for fully-supervised learning and domain adaptation. Nevertheless, these methods can result in unrealistic augmentations that do not benefit out-of-distribution data or may even deteriorate domain generalization performance. To address this limitation, we propose the Clustering Instance Mix (CINMix) technique, which mixes object instances from different scenes under rational geometry constraints to generate diverse and realistic scenes. Due to the unavailability of instance-level labels in semantic segmentation, we apply density-based clustering on each class to roughly separate instances within a

class. After that, multiple instances are sampled and placed in the free location on the floor of an arbitrary scene. Through CINMix, we can generate more source scenes with diverse and cluttered layouts that improve the robustness of our method.

Furthermore, to address the domain shift caused by different point patterns, we introduce simulated noise and occlusion augmentation to the training data following the approach in [8]. As seen in the real scenes in Fig. 1, noise and occlusion alter the geometrical shape of objects. Since 3D semantic segmentation heavily relies on geometrical information, geometrical alteration increases the intra-class variance of the source domain. With the enlarged intra-class variance, the original global classifier that learns a single weight vector for each class may become inadequate in 3D indoor point cloud semantic segmentation. To mitigate this issue, we propose a non-parametric multi-prototypes approach that represents each class with multiple prototypes. This encourages the model to discover diverse and discriminative patterns within each class and optimize the mining of instance-specific information. The multi-prototypes are obtained through clustering, and then updated by moving average. The update process is formulated as the optimal transport problem to uniformly update all the prototypes. During training, the multi-prototypes are used as a non-parametric classifier containing rich intra-class information. During inference, the well-learned multi-prototypes can be used to rectify model predictions from the global classifier by leveraging their encoded instance-specific information.

Our main contributions can be summarized as follows:

- We propose a practical yet challenging setting of domain generalized 3D indoor point cloud semantic segmentation (DG-Indoor-Seg), which trains a robust model on a synthetic source dataset and is able to segment point clouds from other real-world datasets.
- We propose a novel data augmentation technique, *i.e*., clustering instance mix (CINMix), for synthetic-to-real DG-Indoor-Seg. CINMix can generate realistic scenes with complex and cluttered layouts for the source data to narrow the layout domain shift.
- We propose a non-parametric multi-prototypes based classifier to deal with the intra-class variance exacerbated by the point pattern augmentation. The multi-prototypes can be used in both training and inference stages to rectify the model.

2 Related Work

3D Indoor Point Cloud Semantic Segmentation. In the deep learning era, point-based [23, 24, 29] and voxel-based [6, 11, 25] deep neural networks have achieved significant performance under the supervision of fully annotated data. Considering the difficulties in annotating point clouds, semi-supervised [9, 13] and few-shot [34, 39] settings have drawn more attention in recent years. More recently, Ding *et al*. [8] investigate the domain adaptive semantic segmentation in 3D indoor scenes, where labeled source data and unlabeled target data are used to learn a model that can perform well on the target testing data. Different from previous works, in this paper, we first introduce the practical yet challenging domain generalization setting to 3D indoor point cloud semantic segmentation.

Domain Generalization. To tackle the performance degradation in the different domains, domain generalization [15, 22, 35, 36, 37, 40] is widely explored in the community, which learns a robust model with one or multiple source domain(s), aiming to perform well on unseen domains. In recent years, domain generalized semantic segmentation [5, 32, 36, 37]

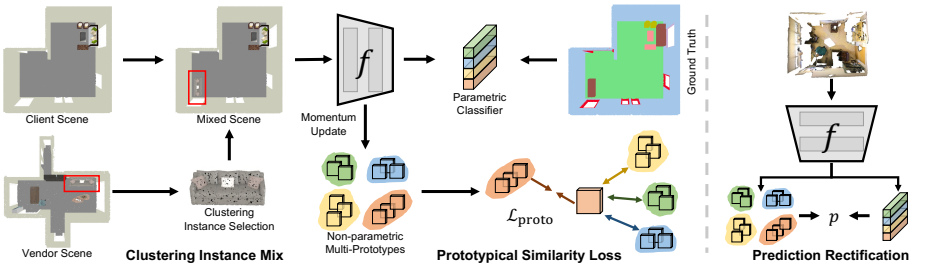

Figure 2: The overall framework of our model for DG-Indoor-Seg. During training, CINMix is first leveraged to generate diverse source samples, and then the model is trained with these samples via the standard cross entropy loss and a prototypical similarity loss based on the momentum-updated multi-prototypes. During inference, the multi-prototypes are used to rectify the model prediction.

