Transfer Learning from Synthetic to Real LiDAR Point Cloud for Semantic Segmentation

Aoran Xiao[1], **Jiaxing Huang**[1], **Dayan Guan**[1], **Fangneng Zhan**[2], **Shijian Lu**[1*]

[1] Singtel Cognitive and Artificial Intelligence Lab for Enterprises, Nanyang Technological University
[2] S-lab, Nanyang Technological University
{aoran.xiao, jiaxing.huang, dayan.guan, fnzhan, shijian.lu}@ntu.edu.sg

Abstract

Knowledge transfer from synthetic to real data has been widely studied to mitigate data annotation constraints in various computer vision tasks such as semantic segmentation. However, the study focused on 2D images and its counterpart in 3D point clouds segmentation lags far behind due to the lack of large-scale synthetic datasets and effective transfer methods. We address this issue by collecting SynLiDAR, a large-scale synthetic LiDAR dataset that contains point-wise annotated point clouds with accurate geometric shapes and comprehensive semantic classes. SynLiDAR was collected from multiple virtual environments with rich scenes and layouts which consists of over 19 billion points of 32 semantic classes. In addition, we design PCT, a novel point cloud translator that effectively mitigates the gap between synthetic and real point clouds. Specifically, we decompose the synthetic-to-real gap into an appearance component and a sparsity component and handle them separately which improves the point cloud translation greatly. We conducted extensive experiments over three transfer learning setups including data augmentation, semi-supervised domain adaptation and unsupervised domain adaptation. Extensive experiments show that SynLiDAR provides a high-quality data source for studying 3D transfer and the proposed PCT achieves superior point cloud translation consistently across the three setups. The dataset is available at https://github.com/xiaoaoran/SynLiDAR.

Introduction

Semantic segmentation of LiDAR sequential point cloud is critical in various scene perception tasks and it has attracted increasing attention from both industry and academia (Behley et al. 2019; Milioto et al. 2019; Pan et al. 2020; Hu et al. 2020; Tang et al. 2020) in recent years. However, training effective segmentation models requires large amounts of annotated point cloud, which are prohibitively time-consuming to collect due to the view change of 3D data and visual inconsistency between LiDAR point cloud and physical world. This can be observed by the small size of existing *real* LiDAR sequential datasets as listed in Table 1.

Inspired by the great success of transfer learning from synthetic to real data in two-dimensional (2D) field (Ros

et al. 2016; Armeni et al. 2016; Gaidon et al. 2016), one possible way to mitigate the data annotation constraint is to leverage synthetic point cloud data that can be collected and annotated automatically by computer engines. However, collecting large-scale synthetic LiDAR sequential point cloud is a nontrivial task which involves a large number of virtual scenes and objects as well as complicated point generation processes. In addition, most existing transfer learning methods (Sun et al. 2019; Saito et al. 2019a; Guan et al. 2021; Huang et al. 2021a,b,c) focus on 2D images which do not work for 3D point cloud. To the best of our knowledge, few researches tackle the challenge of synthetic-to-real transfer of point cloud of nature scenes, largely due to the lack of large-scale synthetic data with accurate geometries and rich semantic annotations.

We address the said issues by creating SynLiDAR, a large-scale LiDAR sequential point cloud dataset for facilitating the research of synthetic-to-real transfer of 3D point cloud data. We collected SynLiDAR from multiple virtual environments that were constructed by professional 3D generalists with advanced graphic tools. Each virtual environment contains configurable object models that are similar to real-world data in both geometry and layout. The dataset is ideal for the study of synthetic-to-real transfer as it consists of comprehensive and diverse point semantics (32 semantic classes with over 19 billion point-wise annotated points) and its collected points are also highly accurate in geometry.

In addition, we designed PCT, a point-cloud translator for mitigating domain gaps between synthetic and real point cloud as illustrated in Fig. 1. The design of PCT is inspired by the observations that point clouds can be viewed as discrete samplings of continuous 3D geometric environments where the domain gaps between synthetic and real point clouds come from either appearance differences (due to environments variations) or sparsity differences (due to sampling variation by sensors). We hence disentangle the domain gap into an appearance component and a sparsity component, and design an appearance translation module (ATM) and a sparsity translation module (STM) to handle the two gap components separately. Specifically, ATM first up-samples synthetic point cloud and translates it to have similar appearance as real point cloud. STM then extracts sparsity features from real point cloud and fuses it with the ATM output to translate synthetic point cloud to have real

*Corresponding author.

(a) A synthetic scene of SynLiDAR.

(b) *Synthetic* point cloud captured in (a).

(c) *Real* point cloud of SemanticKITTI.

(d) Cars

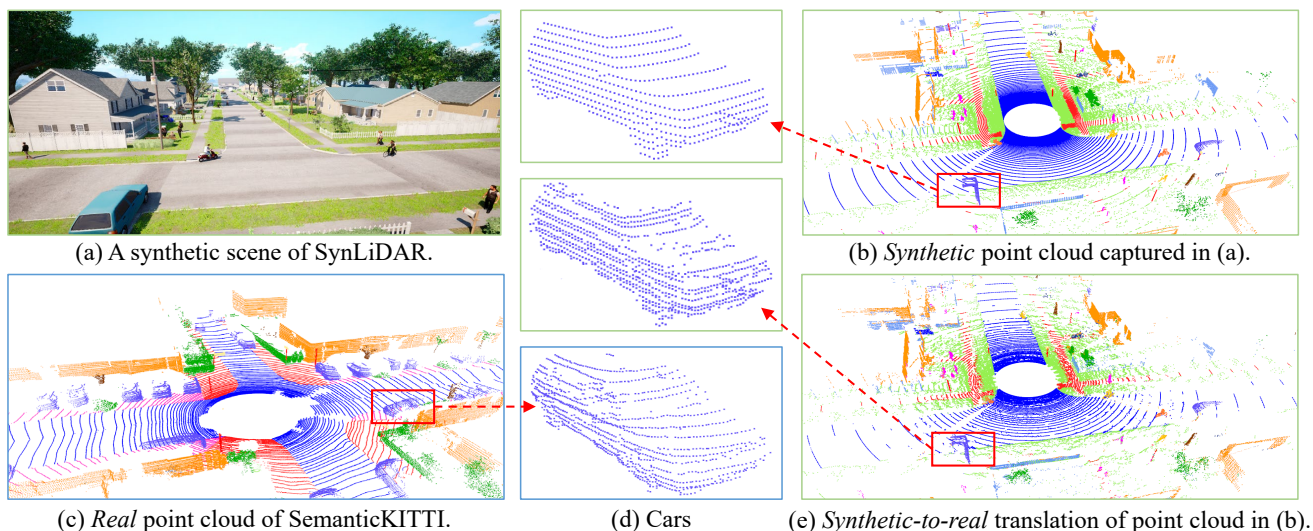(e) *Synthetic-to-real* translation of point cloud in (b).

