Sparse-to-dense Feature Matching: Intra and Inter domain Cross-modal Learning in Domain Adaptation for 3D Semantic Segmentation

Duo Peng[1] Yinjie Lei[1,*] Wen Li[2] Pingping Zhang[3] Yulan Guo[4]
[1]Sichuan University [2]University of Electronic Science and Technology of China
[3]Dalian University of Technology [4]National University of Defense Technology
duo_peng@stu.scu.edu.cn, yinjie@scu.edu.cn, liwenbnu@gmail.com
zhpp@dlut.edu.cn, yulan.guo@nudt.edu.cn

Abstract

*Domain adaptation is critical for success when confronting with the lack of annotations in a new domain. As the huge time consumption of labeling process on 3D point cloud, domain adaptation for 3D semantic segmentation is of great expectation. With the rise of multi-modal datasets, large amount of 2D images are accessible besides 3D point clouds. In light of this, we propose to further leverage 2D data for 3D domain adaptation by intra and inter domain cross modal learning. As for intra-domain cross modal learning, most existing works sample the dense 2D pixel-wise features into the same size with sparse 3D point-wise features, resulting in the abandon of numerous useful 2D features. To address this problem, we propose **D**ynamic **s**parse-to-dense **C**ross **M**odal **L**earning (DsCML) to increase the sufficiency of multi-modality information interaction for domain adaptation. For inter-domain cross modal learning, we further advance **C**ross **M**odal **A**dversarial **L**earning (CMAL) on 2D and 3D data which contains different semantic content aiming to promote high-level modal complementarity. We evaluate our model under various multi-modality domain adaptation settings including day-to-night, country-to-country and dataset-to-dataset, brings large improvements over both uni-modal and multi-modal domain adaptation methods on all settings. Code is available at https://github.com/leolyj/DsCML*

1. Introduction

3D semantic segmentation is a challenging task with plenty of real-world applications, such as particular robotics, autonomous driving and virtual reality. Like other tasks of scene perception, 3D semantic segmentation also faces the challenge of domain shift. For instance, training on one country and testing on another and in different times of the day may lead to significantly dreadful performance. Plenty of domain adaptation methods are proposed to address such domain shift on the task of 2D semantic segmentation [32, 18, 17, 34, 25, 45] but rarely on 3D [44].

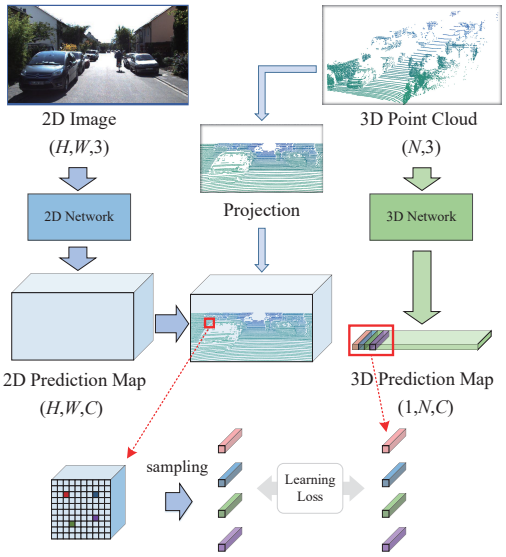


Figure 1. The common strategy of feature processing for 2D-3D cross modal learning. The 2D dense feature map with dense pixel-wise features are sampled to sparse features with the same size of 3D point features. As a result, such sparse-to-sparse feature matching only leverages quite limited 2D features and might cause insufficient 2D-3D information interaction. Specifically, H and W are the height and width of 2D image respectively. N denotes the number of points in point cloud. C is the number of categories for semantic segmentation.

In recent works for multi-modality datasets creation, researchers often incidentally capture 2D images as counterpart when executing the data collection of 3D point clouds. In light of this, Jaritz *et al*. [22] proposes a cross modal learning method to address domain adaptation for 3D semantic segmentation. Through the complementary advantages between 2D and 3D data, the multi-modality domain

*Corresponding Author: Yinjie Lei (yinjie@scu.edu.cn)

adaptation can bring large improvements over uni-modal adaptation methods. And the 2D images are leveraged without labels which means no additional human efforts in labeling work.

While this work [22] already explored multi-modality in domain adaptation, as shown in Fig. 1, it executes the learning between 2D and 3D only towards the matched features by projection from 3D points to 2D image. Multitudes of mismatched features which also contain useful information are discarded. We consider whether 2D and 3D features can be more sufficiently utilized in cross modal learning. This is challenging as the two inputs are heterogeneous and contain different number of elements. Compared to the sparse 3D point clouds, pixels in 2D image are dense and with compact layout. Even if all features of 3D point cloud well matched with the corresponding 2D pixel-wise features, there can be still lots of 2D features are mismatched. That is why numerous works of other perception tasks [37, 9, 24, 6, 21, 26, 39] also capitalize on multi-modality in the same way with [22], i.e., sparse-to-sparse feature matching.

To address such limitation, we propose a strategy namely **D**ynamic **s**parse-to-dense **C**ross **M**odal **L**earning (DsCML) where the sparse point cloud features and dense pixel features can sufficiently interact with each other. Specifically, the proposed DsCML is inspired from the fact that in 2D semantic segmentation, the neighboring pixels are mostly classified to a same category. And all the same-categorized pixel features should be sampled to exchange information with the corresponding 3D point-wise feature. For each 3D point-wise feature, DsCML can dynamically capture the related multiple 2D pixel-wise features with same category. This introduces much richer context information of texture and color in 2D image which is complementary to space information of 3D point cloud. Additionally, a novel sparse-to-dense learning loss is proposed to support the learning of multi-modality where 2D and 3D features differ by orders of magnitude. This DsCML is applied on source and target domain alternately which is the key to domain adaptation.

The method mentioned above is adopted in an intra-domain manner where the 2D and 3D data contain same semantic content. In this paper, we further explore inter-domain cross modal learning for high-level semantic interaction of multi-modality data with different semantic content. Specifically, we introduce cross modal learning to common adversarial strategy by adding discriminator for identification between 2D and 3D features. We coin our method **C**ross **M**odal **A**dversarial **L**earning (CMAL). It enables the mutual learning between 2D and 3D as well as the alignment of feature distribution from different domains, which is the other key to domain adaptation.

