CACE: Sim-to-Real Indoor 3D Semantic Segmentation via Context-aware Augmentation and Consistency Enforcement

Tsung-Yu Chen[1,3] Luyu Yang[2] Tzu-Yu Chuang[3] Shang-Hong Lai[3]
[1] Carnegie Mellon University [2] University of Maryland [3] National Tsing Hua University
{lear1007, yangluyu123, claire510072}@gmail.com lai@cs.nthu.edu.tw

Abstract

*Indoor 3D domain adaptation for semantic segmentation is an understudied task. The first unsupervised sim-to-real benchmark was only proposed recently. Existing methods try to modify the source domain data by simulating the occlusion and noise pattern of the target domain. However, this methodology unrealistically demands a clear definition of the real-world data patterns, and is highly dependent on the simulation quality. In this paper, we propose a novel adaptation framework via **C**ontext-aware **A**ugmentation and **C**onsistency **E**nforcement (CACE). Our CACE framework consists of two modules, a space and context-aware augmentation module that is invariant of target data pattern and domain gaps, and a carefully designed self-supervision module that maximizes the utility of the augmented data. Our CACE surpasses the state-of-the-art method by over 6% on the indoor 3D sim-to-real benchmark 3D-FRONT → ScanNet.*

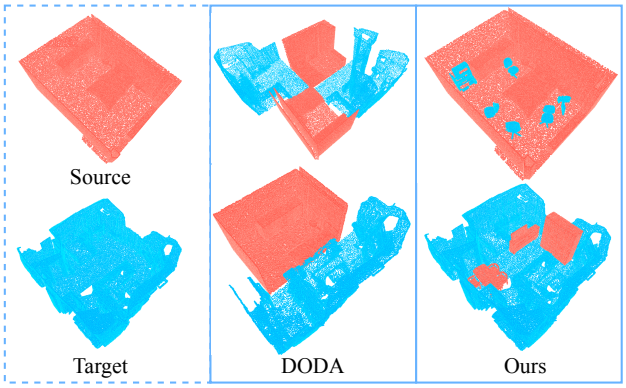

Figure 1. Comparison of augmentation methods for intermediate domains. *First column*: The original source and target domains. *Second column*: Augmented intermediate domains by SOTA method DODA [10] using a cut-and-glue strategy. *Third column*: Augmented results from the proposed method CACE, which is by carefully selecting objects from another domain and arranging objects in a physics-aware manner into the scene.

1. Introduction

Indoor 3D semantic segmentation is a crucial task for object detection and scene reconstruction in indoor environments [1, 14, 16, 27, 28, 31, 40, 49, 52, 69, 76]. Recently, the task has received incredible attention due to its wide applications to robotics [1, 40, 76], virtual reality [16, 28] and human-computer interaction [14, 49, 52]. Like many other 3D tasks [2,4,34,47,54], 3D semantic segmentation is powered by the success of many data-driven approaches. Thus, it is inevitably challenged by the shortage of large-scale annotated data.

An appealing alternative to using real-world data is to train a model using a large amount of synthetic data [5, 10, 39,44], and then generalize it to real-world scenarios. However, the success of this alternative requires tackling the notorious sim-to-real domain gaps, which calls for effective 3D domain adaptation approaches.

3D domain adaptation for indoor scenes is an understudied task. The first unsupervised sim-to-real benchmark on 3D indoor semantic segmentation was only proposed recently [10]. The method tries to mimic the occlusion and noise pattern of the real-world data, in order to yield more transferable simulation data. However, this methodology has two major drawbacks. First, the method requires a clear definition of the real-world data patterns, which is infeasible in reality since the data could contain a mixture of various noise types. In [10], they jitter source points to imitate noises in the real scenes, which might be oversimplified. Second, the method relies heavily on the success of the noise pattern simulation in the first stage, which makes it risky for the model to apply transfer learning.

Therefore, in this paper, we propose an approach that is invariant of the noise patterns of the real data. To achieve this goal, we aim at a method that constructs an intermediate domain that can be generalized to arbitrary sim-to-real domain gaps. We analyze that the sim-to-real domain gaps can be summarized into two categories: room layout and

Figure 2. Example of a scene with broken context integrity. When adding the object class "chair" into a bedroom scene, a reasonable choice is a vanity chair. However, due to the lack of fine-grained class labels, a set of dining chairs got added to the bedroom, breaking the scene context.

object complexity. The room layouts of synthetic scenes are mostly regular and clean, while the real-world layouts are usually cluttered. Moreover, the design of real-world objects is often more complicated in varieties, with personalized object contexts. Instead of directly learning these data patterns, we propose a novel augmentation strategy that integrates objects and scenes from both sim and real domains, with consideration of the object context. We name it Space and Context-aware Augmentation (SCA). As a result, we have circumvented the brutal-force mimic of the real-world data patterns and obtained an intermediate domain that encompasses traits from both the source and target domains, as illustrated in Fig. 1. The proposed augmentation strategy is invariant with noise patterns and can be easily expanded to more complicated scenarios. Moreover, it is not restricted to sim-to-real domain adaptation but can be applied to any two domains with a distribution gap.

However, the augmented scenes from the intermediate domain do not always contribute to the model during training. We observe that some newly constructed scenes are not reasonable and can potentially disrupt the overall context integrity. For example, if we place a dining table set in a bedroom scene (as shown in Fig. 2), the model can be confused about the room type and misjudge the bed as the extension of the dining table since it is adjacent to dining chairs, leading to incorrect predictions. To address this, we identify common objects for scenes with and without augmentations, applying a consistency objective to ensure uniform predictions among common objects. To further refine the use of augmented scenes, we combine it with a mean-teacher model for stable, high-quality pseudo-labels and name it the Consistency-Enforced Self-Supervision (CESS) module. By combining the proposed SCA and CESS modules, we enable the semantic parser to gradually adapt to the target domain while maintaining context integrity and reducing the impact of augmentation side effects.

Our major contributions are summarized as follows: **i)** We propose a novel method SCA to construct an intermediate domain between source and target domains. Our method is simple and easy to generalize, independent of the noise pattern of the target data, and can be applied to arbitrary domain gaps. **ii)** We propose the CESS module that enforces consistency and adapts to the target domain gradually based on a self-supervision framework.

2. Related Work

Unsupervised Domain Adaptation (UDA) is an extensively studied task, in which the goal is to adapt machine learned models from a label-rich source domain to an unlabelled target domain. Many UDA methods [43, 61, 66, 79] try to match the source and target domain distributions by minimizing their divergence in the feature space. Domain adversarial learning [15, 19, 50, 67, 70, 75] is one mainstream approach, in which the disparity between source and target feature spaces is mitigated through a domain discriminator that predicts whether the incoming feature is from the source or target domain. Another popular genre of methods [37, 55, 56] is based on the cluster assumption [18] within the target domain. They follow the assumption that the boundaries of the classifier should not cross high-density target data regions [56]. Some approaches employ self-training and progressively label the data from the target domain with pseudo-labels [21, 36, 38, 45, 80]. A few more recent methods [25, 26, 60, 65, 68] employ the contrastive learning, and show that although the learnt features are not domain-invariant, they still generalize well to the target domain by disentangling domain and class information.