in 2D driving scenes has been studied, where the domain shift lies in the changeable environments and styles. Briefly, one mainstream of these works focuses on designing style-invariant modules, such as whitening [5] and instance normalization [22]. Another mainstream engages in diversifying the source samples with style augmentation [37, 40] and image randomization [12, 32]. More recently, a pioneer work [14] investigates DG in 3D driving scenes, where the distribution shift mainly lies in different LiDAR sensor configurations. Different from previous works on driving scenes, we study domain generalized semantic segmentation on 3D indoor point clouds, where the domain shift lies in the point patterns and layouts.

Data Augmentation for Semantic Segmentation. In addition to standard transformations, such as geometry transformation and color jitter, many advanced data augmentation techniques can improve both in-distribution and out-of-distribution performance. CutMix [33] crops a patch from one sample and mixes it with another sample, which can improve the performance of both semantic segmentation and image classification. ClassMix [20] is proposed for semi-supervised semantic segmentation, which takes pixels of several classes in one image to another one. As for 3D point clouds, Mix3D [19] directly concatenates two samples to train the semantic segmentation model. Cuboid Mixing [8] is designed for domain adaptive 3D point cloud semantic segmentation, which splits each scene into several cuboids and mixes the cuboids from source and target domains. In this paper, we propose the CINMix to generate diverse and realistic layouts within the source domain to overcome the domain shift regarding layout change.

3 Methodology

Problem Definition. Synthetic-to-real domain generalization (DG) focuses on training a robust model with *one* labeled synthetic domain $\mathcal{S}$, where the model is expected to perform well on unseen domains $\{\mathcal{T}_1, \mathcal{T}_2, \cdots\}$ of different real-world distributions. The source and target domains share the same label space (N_C categories).

3.1 Overview

As illustrated in the introduction, the domain gap mainly lies in the different layouts and point patterns. To this end, we propose the Clustering INstance Mix (CINMix) to diversify

Figure 3: Visualization of one thing class (*i.e.*, chair) and its clustered instance groups after density-based class clustering.

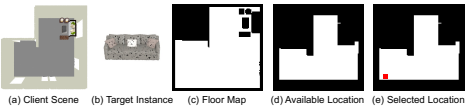

Figure 4: Procedures of geometry constrained mixing.

source layouts and non-parametric multi-prototypes to address the intra-class variance exacerbated by point pattern augmentations. The overall framework of our proposed method is shown in Fig. 2. Specifically, we randomly select the point clouds of two scenes during training, and consider one point cloud scene as vendor scene (to sample instances from it) while treating the other one as client scene (to insert the sampled instances into it). The thing classes of the vendor scene are clustered into several groups, and then the available groups are mixed to the client point cloud scene subjecting to the geometry constraints. After the CINMix step, the mixed sample is augmented with virtual scan simulation [8] and standard geometry augmentation techniques. Then it is fed into the model that consists of a feature encoder and a parametric classifier to generate its prediction, which is supervised by the ground truth via the standard cross entropy loss. In addition, we utilize non-parametric multi-prototypes to compute a prototypical similarity loss as the additional supervision. The non-parametric multi-prototypes are used to model the intra-class variance and is updated by moving average. During inference, the multi-prototypes are used to rectify the prediction from the parametric classifier. With the help of CINMix and multi-prototypes, our model can deal with various unseen scenes in the real world.

3.2 Clustering Instance Mix

Considering the irregular and unordered attributes of point clouds, we propose the CINMix to generate diverse and realistic source scenes under geometry constraints. CINMix includes two steps: the density-based class clustering and geometry constrained mixing.