The main contributions of this paper are summarized as follows:

- To the best of our knowledge, this is the first work to explore cross modal learning in both intra and inter domain for the semantic segmentation problem.
- As for intra-domain cross modal learning, we propose a module named DsCML to establish sufficient relationships of multi-modality features, i.e., sparse-to-dense feature matching.
- As for inter-domain cross modal learning, we propose a method named CMAL to achieve both high-level cross modal interaction and cross domain feature alignment.
- The proposed method is evaluated on various real-to-real adaptation settings (i.e., day-to-night, country-to-country and dataset-to-dataset), obtaining state-of-the-art segmentation performance with both uni-modal and multi-modal methods.

2. Related Work

In this section, we briefly introduce the techniques related to our approach from three parts. We first give the description of Domain Adaptation in Sec. 2.1. Besides, a considerable literature has grown up around the theme of Multi-Modality Learning in Sec. 2.2. Moreover, various relevant approaches about Adversarial Learning are discussed in Sec. 2.3

2.1. Domain Adaptation

In the past few years, Domain adaptation has attracted great interest as its critical success in new, unseen environments. In the early years, several effective methods have been developed such as maximal confusion [41, 12, 11], Maximum Mean Discrepancy (MMD) [30, 31] and synthesizing images with target styles [47, 36]. Some other works advance adversarial training [48, 27, 18, 5, 17, 40, 35, 45] to narrow source-target distribution difference. Besides, as a typical semi-supervised learning scheme, self-training with pseudo-labels also show positive effect for domain adaptation [49, 50, 28, 29] and capture growing interest.

While promising progress has been achieved, most algorithms focus on the single modality adaptation setting. It lacks the consideration of utilizing the complementary of multi-modality data. In addition, mostly methods are concerned with 2D semantic segmentation adaptation, few [44] adopts domain adaptation in 3D segmentation from point clouds. In sight of this limitation, approach [22] on multi-modal input data (i.e., 2D image + 3D point cloud) has been proposed benefiting from multi-modal data. It assumes that both modalities are available on source and target domains. On this basis, in this paper, we aim to handle the problem in multi-modality domain adaption for 3D semantic segmentation.

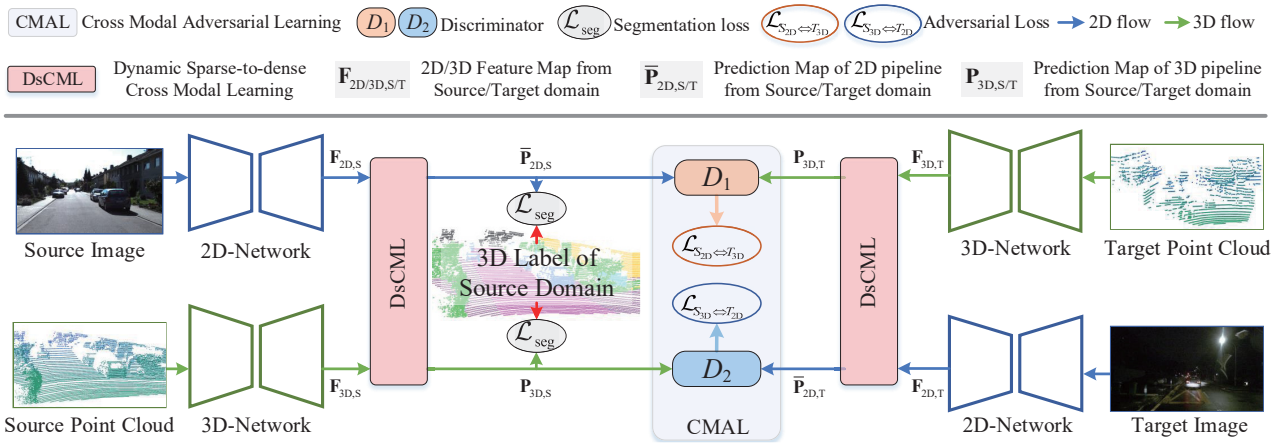


Figure 2. Overall architecture of our approach, which consists of a DsCML module for intra-domain cross modal learning and one adversarial learning for inter-domain cross modal learning. At the end of 2D network and 3D network, DsCML can sufficiently transfer knowledge across multi-modality features (i.e., $\mathbf{F}_{2D,S/T}$ and $\mathbf{F}_{3D,S/T}$) and further generate the prediction maps for 3D semantic segmentation (i.e., $\bar{\mathbf{P}}_{2D,S/T}$ and $\mathbf{P}_{3D,S/T}$). Only the source predictions are supervised by 3D labels. After that, the prediction maps which are from different domains and different modalities are fed into CMAL for higher level cross modal learning.

2.2. Multi-Modality Learning

Taking the advantage of modality complementary is a straightforward and effective way to boost performance. A typical case is the fusion of RGB-Depth images for 2D semantic segmentation [46, 42, 15]. Due to both RGB and Depth images are with the almost same geometrical form, this kind of multi-modality learning is simple to implement. It is challenge to build the bridge for information interaction between 2D and 3D as the heterogeneous data form. A common solution is to filter dense 2D features to sparse point features to enable one-to-one 2D-3D feature matching which is convenient for subsequent processing [37, 8, 9, 24, 6, 21, 39]. However, this type of feature matching leads to the lost of plentiful context information resulting in insufficient 2D-3D interaction. To this end, in this paper, we focus on how to exploit sparse-to-dense feature matching and corresponding learning strategy.

2.3. Adversarial Learning

In domain adaptation, adversarial learning is mainly utilized to narrow the domain gap by reducing the distribution difference between source and target domain. Since the data representation are quite distinct among different feature levels in CNN, methods based on various feature spaces are presented. Among these methods, adaptation on pixel-level [48, 27, 20, 2], feature-level [18, 5, 32, 19], both two levels [17, 43], output-space [40] and label-space [45] are exist.