3D Indoor Semantic Segmentation Domain Adaptation. Compared to the previous efforts [23, 30, 73, 78, 82] on 3D domain adaptation for outdoor scene parsing, indoor semantic segmentation domain adaptation is relatively new and less studied [10, 42]. The main reason that outdoor-based approaches cannot be applied to indoor scenes is that they often adopt LiDAR-specific range image format. Differently, indoor scenes are often constructed by RGB-D sequences [10]. Moreover, the process of scene construction for outdoor and indoor scenes is completely different, resulting in sub-optimal solution applying outdoor approaches. In [10], they try to mimic the data pattern of the real data and yield more transferable simulation data. However, the method has an underlying assumption that the noise and occlusion patterns can be clearly defined, which is unrealistic given how diversified real indoor scenes can be [7, 77]. In this paper, we propose an approach that is invariant to the data patterns and domain gaps.

Point Cloud Mixing and Mean-Teacher. Point cloud mixing and mean-teacher architecture have been explored in point cloud segmentation [48] as well as in other cross-domain tasks such as domain adaptation in point cloud semantic segmentation [59, 83] and indoor domain adapta-

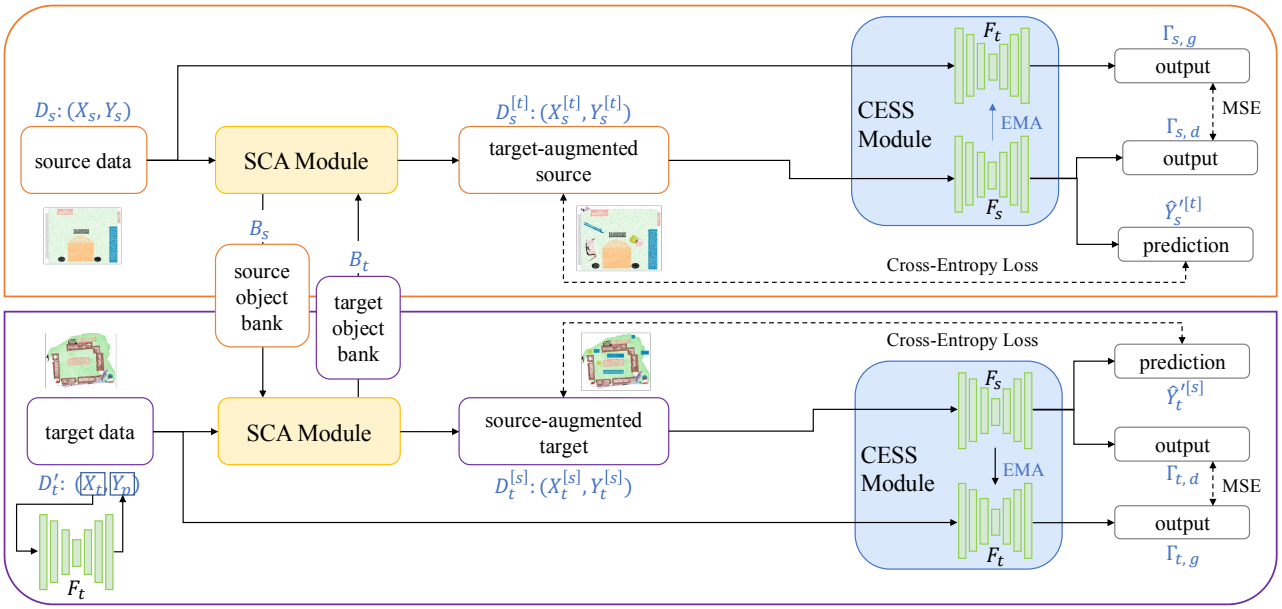

Figure 3. The overall pipeline of CACE. *Top*: Data flow of source data $D_s : (X_s, Y_s)$; *Bottom*: Data flow of target data and its pseudo label $D_t' : (X_t, Y_p)$. Note that the ground truth target domain label Y_t is unavailable in our setting. Given each batch, we first construct the intermediate domain with the SCA module by exchanging objects from the source object bank B_s and target object bank B_t. Thus we obtain the target-augmented source data $D_s^{[t]}$ and the source augmented target data $D_t^{[s]}$. After that, in the CESS module, we train the student model F_s in the supervised branch using $D_s^{[t]}$ and the teacher model F_t in the supervising branch. We perform a similar process with $D_t^{[s]}$. Note that all F_s in the figure share the same weights, and so do all F_t.

tion class-incremental object detection [84]. However, our method differs from theirs in three key ways: (1) Context-aware cluster sampling: Mix3D [48] augmented two scenes by mixing their points directly to create an out-of-context environment for objects; CoSMix [59], CINMix [83] and DA-CIL [84] select objects to mix in a scene-agnostic manner, while our cluster sampling component is context-aware. This means that we consider the spatial relationships between objects when selecting them for mixing, which results in more realistic and informative mixed point clouds. (2) Physics-aware arrangement module: CoSMix, CINMix and DA-CIL naively copy and paste objects into the scene when mixing, which can lead to unrealistic placements (e.g., furniture floating in the air or cars overlapping). Whereas, DODA simply breaks and glues different scenes into a new one as shown in Fig. 1. Our physics-aware arrangement module ensures that mixed objects are placed in a realistic and physically plausible manner. (3) Mean-teacher architecture with consistency loss: CoSMix and DA-CIL both use the mean-teacher architecture to generate pseudo labels and self-supervision. However, we further utilize the mean-teacher architecture with consistency loss to reduce the artefacts produced by argumentation.

3. Proposed Method

As shown in Fig. 3, we present a novel training framework for unsupervised domain adaptation for indoor 3D semantic segmentation, from source to target domain, i.e. from synthetic to real point clouds. We define N_K as the number of common classes in the source and target domain datasets, $N_t^{(i)}$ and $N_s^{(i)}$ as the number of points in the i-th point cloud in the source and target domain datasets, respectively. In the setting of unsupervised domain adaptation, our goal is to train a model with the source domain point clouds and labels $D_s = \{(X_s^{(i)}, Y_s^{(i)})\}_{i=1}^{N_s}$, where $X_s^{(i)} \in \mathbb{R}^{N_s^{(i)} \times d_X}$ and $Y_s^{(i)} \in \{0,1\}^{N_s^{(i)} \times N_K}$ as well as the target domain point clouds without ground-truth labels $D_t = \{X_t^{(i)}\}_{i=1}^{N_t}$ that predicts the semantic labels $\{\hat{Y}_t^{(i)}\}_{i=1}^{N_t}$, where $\hat{Y}_t^{(i)} \in \{0,1\}^{N_t^{(i)} \times N_K}$, for the target domain point cloud D_t. The dimension of point clouds is set to d_X here without loss of generality. We use $d_X = 3$, which is the point location in Euclidean space in our experiment.