Density-based Class Clustering. 2D based ClassMix [20] cannot be used since one class in the 3D space may not be constrained in a fixed size range like that in the 2D images. For example, when putting the "sofa" class in a living room into another kid room, directly mixing them can generate unrealistic results since the living rooms are commonly larger and the sofas can spread across the room. As shown in previous works [18, 26, 28], the geometry is much easier to be controlled when the instance labels are provided. In addition, the thing class, *e.g.*, chair and table, can be easily separated into different parts by density-based clustering in 3D indoor scenes. The clustering parts can contain one instance or multiple nearby instances, but each part is within a relatively small scale, which can be mixed into other scenes. Therefore, we use DBSCAN [9] for each thing class in the vendor scene to get one or multiple instances within a class (Fig. 3).

Geometry Constrained Mixing. Given the clustering instances, where to fuse them into the client scene remains a problem. Since the point clouds of vendor and client scenes are not aligned, directly concatenating them may lead to unrealistic results. To this end, we introduce two major geometry constraints for the mixing: the mixed instances should be (1) on the floor, and (2) have no overlap with the existing classes in the client scene. The overall mixing procedures are shown in Fig. 4. First, a non-class floor map is obtained from the client scene. Then, erosion is applied to the floor map by treating the shape of the clustering

instance as the kernel size. Finally, the clustering instance is inserted into one of the available locations randomly. Note that we mix multiple clustering instances to one client scene and apply random rotation along z axis for each instance to improve the diversity of our CINMix augmentation.

3.3 Non-parametric Multi-prototypes

Different point patterns, represented by noise and occlusion, is another severe domain shift type limiting the generalization ability. We use virtual scan simulation [8] to add noise and occlusion into the synthetic scenes, which can augment the point patterns of these scenes. However, the addition of the noise and occlusion leads the synthetic data to have a larger intra-class variance that is similar to the variance of the complex real-world data (see Fig. 1). To tackle the enlarged intra-class variance, we introduce the non-parametric multi-prototypes. These multi-prototypes can encourage the model to discover intra-class discriminative patterns and optimize the mining of the instance-specific information.

Multi-prototypes Initialization. The multi-prototypes are initialized by clustering the features of source data. Specifically, we first calculate the mean feature of each class c for each augmented source sample x^i:

$$\bar{f}_c^i = \frac{\sum_{j=1}^{n} f(x_j^i) * \mathbb{1}\left(y_j^i == c\right)}{\sum_{j=1}^{n} \mathbb{1}\left(y_j^i == c\right)}, \tag{1}$$

where n denotes the number of points in x^i. $f(\cdot)$ is the feature extractor and y_j^i denotes the label for point x_j^i Given all the class-wise features in the source domain, we cluster features of each class into K clusters and get the K cluster centroids as the multi-prototypes $P_c \in \mathbb{R}^{K\times D}$. K-means++ [3] is adopted in this paper.

$$P_c = \mathrm{KPP}([\bar{f}_c^1, \bar{f}_c^2, \cdots, \bar{f}_c^{N_s}], K), \tag{2}$$

where KPP denotes the K-means++ clustering algorithm and K is the number of clusters. N_s denotes the number of samples in the source data.

Momentum Update via Optimal Transport. Since feature representations are changing along with the training, the initialized multi-prototypes should be updated for accurate representations. Intuitively, each prototype in multi-prototypes can be updated by the feature(s) that are closest to the prototype. However, this might lead to a degenerate solution that only one prototype is updated when all features are close to this specific prototype. Inspired by [31, 41], we adopt the solution of optimal transport to split the features uniformly to K prototypes.

Prototypical Similarity Loss. Given the multi-prototypes $P_c \in \mathbb{R}^{K\times D}$ for class c, we define the probability of one point x_j belonging to the class c as the maximum similarity to the multi-prototypes:

$$s(x_j, c) = \max\left(f(x_j) \cdot P_c^\top\right). \tag{3}$$

Instead of the global representation for each class, $s(x_j, c)$ can model the most similar representation of the same class and the most confusing representation from other classes. Optimizing over such probability can force the model to discover intra-class discriminative

patterns and support the learning of instance-specific details. Consequently, we define the prototypical similarity loss as:

$$\mathcal{L}_{\text{proto}}(x,y) = -\frac{1}{n}\sum_{j=1}^{n} y_j \log \frac{\exp\left(s(x_j,c_j)\right)}{\sum_{k=1}^{N_C}\exp\left(s(x_j,c_k)\right)}, \tag{4}$$

where n and N_C denote the number of point clouds in x and the number of classes, respectively.