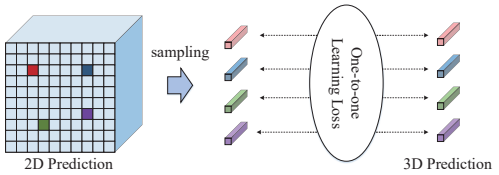
For instance, Chen *et al.* [5] implements a joint global and class-specific adversarial loss at the middle stage feature maps. Zhu *et al.* [48] addresses the adversarial learning on the pixel level, which essentially transfers the style of labeled source images into that of target domain. Hoffman *et al.* [17] proposes method CyCADA where both feature-level and pixel-level adversarial schemes are taken into account. Besides, Tsai *et al.* [40] has proven the effectiveness of output space feature alignment as it jointly promotes the optimization for both classifier and extractor. Moreover, a label-driven adversarial learning is studied in [45] for semantic segmentation. In summary, adversarial learning is a high-level feature constraint towards the holistic data distribution. It enables the learning between two objects with different semantic content, so our method utilizes this advantage.

3. Method

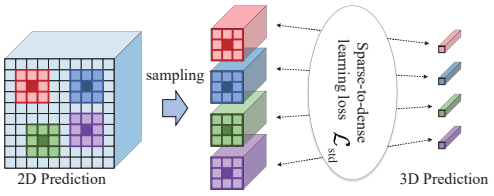
Our approach is presented for 3D semantic segmentation assuming the presence of 2D images and 3D point clouds. In this section, we first describe the architecture overview. Later, we showcase the intra-domain cross modal learning: DsCML in Section 3.2. Finally, the inter-domain cross modal learning: CMAL is introduced in Section 3.3.

3.1. Overview of the Proposed Framework

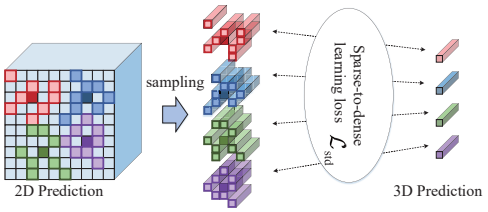
The overall architecture is shown in Fig. 2 (best viewed from both sides to middle). We can briefly describe the main steps as follows. We begin by input the data of source domain S and target domain T into the 2D and 3D network to produce the feature maps before classifier (i.e., $\mathbf{F}_{2D,S}$, $\mathbf{F}_{3D,S}$, $\mathbf{F}_{2D,T}$and $\mathbf{F}_{3D,T}$). Next, the feature maps of each domain are fed into DsCML module for intra-domain cross modal learning. After that, DsCML generates the prediction for 3D semantic segmentation on both source and target domains (i.e., $\bar{\mathbf{P}}_{2D,S}$, $\mathbf{P}_{3D,S}$, $\bar{\mathbf{P}}_{2D,T}$ and $\mathbf{P}_{3D,T}$). It is worthy to mention that DsCML converts the dense 2D feature map into the prediction with the same size as 3D prediction. Hence, we use symbol with superscript (i.e., $\bar{\mathbf{P}}_{2D,S}$



(a) common CML



(b) Ours: sCML



(c) Ours: DsCML

Figure 3. An illustration of feature matching in CML, sCML and DsCML. (a) In common CML, 3D point feature are one-to-one matched with the corresponding 2D pixel-wise features. (b) In sCML, each point features are matched with a square region learning with sparse-to-dense learning loss. (c) In DsCML, the region is deformable which enables the adaptive learning of model for searching the most suitable regions.

and $\overline{\mathbf{P}}_{\mathrm{2D,T}}$) to distinguish it from the 2D image segmentation prediction (i.e., $\mathbf{P}_{\mathrm{2D,S}}$ and $\mathbf{P}_{\mathrm{2D,T}}$). Afterwards, only the source predictions are supervised by the label of source domain. Finally, the source 2D (3D) and target 3D (2D) predictions are fed into the discriminator D_1 (D_2) for adversarial learning, aiming to execute intra-domain cross modal learning.

3.2. Intra-domain cross modal learning: DsCML

The goal of DsCML is to enable sufficient information interaction between the modalities in an intra-domain manner to make them complement each other. This objective can be applied to target domain as its training without access of any annotations. The cross modal learning on target domain enables the domain adaptation for 3D semantic segmentation.

Cross Modal Learning (CML). To better understand how DsCML solves the problem of insufficient cross modal learning, we begin with a common Cross Modal Learning method (CML). As shown in Fig. 3 (a), the loss is implemented on each matched feature pair obtained by sparse-to-sparse matching. Note that we choose 2D and 3D feature maps from the output space where features are already handled by classifier. Hence, it can be formulated as:

$$\mathcal{L} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{K}(\mathrm{Samp}(\mathbf{P}_{\mathrm{2D}})^n, \mathbf{P}_{\mathrm{3D}}^n), \tag{1}$$

where $\mathrm{Samp}(\mathbf{P}_{\mathrm{2D}})^n$ denotes the probability scores of the n-th sampled pixel, $\mathbf{P}_{\mathrm{3D}}^n$ is probability scores of the n-th point, N denotes the number of points in 3D point cloud and $\mathcal{K}(\cdot,\cdot)$ represents the KL distance. We can see that in common CML, only part of the features in 2D feature map are learning with 3D point features, large amount of useful features are discarded.

Sparse-to-dense Cross Modal Learning (sCML). As is well-known, the neighboring pixels mostly belong to a same category in 2D semantic segmentation. In light of this, we try to utilize a square patch of 2D feature map to exchange information with the corresponding point features of 3D point cloud. As shown in Fig. 3 (b), the sparse-to-dense learning loss is the key to implement one-to-many constraints. It can be written as:

$$\mathcal{L}_{std} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{K}(\phi_{max}^n(\mathbf{P}_{\mathrm{2D}}), \mathbf{P}_{\mathrm{3D}}^n) + \mathcal{K}(\phi_{min}^n(\mathbf{P}_{\mathrm{2D}}), \mathbf{P}_{\mathrm{3D}}^n), \tag{2}$$

where $\phi_{max}^n(\mathbf{P}_{\mathrm{2D}})$ denotes the max probability scores in the n-th 2D patch and similarly $\phi_{min}^n(\mathbf{P}_{\mathrm{2D}})$ is the min probability scores in the n-th 2D patch. By constraining the supremum and infimum of a set, we can progressively constrain every element in the patch after several iterations.