We aim to adapt the knowledge learned from the source domain to improve the semantic segmentation performance on the target domain, without using any labelled data in the target domain. Our training framework consists of two modules, Space and Context-aware Augmentation (SCA) and Consistency-Enforced Self-Supervision (CESS). To bridge the domain gap between synthetic and real data, we propose SCA, which creates an intermediate domain that gradually adapts the semantic parser to the target domain. SCA incorporates an augmentation algorithm that strategically places objects from the source domain into the target domain and vice versa. By carefully selecting and arrang-

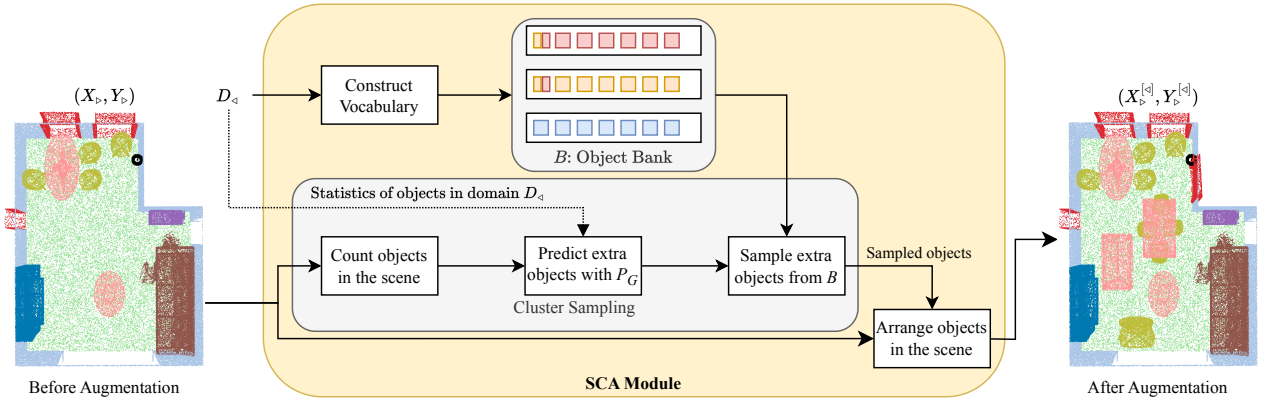

Figure 4. The architecture of SCA module. The *Vocabulary Construction* component extracts objects from all scenes in training scenes in dataset $D_\triangleleft$ and assembles them into a comprehensive vocabulary bank B. Within B, individual queues are designated for each object class, housing all instances of the class within $D_\triangleleft$. We apply online augmentation when training. We first compute the existing objects in a scene via a *Counting Algorithm*. Then we estimate the number of objects to be added to the scene via P_G. The *Extra Object Sampling* component then selectively retrieves objects from the corresponding queues in B, forming an object set C'. Then the *Cluster Arrangement* ensures a physics-aware and contextually informed placement of the objects. Without losing generality, $D_\triangleright$ is the data in $\triangleright$ domain; $(X_\triangleright^{[\triangleleft]}, Y_\triangleright^{[\triangleleft]})$ is a scene in $\triangleright$ domain augmented with objects in $\triangleleft$ domain.

ing objects, SCA preserves both local and global context in the augmented scenes. While our intent is to maintain scene context inside the intermediate domain, the addition of extra clusters can potentially disrupt the scene context. To address this issue, we introduce the CESS module, which incorporates a self-supervision head in the model architecture under a Mean-Teacher paradigm. This module effectively mitigates the detrimental effects of the augmented objects introduced by SCA and generates robust and consistent pseudo labels, enhancing overall model reliability and coherence. By combining the SCA and CESS modules, our approach enables the semantic parser to gradually adapt to the target domain while maintaining context integrity and reducing the impact of augmentation side effects.

We adopt the two-stage training pipeline in [10]. The first stage is the pre-training stage, where F_s is trained with $\{(X_s^{(i)}, Y_s^{(i)})\}_{i=1}^{N_s}$; the second stage is the self-training stage. In the second stage, the structurally identical models F_t and F_s, are initialized with weights of F_s after the pre-training stage. Since the target data labels $\{(Y_p^{(i)})\}_{i=1}^{N_t}$ are not available to the model in our setting, we generate pseudo labels $\{(Y_p^{(i)})\}_{i=1}^{N_t}$ with the prediction of F_t on $\{(X_t^{(i)})\}_{i=1}^{N_t}$ every N_u epoch. $\{(X_s^{(i)}, Y_s^{(i)})\}_{i=1}^{N_s}$ and $\{(X_t^{(i)}, Y_p^{(i)})\}_{i=1}^{N_t}$ will be used in the self-training stage. F_t will be updated via the exponential moving average of F_s.

3.1. Space and Context-aware Augmentation (SCA)

As described in Section 1, the domain gap between synthetic data and real scans can primarily be attributed to two factors: the irregularity of room layout and the complexity of object structure. In order to address the disparity, we propose a novel augmentation algorithm called SCA that effectively emulates the irregular layout, and we establish an intermediate domain by integrating objects and scenes from both domains to better generalize object structures from synthetic to real point clouds.

SCA consists of three distinct sub-modules: **Vocabulary Construction**, **Cluster Sampler**, and **Cluster Arrangement** (Fig. 4). In the vocabulary construction module, a vocabulary bank is constructed by gathering objects from scenes, where clusters of points and their labels are established. During training, the cluster sampler module selects clusters from the vocabulary bank that are later incorporated into each scene. Finally, the cluster arrangement module places the selected clusters within the scene based on cluster labels, dimensions, and the room layout.

Vocabulary construction. To construct the vocabulary bank B, given a dataset, we first remove all the points belonging to walls, floors and ceilings. In each scene, we cluster points in the Euclidean space with an out-of-shelf clustering algorithm (e.g., details discussed in Suppl. S5). We define the set of clusters as $C = \{(X_c^{(i)}, Y_c^{(i)})\}_{i=1}^{N_c}$, where N_c is the total number of clusters in the dataset. Next, we create a series of queues $(Q_1, Q_2, \ldots, Q_{N_K})$, where N_K corresponds to the number of classes present in the dataset. Each queue Q_k contains clusters with points belonging to class k. Formally, Q_k can be defined as follows: $Q_k = \{(X, Y) \in C \mid \exists y \in Y : y = k\}$. In simpler terms, Q_k comprises all clusters in C that are associated with points of class k.

In contrast to CINMix [83], where each cluster contains only points with identical semantic labels, our method preserves local context by grouping neighbouring objects in the Euclidean space. For example, a dining table and surrounding chairs in a dining room are grouped into a single cluster, reflecting real-world layouts.