Figure 1: We create SynLiDAR, a large-scale multiple-class synthetic LiDAR point cloud dataset as illustrated in (b). SynLiDAR contains over 19 billion annotated points of 32 semantic classes which was collected by constructing multiple virtual environments and 3D object models as shown in (a). To make synthetic point cloud more useful for handling real-world LiDAR point cloud as shown in (c), we design a point cloud translator (PCT) that translates synthetic point cloud by decomposing the domain gap into an appearance component and a sparsity component. The translated data in (e) has a closer distribution as real point cloud and is more effective in processing real point cloud. The close-up views in (d) show the translation effects.

sparsity. To the best of our knowledge, PCT is the first translation method for LiDAR point clouds in natural scenes.

The contribution of this work can be summarized in three aspects as listed:

- We create SynLiDAR, a large-scale synthetic LiDAR sequential point cloud dataset that has rich semantic classes and a large number of points with accurate point-wise annotations. SynLiDAR will lay a strong foundation for the study of the under-explored synthetic-to-real transfer in LiDAR point cloud segmentation.

- We examine the major underlying factors of the domain gap between synthetic and real point clouds, and design PCT, a pioneer LiDAR point cloud translator that can transform synthetic point clouds to have similar features and distributions as real point clouds and accordingly mitigate the domain gap effectively.

- We design three experimental setups for the study of synthetic-to-real point cloud transfer: data augmentation (DA), semi-supervised domain adaptation (SSDA) and unsupervised domain adaptations (UDA). We conducted extensive experiments under the three setups which will form valuable bases for the future investigation of synthetic-to-real point cloud transfer.

Related Works
Semantic Segmentation of Point Cloud

3D deep learning has attracted increasing attention, which is important for different LiDAR perception tasks including semantic segmentation. Different approaches have been proposed to segment LiDAR point clouds, including projection-based methods (Wu et al. 2018; Milioto et al. 2019; Xiao

et al. 2021), point-based methods (Qi et al. 2017a,b; Hu et al. 2020), sparse convolutional methods (Graham, Engelcke, and Van Der Maaten 2018; Choy, Gwak, and Savarese 2019a), customized 3D convolutional methods (Thomas et al. 2019; Zhu et al. 2020), etc.

LiDAR Sequential Point Cloud Datasets

LiDAR *sequential* point clouds provide point cloud scans, each containing sparse and incomplete points collected in a sweep by LiDAR sensors. Several real world datasets, including SemanticKITTI (Behley et al. 2019), SemanticPOSS (Pan et al. 2020) and nuScenes-Lidarse (Caesar et al. 2020), have been proposed recently and promote the developments of LiDAR point cloud segmentation researches. However, labeling point-wise semantic annotations is prohibitively time-consuming for LiDAR sequential point clouds. Therefore, existing real point cloud datasets have very limited data sizes as listed in Table 1.

Inspired by the success of 2D synthetic image datasets (Richter et al. 2016), a few pioneer studies (Wu et al. 2018; Hurl, Czarnecki, and Waslander 2019) have explored to collect synthetic point cloud data from GTA-V games. However, 3D meshes in GTA-V games are not accurate and GTA-V games provide only two object classes *Car* and *Pedestrian* (Wu et al. 2018). Its collected synthetic data are thus insufficient for studying fine-grained LiDAR point cloud segmentation. We instead construct a wide range of realistic virtual environments and object models by leveraging graphic tools and professionals. The synthetic point clouds within the SynLiDAR thus capture much more accurate geometries and the rich diversity of semantic labels as in natural scenes.

dataset	format	#scans	#points	#classes	annotation	type
SemanticKITTI (Behley et al. 2019)	point	43,552	4,549M	25	point-wise	real
SemanticPOSS (Pan et al. 2020)	point	2,988	216M	14	point-wise	real
nuScenes-Lidarseg (Caesar et al. 2020)	point	40,000	1,400M	32	point-wise	real
GTA-LiDAR (Wu et al. 2018)	image	121,087	-	2	pixel-wise	synthetic
PreSIL (Hurl, Czarnecki, and Waslander 2019)	point	51,074	3,135M	2	point-wise	synthetic
SynLiDAR (ours)	point	198,396	19,482M	32	point-wise	synthetic

Table 1: Overview of outdoor LiDAR sequential point cloud datasets with semantic annotations: #scans: Number of scans for the datasets; #points: Number of points in millions (M); #classes: Number of semantic classes.

Transfer Learning of Point Cloud

Transfer learning aims to transfer the knowledge from the source domain to the target domain. It is an important tool to solve the inefficient training data problem (Tan et al. 2018). This paper discusses three important transfer learning tasks of point cloud: DA combines multiple labeled datasets for training to reach better performances than training on each single one (Li et al. 2020; Chen et al. 2020); SSDA exploits the knowledge from the source data with annotations and use a certain number of unlabeled examples and a few labeled ones from the target domain to learn a target model; UDA instead uses annotated source data and target data without annotations to learn the target model (Qin et al. 2019; Saleh et al. 2019; Yang et al. 2021). Several pioneer works (Wu et al. 2019; Zhao et al. 2020; Yi, Gong, and Funkhouser 2021) have been proposed for the research of UDA problem in the LiDAR segmentation task. Instead, PCT mitigates domain gap problem in the input space and is effective for different kinds of transfer learning setups.