Dynamic sparse-to-dense Cross Modal Learning (DsCML). As shown in Fig. 3 (c), we further introduce deformable patch to adaptively search the patch with appropriate region. The sparse-to-dense loss $\mathcal{L}_{std}$ can thus be improved as:

$$\mathcal{L}_{std} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{K}(\Phi_{max}^n(\mathbf{P}_{\mathrm{2D}}), \mathbf{P}_{\mathrm{3D}}^n) + \mathcal{K}(\Phi_{min}^n(\mathbf{P}_{\mathrm{2D}}), \mathbf{P}_{\mathrm{3D}}^n), \tag{3}$$

where $\Phi_{max}^n(\mathbf{P}_{\mathrm{2D}})$ denotes the max probability scores in the n-th 2D deformable patch and $\Phi_{min}^n(\mathbf{P}_{\mathrm{2D}})$ is the min one.

Architecture of DsCML. Fig. 4 illustrates the architecture of our DsCML module for intra-domain cross modal learning. Borrowing from the Deformable Convolution [7], we utilize the extracted offset map to enable the deformable max/min/avg pooling. After pooling, we can select the max/min/avg features of each deformable patch with the guidance of projection from 3D to 2D. Finally, the max/min/avg probability scores are obtained through the classifier. The max and min probability scores are utilized

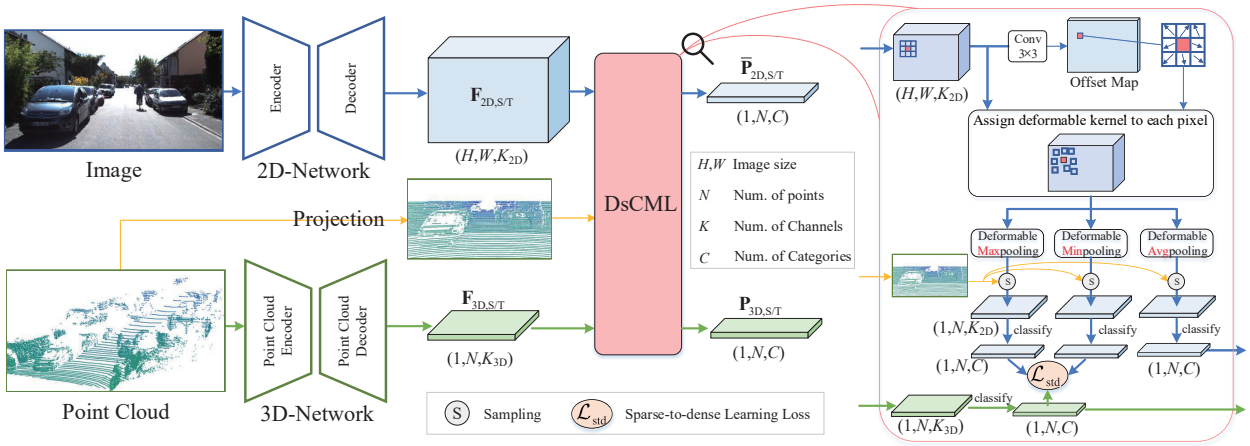


Figure 4. The architecture of our proposed DsCML. With the help of the deformable pooling [7] accompanied with the proposed $\mathcal{L}_{std}$, we can achieve the dynamic sparse-to-dense feature matching.

for cross modal learning in DsCML, while the avg scores are fed into the computation of segmentation loss:

$$\mathcal{L}_{seg} = -\frac{1}{N}\sum_{n=1}^{N} Y_S^n(\log(\Phi_{avg}^n(\mathbf{P}_{2D,S})) + \log(\mathbf{P}_{3D,S}^n)), \tag{4}$$

where Y_S^n denotes the source label of the n-th point and $\Phi_{avg}^n(\mathbf{P}_{2D,S})$ is the average probability scores in the n-th deformable patch of source 2D predicted map.

3.3. Inter-domain cross modal learning: CMAL

Although performing intra-domain cross modal learning can directly improve performance of each modal, there still exist two problems: (a) the learning is mainly towards different modal features with same semantic content, which is a low-level feature alignment; (b) the learning on source and target domain are mutually independent, resulting in the supervision of source label cannot effectively guide the segmentation of target domain. For the latter, Adversarial Learning has shown tremendous adaptation progress by focusing on the mapping between data from different domains in feature space. In light of this, we advance to apply Cross Modal Adversarial Learning (CMAL) to solve both problems with one scheme. In CMAL, the multi-modality learning is execute towards features with different content, different modal and different domain. This high-level feature alignment aims to relieve the distribution difference between source and target as well as 2D and 3D. In this way, the 2D and 3D networks can jointly act 3D semantic segmentation on both domains.

As for CMAL, let $\overline{\mathbf{P}}_{2D,S}, \mathbf{P}_{3D,S} = G(x_{2D,S}, x_{3D,S})$ be the 3D prediction maps of source 2D image $x_{2D,S}$ and 3D point cloud $x_{3D,S}$, where G denotes the multi-modality network with DsCML. Similarly, we process the data in target domain and obtain the predictions $\overline{\mathbf{P}}_{2D,T}$ and $\mathbf{P}_{3D,T}$.

Table 1. The sample number in each split of datasets for all three settings. Note that the training samples in target domain are without labels.

settings	Source	Target		
	train	train	val	test
nuScenes:Day/Night	24745	2779	606	602
nuScenes:USA/Singapore	15695	9665	2770	2929
A2D2/SemanticKITTI	27695	18029	1101	4071

To make the distribution of $\overline{\mathbf{P}}_{2D,S}$ ($\mathbf{P}_{3D,S}$) closer to $\mathbf{P}_{3D,T}$ ($\overline{\mathbf{P}}_{2D,T}$), we use adversarial loss as:

$$\mathcal{L}_{S_{2D}\Leftrightarrow T_{3D}} = -\log(\rho(\overline{\mathbf{P}}_{2D,S})) - \log(1-\rho(\mathbf{P}_{3D,T})), \tag{5}$$

$$\mathcal{L}_{S_{3D}\Leftrightarrow T_{2D}} = -\log(\rho(\mathbf{P}_{3D,S})) - \log(1-\rho(\overline{\mathbf{P}}_{2D,T})), \tag{6}$$

where $\rho(\mathbf{P})$ be the probability that the prediction map $\mathbf{P}$ belongs to the source domain after the identification of discriminator.