Cluster sampling. To generate training samples in the intermediate domain, we develop a cluster sampling strategy. We first sample $N_{c'}$ classes with replacement and select a cluster from the vocabulary bank for each selected class. To preserve global context, our Global-Context-Aware Sampler (GCAS) prioritizes objects based on their likelihood of appearing in a room (e.g., beds are more likely to be sampled in rooms with nightstands than with toilets). We present a probabilistic framework to determine which object class fits best into an existing scene, considering the diversity and number of objects already in place. P_G **estimate** the most suitable class k' to add based on the scene's current object occurrence. We de-queue an object from the queue of the selected class $Q_{k'}$ and **insert** it into the scene using our cluster arrangement algorithm (details provided later). We then **update** the object occurrence and re-queue the object at the end of $Q_{k'}$, ensuring all objects in class k' are cycled through before reuse, promoting balanced and diverse augmentation. This **estimate**, **insert**, and **update** process is repeated $N_{c'}$ times for each scene during augmentation.

We develop P_G using a multinomial Naive Bayes classifier, which takes the number of objects per class as input and outputs the probability a new object of each class fits into the scene. The classifier is trained on a dataset that reflects the distribution of object classes across various scenes, with training samples adjusted to simulate the addition of new objects. This enables us to assess how changes in object distribution impact scene composition, enhancing realistic augmentation. Further details on P_G are provided in Suppl. S2.

Cluster arrangement. To integrate new object classes into a scene, we propose two strategies. The first arranges objects within the room, ensuring they occupy available space reasonably, while the second focuses on wall-mounted objects like windows and doors.

For the first strategy, clusters are randomly placed in available space, avoiding collisions. We generate N_p candidate centre points. Clusters are placed iteratively at each candidate point, and a point is considered valid if the cluster does not collide with other objects in the scene. Details of the validation algorithm are provided in Suppl. S3.

The second strategy arranges clusters on the wall. We estimate the wall's normals $\hat{\mathcal{N}}_{\text{wall}}$ and sample N_q candidate points for cluster placement. The cluster is translated to the candidate point, and its normal is aligned with the wall's normal by rotating around the z-axis. Similar to the first strategy, each candidate point is tested iteratively. Once the cluster is installed at the valid point, a 3D bounding box is created around the cluster. Wall points inside the bounding box are relabeled as the class of the cluster. Details on the algorithm are provided in Suppl. S3.

SCA in pre-training and self-training stages. In the pre-training stage, we construct a in-domain augmented source scene $\{(X_s^{[s](i)}, Y_s^{[s](i)})\}_{i=1}^{N_s}$ by adding a source bank B_s to source domain. The in-domain augmentation approach in the source domain aims to replicate the complex and cluttered room layouts frequently encountered in real-world scenarios, where furniture and objects are distributed haphazardly rather than organized in a tidy, deliberate manner. During the self-training stage, we construct a target bank B_t by utilizing the point clouds X_t from the target domain along with the pseudo labels Y_p generated using the pre-trained semantic parser F_s. To create an intermediate domain between the source domain D_s and the target domain D_t, we perform cross-domain augmentation. This involves creating cross-domain augmented source scenes $D_s^{[t]} = \{(X_s^{[t](i)}, Y_s^{[t](i)})\}_{i=1}^{N_s}$ by adding B_t to D_s and cross-augmented target scenes $D_t^{[s]} = \{(X_t^{[s](i)}, Y_t^{[s](i)})\}_{i=1}^{N_t}$ by adding B_s to D_t'. With the cross-domain augmentation technique, we not only enhance the diversity of room layouts, as in the pre-training stage but also establish an intermediate domain that bridges the gap between the source and target domains.

3.2. Consistency-Enforced Self-Supervision (CESS)

Given that the pseudo labels Y_p generated by the pre-trained semantic parser F_s may contain incorrect predictions, we introduce the Mean-Teacher framework to mitigate the impact of inaccurate supervision, thus enhancing the reliability of the training process. Furthermore, while our SCA module is designed to preserve both global and local context, the introduction of additional clusters that were not originally part of the scene can potentially disrupt the overall context integrity. To address potential inconsistencies introduced by the augmented clusters, we propose a network head to compensate for such effects.

We introduce the CESS module to tackle these issues. This module is specifically designed to generate stable pseudo labels for self-training on the target domain. By enforcing consistency between augmented and unaugmented point clouds, the CESS module helps to maintain the overall contextual coherence.

Our self-training framework. To address the noise of pseudo labels which can potentially harm the self-training process, we employ a Mean-Teacher paradigm [64] with two models: a teacher and a student model. The two models have identical architectures and are initialized with weights acquired from the pre-training stage.

Following [64], the weights of the teacher model are updated using the exponential moving average (EMA) of the student model as follows:

$$\theta_t \leftarrow \alpha\theta_t + (1-\alpha)\theta_s \tag{1}$$

where θ_t and θ_s denote the weights of the teacher and student models, respectively, and α is a decaying factor con-

trolling the weight updates. In contrast, the student model is supervised by the pseudo labels of the real scenes, the ground-truth labels of the synthetic data, and the predictions of the teacher model. The teacher model can be viewed as an ensemble of the student models at different times in training, which provides more consistent and stable predictions.

Common points consistency. To address the potential disruption to the global context integrity introduced by the SCA module, we introduce a compensation mechanism during training, each sample goes through two branches: the supervised branch and the supervising branch. In the supervised branch, the point clouds are pre-processed using the complete augmentation pipeline including SCA, as described in Section 3.3. However, in the supervising branch, the point clouds are not augmented via the SCA module (Fig. 3). The predictions from the supervising branch maintain the integrity of the global context since it is unaugmented. We enforce the consistency between the augmented and unaugmented point clouds of a scene, by comparing the model outputs of the commonly shared points. We utilize the normalized output Γ_g from the supervising branch to supervise the output Γ_d from the supervised branch using mean squared error (MSE):

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N}\left(\Gamma_d^{(i)} - \Gamma_g^{(i)}\right)^2. \tag{2}$$

3.3. Training and Inference

Pre-processing. To enhance the generalization capability of our model and prevent overfitting, we employ various pre-processing techniques commonly used in 3D semantic segmentation, including the adaptation of Virtual Scan Simulation (VSS) as described in [10]. Additionally, we apply other standard pre-processing techniques including random rotation, scaling, and cropping.

Objective Function. During the pre-training stage, we employ the cross-entropy loss to supervise the predictions of F_s using the ground-truth labels of the synthetic data. In the self-training stage, we introduce the consistency loss $\mathcal{L}_c$ to enforce consistency in the output layers of the original data and augmented data. Therefore, the overall objective function contains the cross-entropy loss $\mathcal{L}_s$ in the source domain, cross-entropy loss $\mathcal{L}_t$ on the target domain, and the consistency loss $\mathcal{L}_c$, as defined as Eq. 3.

$$\min \mathcal{L} = \lambda_s\mathcal{L}_s + \lambda_t\mathcal{L}_t + \lambda_c\mathcal{L}_c, \tag{3}$$

$$\mathcal{L}_s = \frac{1}{N_s}\sum_{i=1}^{N_s} CE(\hat{Y}_s^{[t](i)}, Y_s^{[t](i)}),$$

$$\mathcal{L}_t = \frac{1}{N_t}\sum_{i=1}^{N_t} CE(\hat{Y}_t^{[s](i)}, Y_t^{[s](i)}),$$

where λ_s, λ_t, and λ_c are hyper-parameters that control the relative importance of each loss term.