Prediction Rectification. Compared with the global classifier, the non-parametric multi-prototypes contain more intra-class discriminative patterns, which can serve as the rectification for the model prediction to alleviate the instance-specific impact. For example, armchair may be closer to the unified representation of sofa instead of chair, which may lead to an incorrect prediction from the global classifier. However, one of the multi-prototypes of the chair can represent the armchair, and integrating such information can rectify the result. Thus, we leverage the similarity to the multi-prototypes as the weight to rectify the global classifier prediction. The weight is obtained by:

$$w(x_j,c) = \frac{\exp\left(s\left(x_j,c\right)\right)}{\sum_{k=1}^{N_C}\exp\left(s\left(x_j,c_k\right)\right)}. \tag{5}$$

Then the model prediction of the global classifier Φ is rectified by $w(x_j,c)$:

$$p(x_j,c) = w(x_j,c) * \Phi\left(f(x_j),c\right), \tag{6}$$

where $p(x_j,c)$ is the rectified prediction for the point x_j.

Training Objective Equipped with the clustering instance mix and multi-prototypes, the model is optimized by the combination of the cross entropy loss over the global classifier and the proposed prototypical similarity loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(x,y) + \mathcal{L}_{\text{proto}}(x,y). \tag{7}$$

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on the synthetic-to-real domain generalization setting. The synthetic dataset 3D-FRONT [10] is leveraged as the training data, and the model is evaluated on two real-world datasets, *i.e.* ScanNet [7] and S3DIS [2]. 3D-FRONT [10] is a large-scale dataset of 3D indoor scenes, containing 18,968 rooms with 13,151 3D furniture objects. Following [8], 4,995 rooms from 3D-FRONT [10] are adopted as the training data. For the real-world target domains, ScanNet [7] is a large-scale real-world dataset for point cloud scene understanding, containing 1,201 training, 312 validation, and 100 testing scans. S3DIS [2] is a real-world 3D semantic segmentation dataset with 271 scenes from 6 areas. Following [8, 23], 68 samples in the fifth area are used as the validation data.

Evaluation Metric. We use eight shared categories across the three datasets for training and evaluation, and we adopt the mean Intersection-over-Union (mIoU) over the eight categories to evaluate the model performance.

Implementation Details. Following [8], we adopt sparse-convolution-based U-Net model [11] as our backbone. We use AdamW [17] optimizer with an initial learning rate 6×10^{-4} and

Table 1: Domain generalization results on 3D-FRONT → ScanNet & S3DIS benchmark.

Target	Method	wall	floor	chair	sofa	table	door	wind.	bksf.	mIoU
ScanNet	Source Only	70.24	86.41	61.53	31.90	50.95	6.60	3.02	31.93	42.82
	ClassMix [20]	68.52	80.54	52.40	42.67	28.94	7.20	8.36	39.28	40.99
	Cuboid Mixing [8]	73.02	89.04	62.56	36.87	45.89	7.22	0.36	37.59	44.07
	Mix3D [19]	71.87	87.23	60.68	**48.15**	43.15	**8.56**	0.65	27.83	43.52
	VSS [8]	68.80	89.05	57.91	45.49	47.22	8.47	10.09	36.42	45.43
	DODA-S [8]	71.65	**89.23**	60.72	41.78	48.13	8.33	6.76	39.48	45.76
	Ours	**73.42**	88.07	**62.72**	43.56	**52.14**	7.71	**14.20**	**44.44**	**48.28**
S3DIS	Source Only	68.94	92.63	50.86	16.61	47.50	9.00	0.87	22.74	38.64
	ClassMix [20]	70.53	93.68	56.04	7.89	34.25	10.51	0.20	35.52	38.58
	Cuboid Mixing [8]	69.18	88.41	56.83	9.43	34.36	11.47	0.23	24.96	36.86
	Mix3D [19]	70.73	92.64	53.17	9.04	50.51	7.97	1.17	25.88	38.89
	VSS [8]	69.71	94.52	58.11	**26.43**	40.83	**21.16**	23.14	49.24	47.89
	DODA-S [8]	71.33	92.11	59.32	20.02	36.24	13.47	7.63	38.75	42.36
	Ours	**76.00**	**94.66**	**66.22**	17.79	**53.20**	21.12	**29.84**	**51.14**	**51.25**