We optimize the following min-max criterion:

$$\max_{G}\min_{D_1}\mathcal{L}_{S_{2D}\Leftrightarrow T_{3D}}, \tag{7}$$

$$\max_{G}\min_{D_2}\mathcal{L}_{S_{3D}\Leftrightarrow T_{2D}}, \tag{8}$$

where D_1 and D_2 are two discriminators with same architecture to handle different adversarial tasks. We handle Eq. 7 and 8 by alternating optimization between G and D_1 (D_2).

4. Experiments

4.1. Datasets Description

We strictly follow Jaritz's work: xMUDA [22] to implement our method on three real-to-real adaptation settings: day-to-night, country-to-country and dataset-to-dataset. Three public datasets nuScenes [3], A2D2 [13] and SemanticKITTI [1] are leveraged where the LiDAR and camera are synchronized and calibrated. Only 3D annotations

Table 2. Comparison results with both uni-modal and multi-modal adaptation methods for 3D semantic segmentation in different cross-modal domain adaptation settings. We report the result for each network stream in terms of mIoU. The best two results are marked in bold and underline. 'Avg' denotes the result which is obtained by taking the mean of the predicted 2D and 3D probabilities after softmax.

Modality	Method	USA→Singapore (nuScenes)			Day→Night (nuScenes)			USA→Singapore (Lidarseg)			Day→Night (Lidarseg)			A2D2→SemanticKITTI		
		2D	3D	Avg	2D	3D	Avg	2D	3D	Avg	2D	3D	Avg	2D	3D	Avg
	Baseline (Source only)	53.2	46.8	61.2	41.8	41.4	47.6	53.3	48.0	61.6	41.8	43.8	48.0	36.4	37.3	42.2
Uni-modal	MinEnt [43]	53.4	47.0	59.7	44.9	43.5	51.3	53.6	48.6	61.9	44.9	44.3	51.8	38.8	38.0	42.7
	PL [28]	55.5	51.8	61.5	43.7	45.1	48.6	55.4	52.7	62.8	43.9	47.6	50.9	37.4	44.8	47.7
	FCNs in the Wild [18]	53.7	46.8	61.0	42.6	42.3	47.9	54.0	49.2	62.4	42.6	43.9	48.7	37.1	43.5	43.6
	CyCADA [17]	54.9	48.7	61.4	45.7	45.2	49.7	54.9	51.3	62.6	45.5	47.8	49.6	38.2	43.9	43.9
	AdaptSegNet [40]	56.3	47.7	61.8	45.3	44.6	49.6	56.5	49.0	62.0	45.5	45.3	49.3	38.8	44.3	44.2
	CLAN [32]	57.8	51.2	62.5	45.6	43.7	49.2	57.7	52.1	63.1	45.6	45.1	50.1	39.2	44.7	44.5
Multi-modal	xMUDA [22]	59.3	52.0	62.7	46.2	44.2	50.0	61.7	52.6	63.3	47.3	46.0	50.6	36.8	43.3	42.9
	xMUDA+PL [22]	61.1	54.1	63.2	47.1	46.7	50.8	63.0	54.3	64.2	48.4	47.5	51.2	43.7	48.5	49.1
	DsCML	61.3	53.3	63.6	48.0	45.7	51.0	63.3	54.0	64.2	49.8	47.2	51.7	39.6	45.1	44.5
	DsCML + CMAL	63.4	55.6	64.8	49.5	48.2	52.7	**65.6**	56.2	66.1	50.9	49.3	53.2	46.3	50.7	51.0
	DsCML + CMAL + PL	**63.9**	**56.3**	**65.1**	**50.1**	**48.7**	**53.0**	**65.6**	**57.5**	**66.9**	**51.4**	**49.8**	**53.8**	**46.8**	**51.8**	**52.4**

are utilized for 3D semantic segmentation. Specifically, we leverage nuScenes to generate the splits: Day/Night and USA/Singapore for day-to-night and country-to-country adaptation. The other two datasets are utilized for dataset-to-dataset adaptation, i.e, A2D2/SemanticKITTI. Tab. 1 shows the split details of three datasets for three real-to-real adaptation settings.

As for Day/Night, the 3D point clouds captured by LiDAR show small domain difference due to the sensor has a strong robustness to light variations. While the RGB image capture by camera is the opposite. In the setting of USA/Singapore, the 3D domain difference may be larger than that of 2D in some conditions or vice versa. In A2D2/SemanticKITTI, the density (resolution) of point cloud are large different which highly affect the adaptation performance of 3D network. In this case, the image with small domain gap can help to boost the adaptation performance.

4.2. Implementation Details

Dataset Preprocessing: For domain adaptations of Day/Night and USA/Singapore in nuScenes [3], we utilize the accessible 3D bounding boxes annotations to obtain the 3D point-wise labels as xMUDA [22] did. More specifically, for the point lying inside 3D boxes, we assign it the corresponding object label, otherwise it is labeled as background. To make a more convincing evaluation on our approach, we also experiment on nuScenes-Lidarseg [3] (shorten to 'Lidarseg') which contains the point-wise annotation. For domain adaptation of A2D2[13]/SemanticKITTI[1], we select 10 classes which are shared between the two datasets, i.e., Car, Truck, Bike, Person, Road, Sidewalk, Parking, Nature, Building and Other objects. With the help of code released from xMUDA [22], we compute the projection between each 3D point and its corresponding 2D image pixel.

Network Baseline: To make a fair comparison with the only known multi-modal 3D domain adaptation method [22], for 2D network, we adopt ResNet34 [16] pre-trained on ImageNet [10] as the encoder of U-Net [38]. For 3D network, we use SparseConvNet [14] with U-Net architecture and implement downsampling for six times. Meanwhile, a voxel with size of 5cm is adopted in 3D network, which is small enough to ensure only one 3D point exists in each voxel. The source codes and models are trained and evaluated on PyTorch toolbox [33] based on Python 3.7 platform. All proposed models are implemented on one NVIDIA RTX 3090Ti GPU with 24GB RAM and four E-2224 CPUs.

Parameter Settings: In training period, we choose a batch size of 8 and Adaptive Moment Estimation (Adam) [23] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is set to $1e^{-3}$ initially and follows the poly learning rate policy [4] with a poly power of 0.9. Each deformable patch in DsCML is based on 5×5 square patch. The max training iteration is set to 100k.