Domain Translation of Point Cloud

Domain Translation aims to learn meaningful mapping across domains. It is well developed for 2D images between paired domains (Isola et al. 2017), unpaired domains (Zhu et al. 2017a), multiple modalities (Zhu et al. 2017b), etc. For 3D data, some attempt has been reported for translation from images to depth (Liu, Shen, and Lin 2015), from point cloud to depth (Roveri et al. 2018), from point cloud to images (Dai et al. 2020), etc. However, existing generative methods (Li et al. 2019; Xie et al. 2021) mainly focus on 3D objects while the translation between LiDAR point clouds in scenes is largely neglected. We address this challenge for mitigating the gap between synthetic and real point clouds.

The SynLiDAR Dataset

SynLiDAR is collected from multiple realistic virtual scenes constructed by professional 3D generalists using the Unreal Engine 4 platform (UE4 2014) (as shown in Fig. 1 a). These virtual scenes include different types of outdoor environments such as cities, towns, harbour, etc, covering large area of virtual areas. They are constituted by a large number of physically accurate object models that are produced by expert modelers with 3D-Max software, to ensure the high quality of synthetic data. Specifically, accurate coordinates and precise point-wise annotations of point cloud are collected automatically from these virtual environments. Note
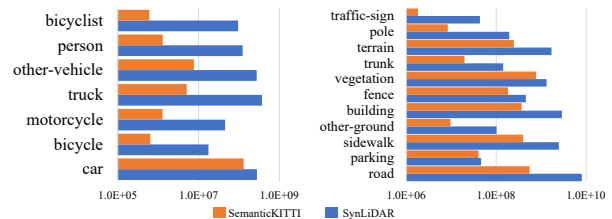

Figure 2: The numbers of annotated points (x-axis) per class (y-axis) for SemanticKITTI and SynLiDAR. Left: *Thing* classes; Right: *Stuff* classes.

that SynLiDAR can be easily expanded by including new virtual scenes, 3D objects of new classes, etc.

Another important attribute of LiDAR point cloud is *intensity*, which is challenging to simulate due to the complicated signal transmission, propagation and reception processes in real environments (Wu et al. 2019). In SynLiDAR, we address this issue by training a rendering model that learns from real LiDAR point cloud and predicts intensity values for SynLiDAR. Specifically, we train an intensity prediction model by using MinkowskiNet (Choy, Gwak, and Savarese 2019b), where coordinates and semantic labels of cloud points in SemanticKITTI (Behley et al. 2019) are used as inputs and the point intensity (SemanticKITTI provided) is used as references.

SynLiDAR has 13 LiDAR point cloud sequences with point-wise annotations. It has 198,396 scans of point cloud with 19 billion points in total, where each scan has around 98,000 points on average. Precise point-wise annotations of 32 semantic classes are provided for fine-grained scene understanding. Tab. 1 shows that SynLiDAR outweighs the existing LiDAR point cloud dataset of semantic segmentation in both point numbers and semantic classes and Fig. 2 compares point numbers for categories of *thing* (countable foreground classes) and *stuff* (uncountable background classes) in SynLiDAR and SemanticKITTI (the largest real point cloud dataset to the best of our knowledge), indicating that SynLiDAR is a truly *large-scale* point cloud dataset and an ideal data source for transfer learning researches on synthetic-to-real point cloud.
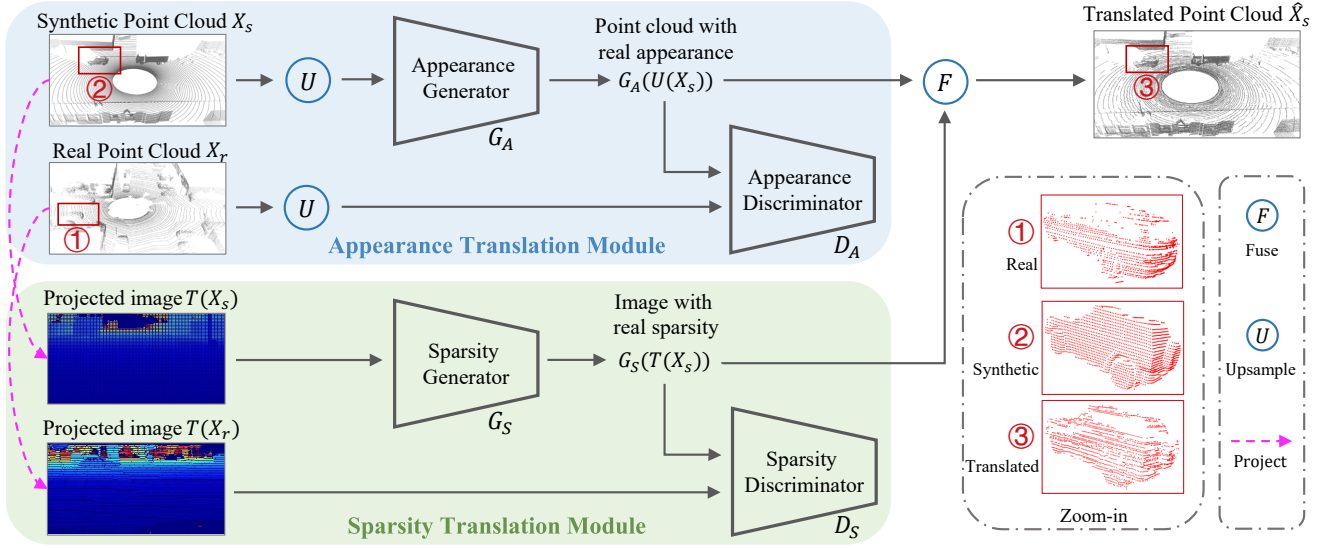
Figure 3: The proposed PCT disentangles point-cloud translation into appearance translation and sparsity translation tasks. Given synthetic point cloud, the appearance translation first learns to reconstruct dense point cloud that have similar appearance as real point cloud. The sparsity translation then learns real sparsity distribution in 2D space and fuses it with the reconstructed point cloud in 3D space. The final translation has similar appearance and sparsity as real point cloud as illustrated.