Table 2: Ablation studies on CINMix and non-parametric multi-prototypes. "Para." denotes the parametric classifier. "N-Para. T" and "N-Para. I" denote using the non-parametric multi-prototypes in training and inference stages, respectively.

No.	CINMix	Para.	N-Para. T	N-Para. I	ScanNet	S3DIS	Mean
1	✗	✓	✗	✗	45.43	47.89	46.66
2	✓	✓	✗	✗	46.23	48.92	47.58
3	✗	✓	✓	✗	46.63	48.29	47.46
4	✗	✓	✓	✓	46.78	49.06	47.92
5	✓	✗	✓	✗	46.88	46.26	46.57
6	✓	✓	✓	✗	48.00	50.02	49.01
7	✓	✓	✓	✓	**48.28**	**51.25**	**49.77**

weight decay 0.01 to optimize the model. The polynomial decay [16] with a power of 0.9 is used as the learning rate scheduler. All models are trained for 100 epochs with a batch size of 32. The number of multi-prototypes K, the smoothness parameter λ, and the update momentum m are set to 3, 20, and 0.999, respectively. For the DBSCAN algorithm in density-based class clustering, we set the maximum distance between neighbor points to 0.2 and the number of points in a neighborhood to 100.

Baselines. Previous works [5, 37, 40] on DG-Seg in 2D images consider the domain shift as image style change. Hence, these works cannot be applied to 3D point clouds. As we are the first work focusing on domain generalized semantic segmentation in 3D indoor scenes, we compare our method with the model trained only with cross entropy loss on the source data (source only). In addition, since data augmentation techniques are commonly adopted to improve the generalization ability, we also compare our framework with point cloud based augmentation methods, including ClassMix [20], Cuboid Mixing [8], Mix3D [19], VSS [8] and DODA-S [8].

4.2 Synthetic-to-Real Domain Generalization

In Tab. 1, we compare our model with baselines on the 3D-FRONT → ScanNet & S3DIS benchmark. We make the following observations. **First**, on the ScanNet dataset, our model

outperforms the source only model on all the categories and achieves the mIoU of 48.28%, yielding an improvement of 4.76% and 2.85% over the Mix3D and VSS models respectively. **Second**, our model achieves the best performance on 3 thing classes (chair, sofa, table, and bookshelf) on the ScanNet dataset, demonstrating the effectiveness of mixing the instances of thing classes and mining the intra-class variance. **Third**, our model also achieves the state-of-the-art performance on S3DIS dataset. In addition, most of previous data augmentation methods fail to gain improvement on S3DIS dataset, while our model obtains a mIoU of 51.25%, outperforming the source only baseline by 12.61%.

4.3 Ablation Studies

CINMix. CINMix is proposed to diversify the layouts in the source domain by extracting instances from the vendor scene and mixing them to the client scene under the two geometry constraints. The instances are obtained from the density-based clustering on the points of thing classes. As shown in the 2nd row of Tab. 2, CINMix can improve the performance on ScanNet and S3DIS datasets by 0.8% and 1.03%, respectively, which demonstrates the effectiveness of CINMix in generating diverse and realistic scenes.