Evaluation: Following previous domain adaptation works, we evaluate the performance of a model on the test set by using the standard PASCAL VOC intersection-overunion (IoU). The mean IoU (mIoU) is the mean of all IoU values over all categories. Specifically, the mIoU can be written as follows:

$$\text{mIoU} = \frac{1}{C} \sum_{i=0}^{C} \frac{TP(i)}{TP(i) + FP(i) + FN(i)}, \tag{9}$$

where C is the overall number of categories, $TP(i)$, $FP(i)$, and $FN(i)$ are values of true positive, false positive and false negative towards the i-th category, respectively.

4.3. Comparative Studies

We evaluate our approach on the above three real-to-real adaptation settings and compare with some representative uni-modal domain adaption methods: MinEnt [43], pseudo-labeling (PL) [28], FCNs in the Wild [18], CyCADA [17], AdaptSegNet [40] and CLAN [32]. These uni-modal domain adaptation methods are evaluated on each modality with the same network baselines as ours, i.e., U-Net with ResNet34 encoder (2D network) and SparseConvNet (3D network). In the output space of 2D pipeline, we sample the features from the feature map outputted by 2D network

according to the projection from 3D to 2D. Since AdaptSegNet [40] adopts adversarial learning in the output space, regarding the 2D pipeline of AdaptSegNet, we are faced with two options: implementing adversarial learning on 2D feature map or sampled point features. Herein, we choose the one with better performance from two options, i.e., adaption on sampled point features. Besides, we compare our approach with the only known multi-modal domain adaption method: xMUDA [22].

All comparison results for 3D semantic segmentation are reported in Tab. 2. We can observe that the only usage of DsCML and CMAL brings a significant adaptation effect on all settings compared to Baseline (source only). It is worth noting that our model with only DsCML can outperform all state-of-the-art uni-modal methods. It proves that the two modalities (2D and 3D) are indeed complementary to each other and our DsCML can consistently improves performance of both modalities. From the comparison with "xMUDA+PL" which is the final approah in [22], our model with both DsCML and CMAL achieves the superior performance and contributes 2.5% (2D), 1.8% (3D) and 1.9% (Avg) mIoU gains on average over all settings. Moreover, the model only with DsCML also outperforms "xMUDA" by 2.1% (2D), 1.4% (3D) and 1.1% (Avg) on average. Some qualitative segmentation examples can be viewed in Fig. 5.

4.4. Ablation Studies

4.4.1 Effects of DsCML and CMAL

Next, we conduct additional experiments to demonstrate the benefits of our proposed methods. In Tab. 2, we detail the performance of each design as well as its mIoU improvement by progressively adding DsCML (intra-domain cross modal learning) and CMAL (inter-domain cross modal learning) from the baseline. We can see that DsCML helps to significantly boots the performance of 2D network by 3.2%∼10%, 3D network by 3.4%∼7.8%, Avg by 2.3%∼3.7%. With the addition of CMAL, our model achieves over 1.1% improvement on 2D network and over 2.1% improvement on 3D network. This confirms the effectiveness of our DsCML and CMAL. Besides, we also conduct experiments with PL to demonstrate the complementarity of our model and PL.

4.4.2 Effects of sparse-to-dense feature matching and deformable patch in DsCML

To evaluate the benefits of the proposed sparse-to-dense feature matching in DsCML, we re-implement our approach with CML and sCML, respectively. The comparison results are reported in Tab. 3. Compared to CML (Tab. 3 a), sCML (Tab. 3 b) can stably boost the results over 0.7% on each modal under three settings. It demonstrates the benefits of sparse-to-dense feature matching. To demonstrate the effect of deformable patch in DsCML, we also conducted an ablation experiment by comparing our DsCML with sCML. According to the comparison between model only sCML (Tab. 3 b) and model only with DsCML (Tab. 3 c), we can see that the proposed method DsCML achieves stable improvements and performs best on all three settings. It means the deformable strategy can find the suitable region of patch to exchange information with 3D point features.

Table 3. Performance comparison on CML, sCML and DsCML. Each improvement is obtained by comparing with the upper model.

	Method	USA→Singapore (nuScenes)			Day→Night (nuScenes)			A2D2→Sem.KITTI		
		2D	3D	Avg	2D	3D	Avg	2D	3D	Avg
(a)	CML	59.6	51.7	62.4	46.3	44.3	49.8	36.4	42.8	42.3
(b)	sCML	60.6 (↑1.0)	52.5 (↑0.8)	63.2 (↑0.8)	47.2 (↑0.9)	45.1 (↑0.8)	50.7 (↑0.9)	38.2 (↑1.8)	44.3 (↑1.5)	44.0 (↑0.7)
(c)	DsCML	61.3 (↑0.7)	53.3 (↑0.8)	63.6 (↑0.4)	48.0 (↑0.8)	45.7 (↑0.6)	51.0 (↑0.3)	39.6 (↑1.4)	45.1 (↑0.8)	44.5 (↑0.5)

Table 4. Ablation on feature alignment styles in CMAL. The best two results are marked in bold and underline.

	Adversary Options	U→S (Lidarseg)			D→N (Lidarseg)			A2D2→Sem.KITTI		
		2D	3D	Avg	2D	3D	Avg	2D	3D	Avg
(a)	$S_{2D} \Leftrightarrow T_{2D}$ & $S_{3D} \Leftrightarrow T_{3D}$	<u>64.8</u>	54.9	<u>64.7</u>	<u>50.1</u>	47.3	52.5	40.1	<u>48.2</u>	48.4
(b)	$S_{2D} \Leftrightarrow T_{3D}$ & $S_{3D} \Leftrightarrow T_{2D}$	**65.6**	**56.2**	**66.1**	**50.9**	<u>49.3</u>	**53.2**	<u>46.3</u>	**50.7**	**51.0**
(c)	Both (a) and (b)	63.8	<u>55.3</u>	64.4	49.6	**49.7**	<u>52.9</u>	**46.6**	47.5	<u>49.2</u>

4.4.3 Effects of cross modal alignment in CMAL

As shown in Tab. 4, we conduct ablation studies on three options: (a) inter-modal alignment, (b) cross-modal alignment and (c) both. Except two scores slightly falling behind the best, (b) shows the best performances in all settings. Compared with (a) which only considers relieving domain gap, (b) can effectively improve adaptation performances by simultaneously narrow the domain and modality gap. Although (c) adopts both schemes, it shows unstable performances as it introduces much more discriminators leading to training complexity and difficulty.