Point Cloud Translation

Similar to most synthetic data, point clouds in SynLiDAR have a clear domain gap with real LiDAR data, and the model trained using SynLiDAR usually experiences clear performance drops while applied to real point clouds. This paper proposes PCT, the first translator of LiDAR data to mitigate the domain gap for the challenging scene semantic segmentation task, as illustrated in this section.

An intuitive idea to mitigate the domain gap is to employ existing 3D generative models (Achlioptas et al. 2018; Li et al. 2019; Huang et al. 2020) to translate synthetic data to have real data distributions. However, these models are designed for 3D objects with uniformly distributed points, and standard supervisions like Chamfer loss or Earth Mover's Distance (EMD) (Fan, Su, and Guibas 2017) fail to regularize LiDAR data distribution of 3D scenes directly. As a result, few studies have attempted to address the problem of LiDAR point cloud translation.

We observed that point clouds are discrete samplings of the continuous geometric world, hence the synthetic-to-real gaps could be disentangled into two components: The appearance component reflects the differences between synthetic and real continuous environments, and the sparsity component shows differences in sampling patterns introduced by different LiDAR sensors. The proposed *PCT* mitigates these two components separately and its translated synthetic point clouds have both similar appearance and sparsity as targeting real-world data.

The pipeline of PCT: Given synthetic point clouds $X_s \in \mathbb{R}^{N_s \times 3}$ and real point clouds $X_r \in \mathbb{R}^{N_r \times 3}$, we aim to translate X_s into $\hat{X}_{s \to r} \in \mathbb{R}^{\hat{N}_s \times 3}$ that have similar appearance and sparsity as X_r. Firstly, the ATM up-samples X_s into $U(X_s)$, and translates it to have certain real appearance by a

generator G_A as $X'_s = G_A(U(X_s))$. Then the STM projects (T) point cloud into images and extracts sparsity features of real point cloud into synthetic data by another generator G_S, *i.e.* $X''_s = G_S(T(X_s))$. Finally, translated point cloud is generated by fusing X'_s and X''_s as $\hat{X}_{s \to r} = F(X'_s, X''_s)$. More details of the two translation modules are provided in the ensuing two subsections.

ATM aims to translate synthetic point cloud to have real appearance and we realize it through a 3D generative adversarial network that consists of a generator G_A and a discriminator D_A, and train them in an adversarial manner. In this stage, we first up-sample both synthetic and real point cloud so as to eliminate the effect of domain-specific sparsity features. The generator aims to produce intermediate representation X'_s (from the up-sampled synthetic data) to have real appearance to fool the discriminator, while the discriminator learns to distinguish X'_s from $U(X_r)$. Specifically, The adversarial learning loss can be formulated as follows:

$$L_A^{adv} = \log(D_A(-G_A(U(X_s)) \log(G_A(U(X_s))))) + \log(1 - D_A(U(X_r) \log(U(X_r)))) \quad (1)$$

We introduce EMD for generator to keep geometries of X_s and X'_s

$$L_A^{emd}(X_s, X'_s) = \sum_{x \in X_s} ||x - \phi(x)||_2 \quad (2)$$

where $\phi : X_s \to X'_s$ is a bijection. The objective function of the ATM can be formulated as:

$$L_A(G_A, D_A) = \arg \min_{G_A} \max_{D_A} (\lambda_A^{adv} L_A^{adv} + \lambda_A^{emd} L_A^{emd}) \quad (3)$$

method	car	bi.cle	mt.cle	truck	oth-v.	pers.	bi.clst	mt.clst	road	parki.	sidew.	oth-g.	build.	fence	veget.	trunk	terra.	pole	traf.	mIoU
baseline	95.7	25.0	57.0	62.1	46.4	63.4	77.3	0.0	93.0	47.9	80.5	2.2	89.7	58.6	89.5	66.5	78.0	64.6	50.1	60.3
Jittering (Qi et al. 2017a)	95.7	27.8	56.2	66.0	45.8	65.3	82.8	0.0	93.0	48.2	79.9	2.5	89.7	62.9	88.9	64.0	77.0	64.8	51.0	61.2
Dropout (Srivastava et al. 2014)	96.0	28.5	57.1	65.1	46.4	64.1	83.6	0.1	93.5	47.6	80.1	2.3	89.3	61.9	90.1	66.9	78.8	65.8	49.1	61.4
PointAug (Li et al. 2020)	95.9	29.2	70.0	76.3	50.0	67.0	84.4	2.4	93.8	48.1	81.2	4.6	89.8	58.4	87.5	65.4	72.7	62.4	50.5	62.6
+SynLiDAR	95.9	33.0	62.8	78.9	50.2	71.4	83.5	0.7	92.3	52.8	79.9	0.1	89.8	59.5	86.3	65.4	72.8	63.6	48.9	62.5
PCT	96.3	38.7	73.4	82.9	56.1	71.1	85.3	1.6	94.1	54.3	81.6	1.3	89.5	59.6	87.8	66.9	73.6	65.4	50.5	64.7

Table 2: Data Augmentation experiments on SemanticKITTI: Combining the training data of SynLiDAR and SemanticKITTI trains more accurate semantic segmentation models. PCT mitigates the domain gap effectively and combining the PCT-translated SynLiDAR with SemanticKITTI further improves the segmentation.