Inference. We adopt the teacher model F_t for inference since the teacher model is a temporal ensemble of the student model that generates more robust predictions. This ensemble mechanism enhances the model's generalization ability and makes it more resilient to potential noise present in pseudo labels, ultimately leading to improved performance on the target domain.

4. Experiments

4.1. Datasets

The **3D-FRONT** dataset is a large-scale repository of synthetic indoor scenes. It contains 18,968 rooms, each furnished with 3D objects from 3D-FUTURE [13]. The room layouts were created by professional designers and contain 31 scene categories and 34 object classes. We use the 4,995 rooms selected by [10] as the training split for fair comparison. **ScanNet** [7] is a large-scale dataset for real-world indoor 3D scene understanding. It contains 1,613 3D scans and 18 categories, each with dense semantic annotations. The dataset has been used for several tasks in 3D scene understanding, such as object classification and detection. It contains various room types ranging from classroom and office to bedroom and kitchen. **S3DIS** [3] is another significant real-world indoor 3D dataset. The dataset contains annotations for 13 categories with 271 scenes from 6 areas. Unlike ScanNet, the room types in this dataset are relatively limited, which were mainly for office use. We use the fifth area as the validation split and the rest as the training split following the setting in [10].

As a domain adaptation task, we use the synthetic 3D-FRONT [12] as the source domain, and the real-world ScanNet [7] and S3DIS [3] as the target domains in our experiment. Since the classes covered by the three datasets are not identical, we select 11 common classes for the UDA experiments with mapping following [10]. The names of classes are listed in the main results tables (3D-FRONT $\rightarrow$ ScanNet in Table 1 and 3D-FRONT $\rightarrow$ S3DIS in Table 2).

4.2. Results

Since UDA for indoor semantic segmentation is an uncharted research area, there are relatively few methods that match our exact setting. We make our best effort to compare our method with similar settings. Due to the absence of official implementations for SqueezeSegV2 [73], CBST [86], and Noisy Student [74] in the indoor UDA task, we reprint the results from the experiments provided in DODA [10]. For methods with similar but not identical settings (i.e., RPCvSD [63] and CINMix [83]), we report two additional mIoUs, denoted as mIoU-R and mIoU-C, which compute only on the classes covered by their respective experiments, besides reporting their results. Unfortunately, due to the lack of publicly available implementations, we cannot com-

Table 1. Adaptation results from 3D-FRONT to ScanNet in terms of mIoU. ST denotes that the model is self-trained with pseudo labels of ScanNet scenes; PT indicates the model is trained only on source data. The oracles are results trained and tested on the same domain.

Method	mIoU	mIoU-R	mIoU-C	wall	floor	cab.	bed	chair	sofa	table	door	wind.	bksf.	desk
Source Only	29.60	38.97	37.04	60.72	82.42	4.44	12.02	61.76	22.31	38.52	05.72	05.12	19.72	12.84
SqueezeSegV2 [73]	29.77	37.19	36.55	61.85	72.74	02.50	16.89	58.79	16.81	38.19	05.08	03.24	35.68	15.72
RPCvSD [63]	-	43.03	-	28.78	70.90	14.24	35.85	61.87	42.15	47.42	-	-	-	-
CINMix [83]	-	-	48.28	**73.42**	88.07	-	-	62.72	43.56	52.14	07.71	14.20	44.44	-
CBST [86]	37.42	48.42	44.15	60.37	81.39	12.18	30.00	68.86	36.22	49.93	07.05	05.82	43.59	16.25
Noisy student [74]	34.67	46.13	43.47	62.63	86.27	01.45	17.13	69.98	37.58	47.87	06.01	01.66	35.79	15.06
DODA (PT) [10]	40.52	51.45	46.24	67.36	90.24	15.98	39.98	63.11	46.38	48.05	07.63	13.98	33.17	19.86
DODA (ST) [10]	51.42	61.91	55.20	72.71	93.86	27.61	64.31	71.64	55.30	58.43	08.21	24.95	56.49	32.06
Ours (PT)	43.48	53.53	49.24	67.58	90.12	16.30	45.28	67.46	48.09	49.92	12.13	17.24	41.34	22.78
Ours (ST)	**58.19**	**66.68**	**59.34**	71.19	**94.05**	**40.53**	**77.17**	**71.76**	**59.35**	**61.21**	**25.62**	**32.70**	**58.82**	**47.73**
Oracle	75.19	81.27	77.15	83.39	95.11	69.62	81.15	88.95	85.11	71.63	47.67	62.74	82.63	59.05

Table 2. Adaptation results from 3D-FRONT to S3DIS in terms of mIoU. The naming convention of this table follows Table 1.

Method	mIoU	mIoU-C	wall	floor	chair	sofa	table	door	window	bkcase.	ceil.	beam	col.
Source Only	36.72	38.68	67.95	88.68	57.69	04.15	38.96	06.99	00.14	44.90	94.42	00.00	00.00
SqueezeSegV2 [73]	36.50	38.33	65.01	89.95	54.29	06.79	45.75	10.23	01.70	32.93	94.81	00.00	00.00
CINMix [83]	-	51.25	76.00	94.66	66.22	17.79	53.20	21.12	29.84	51.14	-	-	-
CBST [86]	42.47	46.41	71.60	92.07	68.09	03.28	60.45	17.13	00.18	58.45	95.87	00.00	00.00
Noisy student [74]	39.44	42.18	68.84	91.78	65.53	06.65	48.67	02.27	00.00	53.67	96.46	00.00	00.00
DODA (PT) [10]	46.85	52.45	70.96	96.12	68.70	25.47	58.47	17.87	27.65	54.39	95.66	00.00	00.00
DODA (ST) [10]	55.54	64.38	76.23	97.17	**76.89**	63.55	**69.04**	25.76	**38.22**	**68.18**	95.85	00.00	00.00
Ours (PT)	49.47	56.03	73.74	97.03	69.09	26.80	52.46	33.93	34.60	60.61	**95.99**	00.00	00.00
Ours (ST)	**57.11**	**66.63**	**77.49**	**97.43**	76.12	**67.06**	63.12	**53.35**	31.36	67.08	95.19	00.00	00.00
Oracle	62.29	69.85	82.82	96.95	78.16	40.37	78.56	56.91	47.90	77.10	96.29	00.41	29.69

pare the performance of all methods mentioned in Section 2 (e.g., CoSMix [59] and DA-CIL [84]) in our setting.