4.4.4 Effects of Sparse-to-dense Loss

As mentioned in Sec. 3.2, the sparse-to-dense loss $\mathcal{L}_{std}$ is proposed to constrain each element of a patch aiming to make them have a same probability distribution with the corresponding 3D point. It accomplishes this goal though constraining the supremum and infimum of each patch iteratively. Averaging is a straightforward way to integrate the information of all elements in patch, and naturally compress the 2D output to the same size of 3D output. One could be curious about whether the performance improvement could also be achieved if we change it to the loss based on mean value of each deformable patch. Specifically speaking, we address this concern by conducting additional experiments where the Eq. 3 is changed as follows:

$$\mathcal{L}'_{std} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{K}(\Phi^n_{avg}(\mathbf{P}_{2D}), \mathbf{P}^n_{3D}), \tag{10}$$

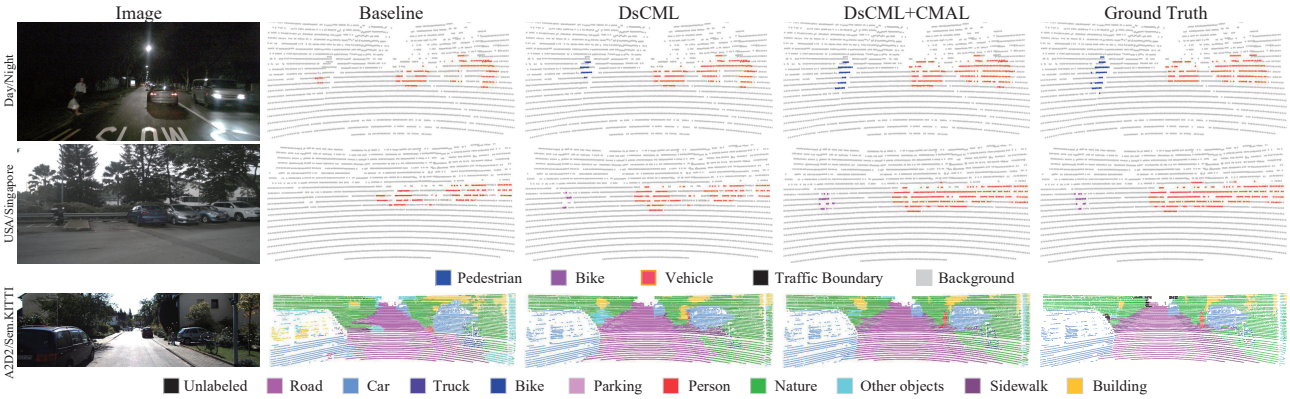


Figure 5. Qualitative 3D semantic segmentation results on three multi-modality adaptation settings: Day/Night, USA/Singapore and A2D2/Sem.KITTI. We show the ensembling results by averaging the softmax outputs of 2D and 3D networks. It can be seen that the segmentation performance is improved by adding the proposed modules progressively.

Table 5. Ablation on sparse-to-dense loss $\mathcal{L}_{std}$. $\mathcal{L}'_{std}$ denotes the loss between 3D point and mean value of each deformable patch. Note that the decline (↓) is obtained by comparing with the first model: DsCML($\mathcal{L}_{std}$).

Method	USA→Singapore (nuScenes)			Day→Night (nuScenes)			A2D2→Sem.KITTI		
	2D	3D	Avg	2D	3D	Avg	2D	3D	Avg
DsCML($\mathcal{L}_{std}$)	61.3	53.3	63.6	48.0	45.7	51.0	39.6	45.1	44.5
DsCML($\mathcal{L}'_{std}$)	60.2 (↓1.1)	52.5 (↓0.8)	62.9 (↓0.7)	47.1 (↓0.9)	44.9 (↓0.8)	50.3 (↓0.7)	37.6 (↓2)	44.3 (↓0.8)	43.6 (↓0.9)
DsCML(w/o $\mathcal{L}_{std}$)	53.4 (↓7.9)	46.9 (↓6.4)	61.5 (↓2.1)	41.8 (↓6.2)	41.6 (↓4.1)	47.6 (↓3.4)	36.5 (↓3.1)	37.3 (↓7.8)	42.3 (↓2.2)

where $\phi^n_{avg}(\mathbf{P}_{2D})$ denotes the average probability scores in the n-th 2D deformable patch. From the comparison results reported in Tab. 5, we observe that with $\mathcal{L}'_{std}$, DsCML performs at least 0.7% worse than that with $\mathcal{L}_{std}$. The performance decline is obviously indicates the effectiveness of our Sparse-to-dense loss which constrains each element in patch. As shown in Fig. 6, constraining on average object of 2D patch is unable to ensure all elements are optimized to a same and correct direction. As mentioned before, $\mathcal{L}_{std}$ is the crucial loss function which enables cross modal learning in DsCML. To demonstrate the effectiveness of loss function between modalities. We conduct experiments using the model without $\mathcal{L}_{std}$. Results are shown the in last row of Tab. 5. As the interaction between 2D and 3D is removed, it shows evident decline. This demonstrates that the loss between 2D and 3D is of great importance for domain adaptation.

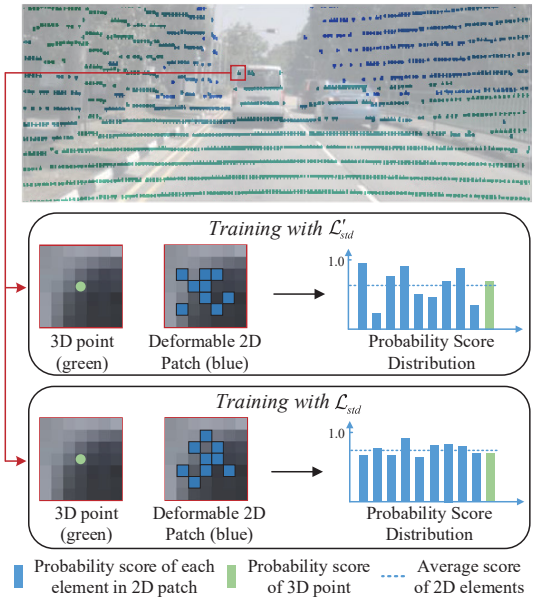


Figure 6. Comparision of probability distribution between learning with $\mathcal{L}'_{std}$ and $\mathcal{L}_{std}$. We show actual probability scores of 2D patch and 3D point after training. As shown above, compared with our $\mathcal{L}_{std}$ which can effectively constrain each 2D element, $\mathcal{L}'_{std}$ may lead to elements with anomalous distribution despite favorable convergence of average distribution.