method	pers.	rider	car	trunk	plants	traf.	pole	garb.	buil.	cone.	fence	bike	grou.	mIoU
baseline	55.6	45.1	66.9	44.4	73.9	45.4	41.6	14.5	76.1	7.9	57.0	54.1	75.3	50.6
Jittering (Qi et al. 2017a)	55.2	48.7	65.1	45.5	75.2	45.9	40.9	18.1	76.4	15.1	57.1	56.4	75.0	51.9
Dropout (Srivastava et al. 2014)	56.9	56.7	67.8	43.3	75.6	40.3	30.7	26.8	75.7	17.7	57.3	55.6	78.6	52.5
PointAug (Li et al. 2020)	62.3	60.7	69.6	39.3	76.0	41.4	33.8	24.1	78.0	13.7	62.2	56.5	79.2	53.6
+SynLiDAR	57.6	59.3	61.1	37.1	76.1	32.7	40.9	34.7	72.7	37.7	57.6	43.3	81.2	53.2
PCT	57.3	61.4	65.8	36.2	77.4	42.5	42.1	49.2	74.5	32.4	55.8	48.9	81.8	55.8

Table 3: Data Augmentation experiments on SemanticPOSS: Combining the training data of SynLiDAR and SemanticPOSS trains more accurate semantic segmentation models. PCT mitigates the domain gap effectively and combining the PCT-translated SynLiDAR with SemanticPOSS further improves the segmentation.

STM aims to transfer sparsity information from real point cloud to synthetic point cloud. However, existing supervisions (Fan, Su, and Guibas 2017) such as Chamfer loss and EMD loss cannot capture sparsity information well as they lack sparsity regulation. To address this problem, we propose to first learn sparsity information in 2D space and then fuse it back into 3D space. Specifically, we first project X_s and X_r into depth images (as $T(X_s)$ and $T(X_r)$ respectively), and then employ an image-to-image translation model to translate $T(X_s)$ to have similar sparsity as $T(X_r)$. The translation network is also GAN-based with a generator G_S and a discriminator D_S. The adversarial learning loss can be formulated as follows

$$L_S^{adv} = \log(D_S(-G_S(T(X_s)) \log(G_S(T(X_s))))) \\ + \log(1 - D_S(T(X_r) \log(T(X_r)))) \quad (4)$$

To preserve the geometry information during the translation, we further include a geometry consistency loss to ensure that the translated depth image preserves similar geometry as the original image:

$$L_S^{geo} = ||\overline{T(X_s)} - \overline{G_S(T(X_s))}||_2 \\ + ||\overline{T(X_r)} - \overline{G_S(T(X_r))}||_2 \quad (5)$$

where $\overline{A} - \overline{B}$ means that the distance is computed for pixels existing in both A and B only. The overall objective of the sparsity translation module can be formulated by:

$$L_S(G_S, D_S) = \arg \min_{G_S} \max_{D_S} (\lambda_S^{adv} L_S^{adv} + \\ \lambda_S^{geo} L_S^{geo}) \quad (6)$$

Finally, the translated image $G_S(T(X_s))$ with real sparsity information are projected back into 3D space for fusion. Specifically, $G_S(T(X_s))$ are used for guidance to drop out points in X'_s. The semantic labels of the translated point cloud are assigned according to labels of neighboring points in the original point cloud.

Experiments

We evaluate how SynLiDAR benefits semantic segmentation over multiple real point cloud datasets and how PCT mitigates domain gaps between SynLiDAR and real point cloud data. We conducted experiments on three transfer setups including DA, SSDA, and UDA as introduced in section 2.3. All experiments were conducted by using state-of-the-art 3D semantic segmentation network MinkowskiNet (Choy, Gwak, and Savarese 2019b).

Datasets and Implementation Details

We conduct experiments over two real-world point cloud datasets. The first is *SemanticKITTI* (Behley et al. 2019) that consists of 43,552 scans of annotated sequential LiDAR point cloud with 19 semantic classes. It is the largest real-world sequential LiDAR point cloud dataset for semantic segmentation to the best of our knowledge. We follow the commonly-used protocol that splits sequences 00-07, 09-10 for training and sequence 08 for validation. The second is *SemanticPOSS* (Pan et al. 2020) which is collected in a university campus. It consists of 2,988 annotated point cloud scans of 14 semantic classes. We follow the benchmark setting by using sequence 03 for validation and the rest for

Method	car	bi.cle	mt.cle	truck	oth-v.	pers.	bi.clst	mt.clst	road	parki.	sidew.	oth-g.	build.	fence	veget.	trunk	terra.	pole	traf.	mIoU
S+T	56.2	3.0	15.1	1.0	5.0	20.2	42.1	2.8	52.1	0.7	19.8	0.0	41.3	5.8	62.1	34.0	42.0	24.6	1.4	22.6
MMD (Tzeng et al. 2014)	56.4	3.3	13.3	1.5	6.1	21.4	34.6	1.6	54.3	0.4	21.4	0.0	50.2	5.8	61.2	37.0	44.9	31.6	2.2	23.5
MME (Saito et al. 2019b)	51.0	5.6	13.1	1.3	7.3	15.1	54.4	4.4	43.1	0.2	28.3	0.0	60.7	13.3	66.1	30.1	39.9	24.8	6.6	24.5
APE (Kim and Kim 2020)	58.6	6.2	16.6	3.1	11.3	14.2	35.8	3.7	61.5	1.7	30.3	0.0	54.7	15.4	64.6	20.0	45.5	23.9	9.1	25.1
PCT	56.0	7.0	17.1	2.8	9.9	23.7	43.7	5.6	55.3	0.8	22.9	0.0	50.1	8.4	65.3	23.1	43.5	28.8	7.5	24.8
APE + PCT	58.1	7.3	17.8	2.6	13.9	24.7	46.5	5.1	60.5	1.9	31.3	0.0	56.8	14.6	67.9	23.7	44.3	26.1	9.3	27.0

Table 4: Experiments on semi-supervised domain adaptation with SynLiDAR (as source) and SemanticKITTI (as target): PCT translates SynLiDAR and mitigates domain gaps in the input space effectively. It is complementary to APE and combining them outperforms the baseline SynLiDAR + SemanticKITTI (*i.e.*, S+T) by large margins.