The main results of **3D-FRONT → ScanNet** are reported in Table 1. Our method outperforms the state-of-the-art approach DODA [10] by a significant 6%. We attribute the performance boost to the effectiveness of SCA, which not only creates robust intermediate domains but also preserves local context. While our method surpasses the SOTA approach across all classes, it's noteworthy that introducing the CESS module leads to a noticeable performance drop specifically in the wall class. The reason might be that F_t often incorrectly classifies irregular walls as doors and CESS implicitly distribute more weights on loss for the building envelope (wall, floor, ceiling) since they always exist in both $\hat{Y}$ and Γ_d. Other classes might exist in Y_c, thus less vulnerable to wrong pseudo labels. Addressing these issues could be a focus for future enhancements in our work.

The results of **3D-FRONT → S3DIS** are reported in Table 2. The proposed method achieves superior results compared to the SOTA [10]. The improvement in this domain adaptation experiment is less significant (+1.57%) because most scenes in 3D-FRONT are home scenes and S3DIS are mostly scenes on campus. With little overlapping in scene types, adding clusters to one another is more likely to break the context integrity as illustrated in Figure 2.

4.3. Ablation Study

In this section, we examine the effectiveness and design choice for SCA and CESS by conducting ablation studies on the 3D-FRONT → ScanNet setting for simplicity.

Table 3. Ablation study of the augmentation algorithm in DODA (TACM) [10], SCA and CESS modules. The first two rows are reprinted from DODA [10].

TACM	SCA	CESS	mIoU
			48.13
✓			51.42
	✓		55.79
	✓	✓	**58.19**

Effectiveness. Table 3 compares the SOTA augmentation technique TACM proposed in DODA [10] with SCA, showing a 4.37% improvement. We evaluate the effectiveness of the SCA and CESS modules by integrating them into the training pipeline incrementally. The experimental results demonstrate that performance improves by an additional 7.66% and 2.40% with each module's inclusion.

SCA: Sampling Policy. We design Uniform Sampler and Context-Reversed Sampler (CRS) to analyse the impact of the sampler and the importance of preserving the global context. Uniform Sampler samples the classes of additional clusters from a uniform distribution $P_U = \frac{1}{N_K}$. Context-Reversed Sampler is designed to break the global context of the augmented scene by adding clusters that are less likely to appear in the scene. To achieve the goal, classes are sampled with the probability P_R in inverse proportion to the probability with P_G. Formally, it is defined as

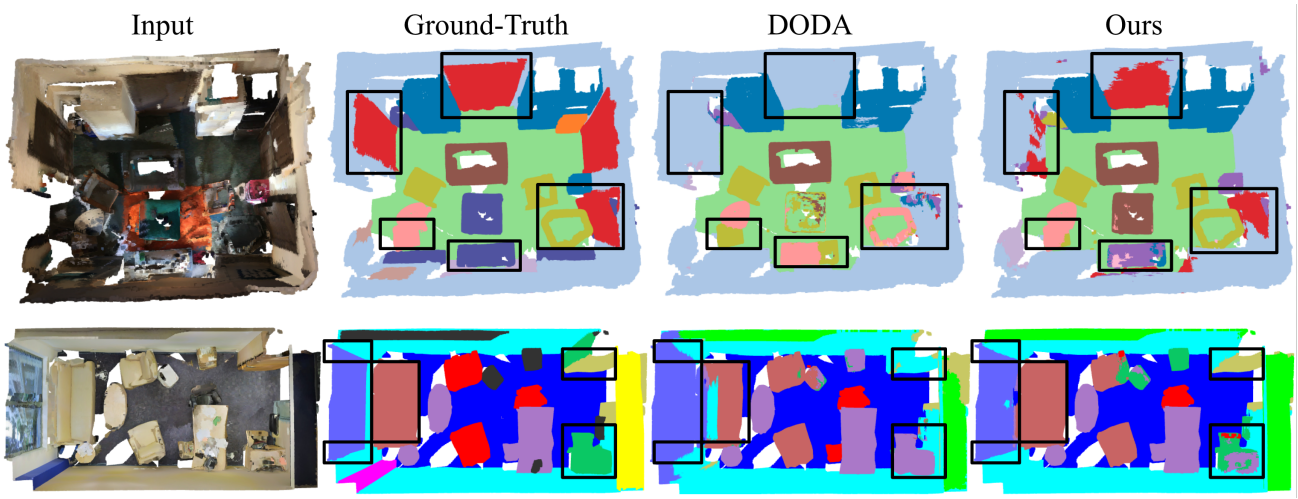

Figure 5. The first and the second rows are the qualitative comparisons on ScanNet and S3DIS respectively. The bounding boxes highlight the parts that our method significantly outperforms the SOTA method DODA [10]. More results can be found in the supplementary material.

$$P_R = \frac{1}{P_G \sum_{k=1}^{N_K} \frac{1}{P_G(k)}} \in [0,1]^K \quad (4)$$

Table 4 shows the mIoU of GCAS is 4.79%/3.97% higher than those of CRS/Uniform Sampler, which shows the help of the global context to a 3D semantic parser.

CESS: Supervision Target. The results of imposing the consistency loss on different domains are shown in rows 1-4 in Table 5. We observe an enhancement in performance when applying the consistency loss to scenarios involving $D_s^{[t]}$, illustrating its effectiveness in mitigating the side effects introduced by SCA. However, its impact deteriorates in setups exclusively using $D_t^{[s]}$, as CESS tends to weigh inaccurately pseudo-labelled points more in this context. This issue is successfully addressed by the implementation of the Mean-Teacher structure (discussed below).

CESS: Mean-Teacher. The integration of the Mean-Teacher integration enhances performance across all three intermediate domain setups (Table 5), with a more significant improvement (+2.26% mIoU) for $D_t^{[s]}$ (adding synthetic objects to real scenes) than $D_s^{[t]}$ (+0.70% mIoU). This difference is attributed to the structure's ability to generate stable supervising signals, as pseudo-labelled points constitute a larger portion of the former setup.

Table 4. Ablation study of the Context-Reversed Sampler (CRS), Uniform Sampler and Global-Context-Aware Sampler (GCAS). The metric is mIoU.

Sampler	Pre-train	Self-train
CRS	40.04	50.93
Uniform	42.63	51.75
GCAS	**43.48**	**55.79**

Table 5. Ablation study of applying the consistency loss on $D_t^{[s]}$, $D_s^{[t]}$, or both. The metric is mIoU.