5. Conclusion

In this paper, we present a multi-modality domain adaptation method for 3D semantic segmentation, which adopts both intra and inter domain cross modal learning. As for intra-domain CML, we advance Dynamic sparse-to-dense Cross Modal Learning (DsCML) to address the problem of insufficient information interaction between dense 2D and sparse 3D features. With the help of sparse-to-dense learning loss, DsCML builds effective consistency constraint between the two heterogeneous data. The design of deformable patch in DsCML enables the network to adaptively search the most suitable 2D region for knowledge transfer with 3D point. As for inter-domain CML, we utilize Cross Modal Adversarial Learning (CMAL) between output features which are both domain-different and modal-different aiming to introduce a higher level modality complementarity. Extensive experiments indicate that our approach achieves the superior performance to both uni-modal and multi-modal methods.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 9297–9307, 2019.

[2] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 3722–3731, 2017.

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020.

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[5] Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1992–2001, 2017.

[6] Hung-Yueh Chiang, Yen-Liang Lin, Yueh-Cheng Liu, and Winston H Hsu. A unified point-based framework for 3d segmentation. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 155–163. IEEE, 2019.

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 764–773, 2017.

[8] Varuna De Silva, Jamie Roche, and Ahmet Kondoz. Fusion of lidar and camera sensor data for environment sensing in driverless vehicles. 2017.

[9] Varuna De Silva, Jamie Roche, and Ahmet Kondoz. Robust fusion of lidar and wide-angle camera data for autonomous mobile robots. *Sensors*, 18(8):2730, 2018.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. Ieee, 2009.

[11] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, pages 1180–1189. PMLR, 2015.

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

[13] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020.

[14] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 9224–9232, 2018.

[15] Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 213–228. Springer, 2016.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1989–1998. PMLR, 2018.

[18] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[19] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11108–11117, 2020.

[20] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Learning semantic segmentation of large-scale point clouds with random sampling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[21] Maximilian Jaritz, Jiayuan Gu, and Hao Su. Multi-view pointnet for 3d scene understanding. In *Proceedings of the International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 0–0, 2019.

[22] 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 Computer Vision and Pattern Recognition (CVPR)*, pages 12605–12614, 2020.

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[24] G Ajay Kumar, Jin Hee Lee, Jongrak Hwang, Jaehyeong Park, Sung Hoon Youn, and Soon Kwon. Lidar and camera fusion approach for object distance estimation in self-driving vehicles. *Symmetry*, 12(2):324, 2020.

[25] Yinjie Lei, Duo Peng, Pingping Zhang, Qiuhong Ke, and Haifeng Li. Hierarchical paired channel fusion network for street scene change detection. *IEEE Transactions on Image Processing*, 30:55–67, 2020.

[26] Yinjie Lei, Ziqin Zhou, Pingping Zhang, Yulan Guo, Zijun Ma, and Lingqiao Liu. Deep point-to-subspace metric learn-

ing for sketch-based 3d shape retrieval. *Pattern Recognition*, 96:106981, 2019.

[27] Peilun Li, Xiaodan Liang, Daoyuan Jia, and Eric P Xing. Semantic-aware grad-gan for virtual-to-real urban scene adaption. *arXiv preprint arXiv:1801.01726*, 2018.

[28] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 6936–6945. IEEE, 2019.

[29] Yan Liu, Lingqiao Liu, Peng Wang, Pingping Zhang, and Yinjie Lei. Semi-supervised crowd counting via self-training on surrogate tasks. In *European Conference on Computer Vision*, pages 242–259. Springer, 2020.

[30] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning (ICML)*, pages 97–105. PMLR, 2015.

[31] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Unsupervised domain adaptation with residual transfer networks. *arXiv preprint arXiv:1602.04433*, 2016.

[32] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 2507–2516. IEEE, 2019.

[33] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[34] Duo Peng, Yinjie Lei, Lingqiao Liu, Pingping Zhang, and Jun Liua. Global and local texture randomization for synthetic-to-real semantic segmentation. *IEEE Transactions on Image Processing*, 2021.

[35] Jian Peng, Bo Tang, Hao Jiang, Zhuo Li, Yinjie Lei, Tao Lin, and Haifeng Li. Overcoming long-term catastrophic forgetting through adversarial neural pruning and synaptic consolidation. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[36] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: towards real-time object detection with region proposal networks. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(6):1137–1149, 2016.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

[39] Hang Su, Varun Jampani, Deqing Sun, Subhransu Maji, Evangelos Kalogerakis, Ming-Hsuan Yang, and Jan Kautz. Splatnet: Sparse lattice networks for point cloud processing. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 2530–2539, 2018.

[40] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 7472–7481, 2018.

[41] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 4068–4076, 2015.

[42] Abhinav Valada, Rohit Mohan, and Wolfram Burgard. Self-supervised model adaptation for multimodal semantic segmentation. *International Journal of Computer Vision (IJCV)*, pages 1–47, 2019.

[43] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 2517–2526, 2019.

[44] 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 *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 4376–4382. IEEE, 2019.

[45] Jinyu Yang, Weizhi An, Sheng Wang, Xinliang Zhu, Chaochao Yan, and Junzhou Huang. Label-driven reconstruction for domain adaptation in semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 480–498. Springer, 2020.

[46] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 202–211, 2017.

[47] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 597–613. Springer, 2016.

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 2223–2232. IEEE, 2017.

[49] 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. Springer, 2018.

[50] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *Proceedings of the Computer Vision and Pattern Recognition (CVPR)*, pages 5982–5991. IEEE, 2019.