Method	pers.	rider	car	trunk	plants	traf.	pole	garb.	buil.	cone.	fence	bike	grou.	mIoU
S+T	25.2	36.1	18.2	12.8	58.6	1.7	30.5	5.6	25.7	3.0	12.0	10.6	75.6	24.3
MMD (Tzeng et al. 2014)	25.5	35.7	28.9	6.7	64.3	1.7	23.2	5.6	53.3	3.3	30.2	13.9	70.4	27.9
MME (Saito et al. 2019b)	33.2	40.2	25.0	11.0	61.9	0.4	31.2	7.3	56.1	5.7	37.1	6.7	71.2	29.8
APE (Kim and Kim 2020)	34.3	40.1	21.5	16.3	62.6	0.9	31.1	2.3	55.9	13.3	34.3	9.6	71.6	30.3
PCT	25.8	36.8	27.8	11.3	62.2	1.9	31.2	5.2	58.7	2.6	34.3	8.5	68.7	28.8
APE + PCT	34.7	36.3	27.2	15.8	62.9	0.8	31.6	8.7	62.3	9.8	35.1	9.3	70.9	31.2

Table 5: Experiments on semi-supervised domain adaptation with SynLiDAR (as source) and SemanticPOSS (as target): PCT translates SynLiDAR and mitigates domain gaps in the input space effectively. It is complementary to APE and combining them outperforms the baseline SynLiDAR + SemanticPOSS (*i.e.*, S+T) by large margins.

training. We *ignore* extra classes of SynLiDAR by mapping them as 'unlabeled' for each real dataset. Mean Intersection of Union (mIoU) is used as the evaluation metric.

For point cloud translation, parameters λ_A^{adv}, λ_A^{emd}, λ_S^{adv}, λ_S^{geo} are set at 0.01, 1, 5, 1, respectively. For semantic segmentation by MinkowskiNet, the maximum point number of each scan is 80,000 for SemanticKITTI and 50,000 for SemanticPOSS; The voxel size is 0.05 and we use coordinates and intensity of point cloud as input features.

Experiments on Data Augmentation

We first evaluate how SynLiDAR augments real point cloud data as compared with state-of-the-art data augmentation methods as shown in Tables 2 and 3. We can observe that incorporating SynLiDAR helps train better models for both SemanticKITTI (+2.2%) and SemanticPOSS (+2.6%). It also shows that SynLiDAR shares good similarity with real-world dataset and provides a high-quality data source for transfer learning of LiDAR point cloud.

We also evaluate how PCT mitigates the domain gap and improves data augmentation. Experiments show that including PCT-translated SynLiDAR in training improves the mIoU by 2.2% and 2.6%, respectively, for SemanticKITTI and SemanticPOSS. The mIoU improvements clearly demonstrate the effectiveness of PCT in mitigating domain gap between SynLiDAR and the two real datasets.

We further evaluate by combining SynLiDAR and PCT-translated SynLiDAR with different portions of Semantic-POSS, aiming to examine how SynLiDAR could alleviate data collection and annotation efforts. Fig. 4 shows experiment results. We can see that incorporating SynLiDAR con-
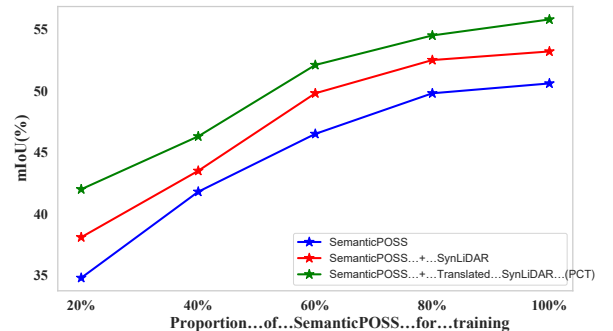

Figure 4: SynLiDAR can effectively augment real-world LiDAR point cloud (SemanticPOSS) in a point cloud segmentation task. The PCT translated SynLiDAR further improves the augmentation consistently by large margins.

sistently improves the segmentation under different portions of SemanticPOSS, and it can save up to 40% SemanticPOSS without sacrificing segmentation performance. In addition, including the PCT-translated SynLiDAR further improves the segmentation consistently under similar setups.

Experiments on SSDA

In this subsection we evaluate the effectiveness of PCT in another setup of semi-supervised domain adaptation (SSDA) with SynLiDAR (as source) and real datasets (as target). We follow the setting of (Saito et al. 2019a) with three parts of training data, *i.e.* labeled source samples, unlabeled tar-

Method	car	bi.cle	mt.cle	truck	oth-v.	pers.	bi.clst	mt.clst	road	parki.	sidew.	oth-g.	build.	fence	veget.	trunk	terra.	pole	traf.	mIoU
Source-Only	42.0	5.0	4.8	0.4	2.5	12.4	43.3	1.8	48.7	4.5	31.0	0.0	18.6	11.5	60.2	30.0	48.3	19.3	3.0	20.4
ADDA (Tzeng et al. 2017)	52.5	4.5	11.9	0.3	3.9	9.4	27.9	0.5	52.8	4.9	27.4	0.0	61.0	17.0	57.4	34.5	42.9	23.2	4.5	22.8
Ent-Min (Vu et al. 2019)	58.3	5.1	14.3	0.6	1.8	14.3	44.5	0.5	50.4	4.3	34.8	0.0	48.3	19.7	67.5	34.8	52.0	33.0	6.1	25.5
ST (Zou et al. 2019)	62.0	5.0	12.4	1.3	9.2	16.7	44.2	0.4	53.0	2.5	28.4	0.0	57.1	18.7	69.8	35.0	48.7	32.5	6.9	26.5
PCT	53.4	5.4	7.4	0.8	10.9	12.0	43.2	0.3	50.8	3.7	29.4	0.0	48.0	10.4	68.2	33.1	40.0	29.5	6.9	23.9
ST+PCT	70.8	7.3	13.1	1.9	8.4	12.6	44.0	0.6	56.4	4.5	31.8	0.0	66.7	23.7	73.3	34.6	48.4	39.4	11.7	28.9

Table 6: Experiments on unsupervised domain adaptation with SynLiDAR (as source) and SemanticKITTI (as target): PCT translates SynLiDAR and mitigates domain gaps in the input space effectively. It complements ST and combining them outperforms the baseline (*i.e.*, source-only) by large margins.