Domain w/ Consistency Loss	w/o EMA	w/ EMA
$D_t^{[s]}$	55.04	57.30
$D_s^{[t]}$	56.38	57.08
$D_t^{[s]}$ and $D_s^{[t]}$	**57.09**	**58.19**

5. Conclusions

We presented a novel approach for unsupervised domain adaptation from synthetic to real in 3D indoor scenes. The proposed method combined Space and Context-aware Augmentation (SCA) and Consistency-Enforced Self-supervision (CESS) to address challenges in adapting to real-world data. SCA effectively creates intermediate domains by augmenting data to simulate real scene irregularities. The careful consideration of object clustering and arrangement preserves context and enhances training data quality. The CESS module ensures prediction consistency by aligning augmented and unaugmented inputs and using the Mean-Teacher framework. Our experimental results demonstrated the superiority of the proposed approach over existing methods, highlighting its potential for practical applications. Addressing the challenges, such as wall mislabelling, could be a focus for future work.

Acknowledgements

This work was supported in part by the National Science and Technology Council, Taiwan under grants NSTC 111-2221-E-007-106-MY3 and NSTC 113-2634-F-007 002.

References

[1] Inigo Alonso, Luis Riazuelo, Luis Montesano, and Ana C Murillo. 3d-mininet: Learning a 2d representation from point clouds for fast and efficient 3d lidar semantic segmentation. *IEEE Robotics and Automation Letters*, 5(4):5432–5439, 2020. 1

[2] Inigo Alonso, Luis Riazuelo, Luis Montesano, and Ana C Murillo. Domain adaptation in lidar semantic segmentation by aligning class distributions. *arXiv preprint arXiv:2010.12239*, 2020. 1, 16

[3] 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. 6

[4] Lucas Caccia, Herke Van Hoof, Aaron Courville, and Joelle Pineau. Deep generative modeling of lidar data. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5034–5040. IEEE, 2019. 1, 16

[5] Suyi Chen, Hao Xu, Ru Li, Guanghui Liu, Chi-Wing Fu, and Shuaicheng Liu. Sira-pcr: Sim-to-real adaptation for 3d point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14394–14405, October 2023. 1

[6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 14

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 2, 6

[8] Angela Dai and Matthias Nießner. 3dmv: Joint 3d-multi-view prediction for 3d semantic scene segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–468, 2018. 16

[9] Angela Dai, Daniel Ritchie, Martin Bokeloh, Scott Reed, Jürgen Sturm, and Matthias Nießner. Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2018. 16

[10] Runyu Ding, Jihan Yang, Li Jiang, and Xiaojuan Qi. Doda: Data-oriented sim-to-real domain adaptation for 3d semantic segmentation. In *ECCV*, 2022. 1, 2, 4, 6, 7, 8, 14, 15, 16, 17

[11] 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 *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 1996. 14

[12] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 6

[13] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *Int. J. Comput. Vision*, 2021. 6

[14] Wolfgang Fuhl, Gjergji Kasneci, and Enkelejda Kasneci. Teyed: Over 20 million real-world eye images with pupil, eyelid, and iris 2d and 3d segmentations, 2d and 3d landmarks, 3d eyeball, gaze vector, and eye movement types. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 367–375. IEEE, 2021. 1

[15] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(1):2096–2030, 2016. 2

[16] Santiago González Izard, Ramiro Sánchez Torres, Oscar Alonso Plaza, Juan Antonio Juanes Mendez, and Francisco José García-Peñalvo. Nextmed: automatic imaging segmentation, 3d reconstruction, and 3d model visualization platform using augmented and virtual reality. *Sensors*, 20(10):2962, 2020. 1

[17] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 14

[18] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *NIPS*, 2005. 2

[19] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, 2018. 2

[20] Binh-Son Hua, Minh-Khoi Tran, and Sai-Kit Yeung. Pointwise convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 984–993, 2018. 16

[21] Chengjie Huang, Vahdat Abdelzad, Sean Sedwards, and Krzysztof Czarnecki. Soap: Cross-sensor domain adaptation for 3d object detection using stationary object aggregation pseudo-labelling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3352–3361, 2024. 2

[22] Qiangui Huang, Weiyue Wang, and Ulrich Neumann. Recurrent slice networks for 3d segmentation of point clouds. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2626–2635, 2018. 16

[23] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12605–12614, 2020. 2, 16

[24] Peng Jiang and Srikanth Saripalli. Lidarnet: A boundary-aware domain adaptation model for lidar point cloud semantic segmentation. *arXiv preprint arXiv:2003.01174*, 2020. 16

[25] Guoliang Kang, Lu Jiang, Yunchao Wei, Yi Yang, and Alexander Hauptmann. Contrastive adaptation network for single-and multi-source domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):1793–1804, 2020. 2

[26] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902, 2019. 2

[27] Nikhil Keetha, Jay Karhade, Krishna Murthy Jatavallabhula, Gengshan Yang, Sebastian Scherer, Deva Ramanan, and Jonathon Luiten. Splatam: Splat track & map 3d gaussians for dense rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21357–21366, 2024. 1

[28] Abderrazzaq Kharroubi, Rafika Hajji, Roland Billen, and Florent Poux. Classification and integration of massive 3d points clouds in a virtual reality (vr) environment. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42(W17), 2019. 1

[29] Roman Klokov and Victor Lempitsky. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *Proceedings of the IEEE international conference on computer vision*, pages 863–872, 2017. 16

[30] Lingdong Kong, Niamul Quader, and Venice Erin Liong. Conda: Unsupervised domain adaptation for lidar segmentation via regularized domain concatenation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9338–9345. IEEE, 2023. 2

[31] Chun-Chieh Ku, Tsung-Yu Chen, and Shang-Hong Lai. Boosting unsupervised domain adaptation for 3d object detection in point clouds with 2d image semantic information. In *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pages 1–6. IEEE, 2023. 1

[32] Abhijit Kundu, Yin Li, Frank Dellaert, Fuxin Li, and James M Rehg. Joint semantic segmentation and 3d reconstruction from monocular video. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 703–718. Springer, 2014. 16

[33] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4558–4567, 2018. 16

[34] Ferdinand Langer, Andres Milioto, Alexandre Haag, Jens Behley, and Cyrill Stachniss. Domain transfer for semantic segmentation of lidar data using deep neural networks. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8263–8270. IEEE, 2020. 1, 16

[35] Felix Järemo Lawin, Martin Danelljan, Patrik Tosteberg, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Deep projective 3d semantic segmentation. In *Computer Analysis of Images and Patterns: 17th International Conference, CAIP 2017, Ystad, Sweden, August 22-24, 2017, Proceedings, Part I 17*, pages 95–107. Springer, 2017. 16

[36] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, 2013. 2

[37] Seungmin Lee, Dongwan Kim, Namil Kim, and Seong-Gyun Jeong. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In *ICCV*, 2019. 2

[38] Jichang Li, Guanbin Li, and Yizhou Yu. Inter-domain mixup for semi-supervised domain adaptation. *Pattern Recognition*, 146:110023, 2024. 2

[39] Xianzhi Li, Rui Cao, Yidan Feng, Kai Chen, Biqi Yang, Chi-Wing Fu, Yichuan Li, Qi Dou, Yun-Hui Liu, and Pheng-Ann Heng. A sim-to-real object recognition and localization framework for industrial robotic bin picking. *IEEE Robotics and Automation Letters*, 7(2):3961–3968, 2022. 1