Non-parametric Multi-prototypes. We propose non-parametric multi-prototypes to address the intra-class variance in the 3D indoor scenes. The proposed multi-prototypes are used in both training and inference stages to learn instance-specific details and rectify the model prediction, respectively. **First**, comparing the 3rd row with the baseline model, introducing non-parametric multi-prototypes in the training stage can improve the performance on both of the target datasets, yielding an improvement of 0.8% on the average mIoU. **Second**, we compare using the parametric and non-parametric classifiers separately in the 2nd and 5th rows of Tab. 2. The non-parametric classifier can improve the performance on ScanNet but achieves worse performance than the parametric one on S3DIS. We conjecture that the test set of S3DIS is less complex than that of ScanNet, and thus the target data may be biased to some of the prototypes without the guidance of global representation, leading to incorrect predictions. **Third**, when combining the parametric and non-parametric classifiers, the model can gain consistent improvement no matter using CINMix or not. In addition, the improvement of the combined classifiers is more significant when CINMix is leveraged. This is because CINMix generates more diverse scenes, which can boost the capacity of the non-parametric classifier. **Finally**, the non-parametric multi-prototypes can also be used in the inference stage to rectify the model prediction by alleviating the impact of instance-specific information. As shown in the last row of Tab. 2, the rectification can slightly improve the performance on ScanNet by 0.28% in mIoU but gains great improvement on S3DIS by 1.23% in mIoU. The reason is that with the guidance of global representation, the multi-prototypes can better address instance-specific details of S3DIS instead of biased to some specific prototypes. All the results demonstrate the effectiveness of non-parametric multi-prototypes in ameliorating the intra-class variance and improving the generalization ability.

Table 3: Comparison with different augmentation.

Augmentation	ScanNet	S3DIS	Mean
VSS+N-Para.	46.78	49.06	47.92
+ClassMix [21]	46.63	45.74	46.19
+Cuboid Mixing [8]	47.74	45.58	46.66
+Mix3D [13]	47.99	48.22	48.11
+CINMix	**48.28**	**51.25**	**49.77**

Table 4: Different degree of clustering constraints.

Degree of constraints	ScanNet	S3DIS	Mean
Strict	45.47	47.41	46.44
Loose	48.11	50.92	49.52
Appropriate	**48.28**	**51.25**	**49.77**

(a) Strict Constraints (b) Loose Constraints (c) Appropriate Constraints

Figure 5: Visualization of different clustering constraints. The two beds are in one room scene.

5 Evaluation

Data Augmentation Techniques. To verify the effectiveness of CINMix, we further compare different data augmentation techniques when equipped with our framework, *i.e.*, using VSS and non-parametric classifier (VSS+N-Para.). In Tab. 3, compared with the model without additional augmentation, all four augmentation techniques can improve the performance on ScanNet. However, ClassMix, Cuboid Mixing and Mix3D deteriorate the performance on S3DIS, since the complete scenes in S3DIS suffer from the missing and incomplete point cloud data introduced by these augmentations. Compared with the above methods, our CINMix can generate diverse and realistic source samples, improving the performance on both datasets by a large margin.

Clustering Constraints Analysis. We empirically set the maximum distance $M_p = 0.2$ and neighborhood points number $N_p = 100$ of DBSCAN [9] to generate complete instances within one class. To analyze the sensitivity of the two hyper-parameters, we visualize the clustering results under strict ($N_p = 300$), loose ($M_p = 0.5$) and appropriate constraints in Fig. 5, where we use each color denotes one cluster. Large object can be separated into several parts under strict constraints and some parts are even viewed as invalid points (white points), while multiple instances are clustered together under loose constraints (two beds are viewed as one instance). However, as illustrated in Tab. 4, since the instances of one thing class are commonly not quite close, leveraging relatively loose constraints do not have strong influence on the model performance. In contrast, strict constraints can readily separate one instances into multiple meaningless parts, and utilizing such parts to represent this class can drastically impair the performance.

6 Conclusion

In this paper, we introduce a new setting of domain generalized semantic segmentation in 3D indoor scenes, which trains a robust model with synthetic data and aims at segmenting real-world data from unseen domains. By investigating this challenging but practical problem, we identify the primary domain gap between synthetic and real-world data as differences in layouts and point patterns. To address the domain shift, we propose two novel techniques: clustering instance mix augmentation to diversify the source layouts, and non-parametric multi-prototypes to handle the enlarged intra-class variance resulting from augmented point patterns. Our proposed method demonstrates significant generalization ability and outperforms baseline models by a large margin on two real-world benchmark datasets.