Method	pers.	rider	car	trunk	plants	traf.	pole	garb.	buil.	cone.	fence	bike	grou.	mIoU
Source-Only	3.7	25.1	12.0	10.8	53.4	0.0	19.4	12.9	49.1	3.1	20.3	0.0	59.6	20.1
ADDA (Tzeng et al. 2017)	27.5	35.1	18.8	12.4	53.4	2.8	27.0	12.2	64.7	1.3	6.3	6.8	55.3	24.9
Ent-Min (Vu et al. 2019)	24.2	32.2	21.4	18.9	61.0	2.5	36.3	8.3	56.7	3.1	5.3	4.8	57.1	25.5
ST (Zou et al. 2019)	23.5	31.8	22.0	18.9	63.2	1.9	41.6	13.5	58.2	1.0	9.1	6.8	60.3	27.1
PCT	13.0	35.4	13.7	10.2	53.1	1.4	23.8	12.7	52.9	0.8	13.7	1.1	66.2	22.9
ST + PCT	28.9	34.8	27.8	18.6	63.7	4.9	41.0	16.6	64.1	1.6	12.1	6.6	63.9	29.6

Table 7: Experiments on unsupervised domain adaptation with SynLiDAR (as source) and SemanticPOSS (as target): PCT translates SynLiDAR and mitigates domain gaps in the input space effectively. It complements ST and combining them outperforms the baseline (*i.e.*, source-only) by large margins.

get samples and 1 labeled target sample that are randomly selected for 1-shot SSDA-based semantic segmentation.

Tables 4 and 5 show experimental results of PCT and other state-of-the-art SSDA methods. As we can see, training labeled SynLiDAR and one-shot real sample with supervised loss (S+T) does not perform well for both two real datasets due to the domain gap. Including PCT-translated SynLiDAR in training improved the mIoU by 2.2% and 4.5% for SemanticKITTI and SemanticPOSS, respectively. Since PCT mitigates the domain gap in the input space while the state-of-the-art method APE (Kim and Kim 2020) does in the feature space, these two methods are complementary and combining them reached new state-of-the-art performances at 27.0% and 31.2%, respectively.

Experiments on UDA

In this subsection, we focus on unsupervised domain adaptation (UDA) for point cloud segmentation. Different from SSDA, in this setup we only use labeled source data (SynLiDAR) and unlabeled target data (real datasets) for training.

As we can see from Tables 6 and 7, training on labeled source data as source-only failed to learn satisfactory segmentation models due to presence of the domain gap. State-of-the-art UDA methods mitigate the domain gap in either feature space (ADDA) or output space (Ent-Min, ST) and improved the segmentation performance effectively. On the other hand, PCT mitigates the domain gap in input space and including its translated SynLiDAR improved mIoU by 3.5% and 2.8% for SemanticKITTI and SemanticPOSS, respectively. It also complements ST and their combination achieved new state-of-the-art performances at 28.9% and

29.6%, respectively. The experiment results further indicate that PCT effectively reduced the domain gap between SynLiDAR and two real datasets.

Conclusion

This paper presents SynLiDAR and a point cloud translation method PCT for synthetic-to-real transfer learning. SynLiDAR is a large-scale synthetic LiDAR sequential point cloud dataset that contains 19 billion points with point-wise annotations of 32 semantic classes. PCT translates synthetic point clouds to have similar appearance and sparsity as real point clouds. Extensive experiments showed that SynLiDAR shares high geometrical similarities with real-world data, which effectively boosts semantic segmentation while combining with different proportions of real data. PCT mitigates synthetic-to-real gaps effectively and its translated data further improves point cloud segmentation consistently in three transfer learning setups.

Acknowledgments

References

Achlioptas, P.; Diamanti, O.; Mitliagkas, I.; and Guibas, L. 2018. Learning representations and generative models for

3d point clouds. In *International conference on machine learning*, 40–49. PMLR.

Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1534–1543.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, 9297–9307.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.

Chen, Y.; Hu, V. T.; Gavves, E.; Mensink, T.; Mettes, P.; Yang, P.; and Snoek, C. G. 2020. Pointmixup: Augmentation for point clouds. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, 330–345. Springer.

Choy, C.; Gwak, J.; and Savarese, S. 2019a. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3075–3084.

Choy, C.; Gwak, J.; and Savarese, S. 2019b. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3075–3084.

Dai, P.; Zhang, Y.; Li, Z.; Liu, S.; and Zeng, B. 2020. Neural point cloud rendering via multi-plane projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7830–7839.

Fan, H.; Su, H.; and Guibas, L. J. 2017. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 605–613.

Gaidon, A.; Wang, Q.; Cabon, Y.; and Vig, E. 2016. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4340–4349.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9224–9232.

Guan, D.; Huang, J.; Lu, S.; and Xiao, A. 2021. Scale variance minimization for unsupervised domain adaptation in image segmentation. *Pattern Recognition*, 112: 107764.

Hu, Q.; Yang, B.; Xie, L.; Rosa, S.; Guo, Y.; Wang, Z.; Trigoni, N.; and Markham, A. 2020. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11108–11117.

Huang, J.; Guan, D.; Xiao, A.; and Lu, S. 2021a. Cross-view regularization for domain adaptive panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10133–10144.

Huang, J.; Guan, D.; Xiao, A.; and Lu, S. 2021b. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. *Advances in Neural Information Processing Systems*, 34.