[40] Xuyou Li, Shitong Du, Guangchun Li, and Haoyu Li. Integrate point-cloud segmentation with 3d lidar scan-matching for mobile robot localization and mapping. *Sensors*, 20(1):237, 2019. 1

[41] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018. 16

[42] Yixun Liang, Hao He, Shishi Xiao, Hao Lu, and Yingcong Chen. Label name is mantra: Unifying point cloud segmentation across heterogeneous datasets. *arXiv preprint arXiv:2303.10585*, 2023. 2

[43] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 2

[44] Jingpei Lu, Florian Richter, and Michael C. Yip. Markerless camera-to-robot pose estimation via self-supervised sim-to-real transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21296–21306, June 2023. 1

[45] David McClosky, Eugene Charniak, and Mark Johnson. Effective self-training for parsing. In *ACL*, 2006. 2

[46] Pietro Morerio, Jacopo Cavazza, and Vittorio Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. *arXiv preprint arXiv:1711.10288*, 2017. 16

[47] Kazuto Nakashima and Ryo Kurazume. Learning to drop points for lidar scan synthesis. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 222–229. IEEE, 2021. 1, 16

[48] Alexey Nekrasov, Jonas Schult, Or Litany, Bastian Leibe, and Francis Engelmann. Mix3D: Out-of-Context Data Augmentation for 3D Scenes. In *International Conference on 3D Vision (3DV)*, 2021. 2, 3, 15, 16

[49] Munir Oudah, Ali Al-Naji, and Javaan Chahl. Hand gesture recognition based on computer vision: a review of techniques. *journal of Imaging*, 6(8):73, 2020. 1

[50] Zhongyi Pei, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Multi-adversarial domain adaptation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[51] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 16

[52] Mohammad Rezaei, Farnaz Farahanipad, Alex Dillhoff, Ramez Elmasri, and Vassilis Athitsos. Weakly-supervised hand part segmentation from depth images. In *The 14th PErvasive Technologies Related to Assistive Environments Conference*, pages 218–225, 2021. 1

[53] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3577–3586, 2017. 16

[54] Christoph B Rist, Markus Enzweiler, and Dariu M Gavrila. Cross-sensor deep domain adaptation for lidar detection and segmentation. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1535–1542. IEEE, 2019. 1, 16

[55] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Adversarial dropout regularization. *arXiv preprint arXiv:1711.01575*, 2017. 2

[56] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, 2018. 2

[57] Khaled Saleh, Ahmed Abobakr, Mohammed Attia, Julie Iskander, Darius Nahavandi, Mohammed Hossny, and Saeid Nahvandi. Domain adaptation for vehicle detection from bird's eye view lidar point cloud data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 16

[58] Ahmad El Sallab, Ibrahim Sobh, Mohamed Zahran, and Nader Essam. Lidar sensor modeling and data augmentation with gans for autonomous driving. *arXiv preprint arXiv:1905.07290*, 2019. 16

[59] Cristiano Saltori, Fabio Galasso, Giuseppe Fiameni, Nicu Sebe, Elisa Ricci, and Fabio Poiesi. Cosmix: Compositional semantic mix for domain adaptation in 3d lidar segmentation. In *European Conference on Computer Vision*, pages 586–602. Springer, 2022. 2, 3, 7

[60] Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 19847–19878. PMLR, 2022. 2

[61] Rui Shu, Hung H Bui, Hirokazu Narui, and Stefano Ermon. A dirt-t approach to unsupervised domain adaptation. In *ICLR*, 2018. 2

[62] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017. 16

[63] Y. Song, Z. Sun, Y. Wu, Y. Sun, S. Luo, and Q. Li. Learning semantic segmentation on unlabeled real-world indoor point clouds via synthetic data. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022. 6, 7

[64] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017. 5

[65] Mamatha Thota and Georgios Leontidis. Contrastive domain adaptation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 2209–2218, 2021. 2

[66] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 2

[67] Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5495–5504, 2018. 2

[68] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022. 2

[69] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 1

[70] Ximei Wang, Liang Li, Weirui Ye, Mingsheng Long, and Jianmin Wang. Transferable attention for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5345–5352, 2019. 2

[71] 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 Transactions on Graphics (tog)*, 38(5):1–12, 2019. 16

[72] Ze Wang, Sihao Ding, Ying Li, Minming Zhao, Sohini Roychowdhury, Andreas Wallin, Guillermo Sapiro, and Qiang Qiu. Range adaptation for 3d object detection in lidar. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 16

[73] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *ICRA*, 2019. 2, 6, 7, 16

[74] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10687–10698, 2020. 6, 7

[75] Luyu Yang, Yogesh Balaji, Ser-Nam Lim, and Abhinav Shrivastava. Curriculum manager for source selection in multi-source domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 608–624. Springer, 2020. 2

[76] Lei Yang, Yanhong Liu, Jinzhu Peng, and Zize Liang. A novel system for off-line 3d seam extraction and path planning based on point cloud segmentation for arc welding robot. *Robotics and Computer-Integrated Manufacturing*, 64:101929, 2020. 1

[77] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 2

[78] Zhimin Yuan, Wankang Zeng, Yanfei Su, Weiquan Liu, Ming Cheng, Yulan Guo, and Cheng Wang. Density-guided translator boosts synthetic-to-real unsupervised domain adaptive segmentation of 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23303–23312, June 2024. 2

[79] Haojie Zhang, Yongyi Su, Xun Xu, and Kui Jia. Improving the generalization of segmentation foundation model under distribution shift via weakly supervised adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23385–23395, June 2024. 2

[80] Zhanwei Zhang, Minghao Chen, Shuai Xiao, Liang Peng, Hengjia Li, Binbin Lin, Ping Li, Wenxiao Wang, Boxi Wu, and Deng Cai. Pseudo label refinery for unsupervised domain adaptation on cross-dataset 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15291–15300, 2024. 2

[81] Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pages 7523–7532. PMLR, 2019. 16

[82] Sicheng Zhao, Yezhen Wang, Bo Li, Bichen Wu, Yang Gao, Pengfei Xu, Trevor Darrell, and Kurt Keutzer. epointda: An end-to-end simulation-to-real domain adaptation framework for lidar point cloud segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3500–3509, 2021. 2

[83] Yuyang Zhao, Na Zhao, and Gim Hee Lee. Synthetic-to-real domain generalized semantic segmentation for 3d indoor point clouds, 2022. 2, 3, 4, 6, 7, 15, 16

[84] Ziyuan Zhao, Mingxi Xu, Peisheng Qian, Ramanpreet Singh Pahwa, and Richard Chang. Da-cil: Towards domain adaptive class-incremental 3d object detection. *arXiv preprint arXiv:2212.02057*, 2022. 3, 7

[85] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 14

[86] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018. 6, 7