Acknowledgements Na Zhao was a visitor at NUS when this work was done. This work is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021).

References

[1] Rareş Ambruş, Sebastian Claici, and Axel Wendt. Automatic room segmentation from unstructured 3-d data of indoor environments. *IEEE Robotics and Automation Letters*, 2017.

[2] 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 *CVPR*, 2016.

[3] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *ACM SODA*, 2007.

[4] Mingmei Cheng, Le Hui, Jin Xie, and Jian Yang. Sspc-net: Semi-supervised semantic 3d point cloud segmentation network. In *AAAI*, 2021.

[5] Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne T Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. In *CVPR*, 2021.

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

[7] 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.

[8] Runyu Ding, Jihan Yang, Li Jiang, and Xiaojuan Qi. Doda: Data-oriented sim-to-real domain adaptation for 3d semantic segmentation. In *ECCV*, 2022.

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

[10] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *ICCV*, 2021.

[11] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018.

[12] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *CVPR*, 2021.

[13] Li Jiang, Shaoshuai Shi, Zhuotao Tian, Xin Lai, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In *ICCV*, 2021.

[14] Hyeonseong Kim, Yoonsu Kang, Changgyoon Oh, and Kuk-Jin Yoon. Single domain generalization for lidar semantic segmentation. In *CVPR*, 2023.

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *AAAI*, 2018.

[16] Wei Liu, Andrew Rabinovich, and Alexander C Berg. Parsenet: Looking wider to see better. In *CoRR*, 2015.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[18] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *CVPR*, 2019.

[19] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3d: Out-of-context data augmentation for 3d scenes. In *3DV*, 2021.

[20] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *WACV*, 2021.

[21] Fei Pan, Sungsu Hur, Seokju Lee, Junsik Kim, and In So Kweon. Ml-bpm: Multi-teacher learning with bidirectional photometric mixing for open compound domain adaptation in semantic segmentation. In *ECCV*, 2022.

[22] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *ECCV*, 2018.

[23] 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.

[24] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Abed Al Kader Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In *NeurIPS*, 2022.

[25] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *CVPR*, 2017.

[26] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *ECCV*, 2022.

[27] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *WACV*, 2021.

[28] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, 2021.

[29] 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 TOG*, 2019.

[30] Chao Yin, Boyu Wang, Vincent JL Gan, Mingzhu Wang, and Jack CP Cheng. Automated semantic segmentation of industrial point clouds using respointnet++. *Automation in Construction*, 2021.

[31] Asano YM., Rupprecht C., and Vedaldi A. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020.

[32] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *ICCV*, 2019.

[33] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019.

[34] Na Zhao, Tat-Seng Chua, and Gim Hee Lee. Few-shot 3d point cloud semantic segmentation. In *CVPR*, 2021.

[35] Yuyang Zhao, Zhun Zhong, Fengxiang Yang, Zhiming Luo, Yaojin Lin, Shaozi Li, and Sebe Nicu. Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification. In *CVPR*, 2021.

[36] Yuyang Zhao, Zhun Zhong, Zhiming Luo, Gim Hee Lee, and Nicu Sebe. Source-free open compound domain adaptation in semantic segmentation. *IEEE TCSVT*, 2022.

[37] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning for domain generalized semantic segmentation. In *ECCV*, 2022.

[38] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning: A unified framework for visual domain generalization. *IJCV*, 2023.

[39] Ziyu Zhao, Zhenyao Wu, Xinyi Wu, Canyu Zhang, and Song Wang. Crossmodal few-shot 3d point cloud semantic segmentation. In *ACM MM*, 2022.

[40] Zhun Zhong, Yuyang Zhao, Gim Hee Lee, and Nicu Sebe. Adversarial style augmentation for domain generalized urban-scene segmentation. In *NeurIPS*, 2022.

[41] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *CVPR*, 2022.