Huang, J.; Guan, D.; Xiao, A.; and Lu, S. 2021c. Rda: Robust domain adaptation via fourier adversarial attacking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8988–8999.

Huang, Z.; Yu, Y.; Xu, J.; Ni, F.; and Le, X. 2020. PF-Net: Point fractal network for 3D point cloud completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7662–7670.

Hurl, B.; Czarnecki, K.; and Waslander, S. 2019. Precise synthetic image and lidar (presil) dataset for autonomous vehicle perception. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2522–2529. IEEE.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Kim, T.; and Kim, C. 2020. Attract, perturb, and explore: Learning a feature alignment network for semi-supervised domain adaptation. In *European Conference on Computer Vision*, 591–607. Springer.

Li, R.; Li, X.; Fu, C.-W.; Cohen-Or, D.; and Heng, P.-A. 2019. Pu-gan: a point cloud upsampling adversarial network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7203–7212.

Li, R.; Li, X.; Heng, P.-A.; and Fu, C.-W. 2020. Pointaugment: an auto-augmentation framework for point cloud classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6378–6387.

Liu, F.; Shen, C.; and Lin, G. 2015. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5162–5170.

Milioto, A.; Vizzo, I.; Behley, J.; and Stachniss, C. 2019. RangeNet++: Fast and accurate LiDAR semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4213–4220. IEEE.

Pan, Y.; Gao, B.; Mei, J.; Geng, S.; Li, C.; and Zhao, H. 2020. SemanticPOSS: A Point Cloud Dataset with Large Quantity of Dynamic Instances. *arXiv preprint arXiv:2002.09147*.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv preprint arXiv:1706.02413*.

Qin, C.; You, H.; Wang, L.; Kuo, C.-C. J.; and Fu, Y. 2019. Pointdan: A multi-scale 3d domain adaption network for point cloud representation. *arXiv preprint arXiv:1911.02744*.

Richter, S. R.; Vineet, V.; Roth, S.; and Koltun, V. 2016. Playing for data: Ground truth from computer games. In *European conference on computer vision*, 102–118. Springer.

Ros, G.; Sellart, L.; Materzynska, J.; Vazquez, D.; and Lopez, A. M. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3234–3243.

Roveri, R.; Rahmann, L.; Oztireli, C.; and Gross, M. 2018. A network architecture for point cloud classification via automatic depth images generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4176–4184.

Saito, K.; Kim, D.; Sclaroff, S.; Darrell, T.; and Saenko, K. 2019a. Semi-supervised domain adaptation via minimax entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8050–8058.

Saito, K.; Kim, D.; Sclaroff, S.; Darrell, T.; and Saenko, K. 2019b. Semi-supervised Domain Adaptation via Minimax Entropy. *ICCV*.

Saleh, K.; Abobakr, A.; Attia, M.; Iskander, J.; Nahavandi, D.; Hossny, M.; and Nahvandi, S. 2019. 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*, 0–0.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1): 1929–1958.

Sun, R.; Zhu, X.; Wu, C.; Huang, C.; Shi, J.; and Ma, L. 2019. Not all areas are equal: Transfer learning for semantic segmentation via hierarchical region selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4360–4369.

Tan, C.; Sun, F.; Kong, T.; Zhang, W.; Yang, C.; and Liu, C. 2018. A survey on deep transfer learning. In *International conference on artificial neural networks*, 270–279. Springer.

Tang, H.; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; and Han, S. 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In *European Conference on Computer Vision*, 685–702. Springer.

Thomas, H.; Qi, C. R.; Deschaud, J.-E.; Marcotegui, B.; Goulette, F.; and Guibas, L. J. 2019. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6411–6420.

Tzeng, E.; Hoffman, J.; Saenko, K.; and Darrell, T. 2017. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7167–7176.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

UE4. 2014. Unreal game engine. Accessed Aug 18, 2021.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2517–2526.

Wu, B.; Wan, A.; Yue, X.; and Keutzer, K. 2018. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 1887–1893. IEEE.

Wu, B.; Zhou, X.; Zhao, S.; Yue, X.; and Keutzer, K. 2019. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 International Conference on Robotics and Automation (ICRA)*, 4376–4382. IEEE.

Xiao, A.; Yang, X.; Lu, S.; Guan, D.; and Huang, J. 2021. FPS-Net: A Convolutional Fusion Network for Large-Scale LiDAR Point Cloud Segmentation. *arXiv preprint arXiv:2103.00738*.

Xie, C.; Wang, C.; Zhang, B.; Yang, H.; Chen, D.; and Wen, F. 2021. Style-based Point Generator with Adversarial Rendering for Point Cloud Completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4619–4628.

Yang, J.; Shi, S.; Wang, Z.; Li, H.; and Qi, X. 2021. ST3D: Self-training for Unsupervised Domain Adaptation on 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10368–10378.

Yi, L.; Gong, B.; and Funkhouser, T. 2021. Complete & label: A domain adaptation approach to semantic segmentation of LiDAR point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15363–15373.

Zhao, S.; Wang, Y.; Li, B.; Wu, B.; Gao, Y.; Xu, P.; Darrell, T.; and Keutzer, K. 2020. ePointDA: An End-to-End Simulation-to-Real Domain Adaptation Framework for LiDAR Point Cloud Segmentation. *arXiv preprint arXiv:2009.03456*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017a. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, J.-Y.; Zhang, R.; Pathak, D.; Darrell, T.; Efros, A. A.; Wang, O.; and Shechtman, E. 2017b. Toward multimodal image-to-image translation. *arXiv preprint arXiv:1711.11586*.

Zhu, X.; Zhou, H.; Wang, T.; Hong, F.; Ma, Y.; Li, W.; Li, H.; and Lin, D. 2020. Cylindrical and Asymmetrical 3D Convolution Networks for LiDAR Segmentation. *arXiv preprint arXiv:2011.10033*.

Zou, Y.; Yu, Z.; Liu, X.; Kumar, B.; and Wang, J. 2019. Confidence regularized self-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5982–5991.