ISPRS Journal of Photogrammetry and Remote Sensing 176 (2021) 211–221

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

Adversarial unsupervised domain adaptation for 3D semantic segmentation with multi-modal learning

Wei Liu[a], Zhiming Luo[b,*], Yuanzheng Cai[c], Ying Yu[a], Yang Ke[d], José Marcato Junior[e], Wesley Nunes Gonçalves[e], Jonathan Li[d]

[a] School of Software, East China Jiaotong University, Nanchang 330013, China
[b] Artificial Intelligence Department, Xiamen University, Xiamen 361005, China
[c] Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), Fuzhou 350121, China
[d] University of Waterloo, Waterloo N2L 3G1, Canada
[e] Federal University of Mato Grosso do Sul, Campo Grande 79070 900, Brazil

ARTICLE INFO

Keywords:
Semantic segmentation
Point cloud
Domain adaptation
Adversarial learning
Multi-modal learning

ABSTRACT

Semantic segmentation in 3D point-clouds plays an essential role in various applications, such as autonomous driving, robot control, and mapping. In general, a segmentation model trained on one source domain suffers a severe decline in performance when applied to a different target domain due to the cross-domain discrepancy. Various Unsupervised Domain Adaptation (UDA) approaches have been proposed to tackle this issue. However, most are only for uni-modal data and do not explore how to learn from the multi-modality data containing 2D images and 3D point clouds. We propose an Adversarial Unsupervised Domain Adaptation (AUDA) based 3D semantic segmentation framework for achieving this goal. The proposed AUDA can leverage the complementary information between 2D images and 3D point clouds by cross-modal learning and adversarial learning. On the other hand, there is a highly imbalanced data distribution in real scenarios. We further develop a simple and effective threshold-moving technique during the final inference stage to mitigate this issue. Finally, we conduct experiments on three unsupervised domain adaptation scenarios, *ie.,* Country-to-Country (USA →Singapore), Day-to-Night, and Dataset-to-Dataset (A2D2 →SemanticKITTI). The experimental results demonstrate the effectiveness of proposed method that can significantly improve segmentation performance for rare classes. Code and trained models are available at https://github.com/weiliu-ai/auda.

1. Introduction

The main goal of 3D Semantic segmentation is to estimate the semantic class (*e.g.*, bike, car, building, road) for each point in the 3D point cloud (Serna and Marcotegui, 2014; Liu et al., 2016; Grilli et al., 2017; Dai et al., 2018; Luo et al., 2020). It plays an important role in many applications, such as autonomous driving, robot control, and mapping. Recently, advances in the 3D semantic segmentation significantly improved the results with convolutional neural networks (CNN) (Wu et al., 2018; Wu et al., 2019; Wang et al., 2019). However, these methods still suffer limitations from the following two aspects. Firstly, training a 3D semantic segmentation network usually requires massive amounts of labeled 3D point cloud data, which is time-consuming and labor-intensive to collect. Secondly, models trained on one dataset usually encounter a significant performance degradation on a new domain due to the appearance distribution gap or shift between different datasets, such as different illumination conditions and different locations.

To reduce the domain gap, various Unsupervised Domain Adaptation (UDA) approaches (Qin et al., 2019; Wu et al., 2019; Jaritz et al., 2020; Luo et al., 2020) have been proposed. There are also a few UDA studies (Qin et al., 2019; Wu et al., 2019; Luo et al., 2020) have been proposed for the 3D segmentation based on a single modality. However, in practice, 3D datasets are often multi-modal, typically consisting of 3D point clouds and 2D images. On the other side, related researches (Guo et al., 2019; Vu et al., 2019; Feng et al., 2020) have shown that exploiting the complementary information between the 2D and the 3D modalities is beneficial for the segmentation task.

Captured by different devices, the 2D and 3D modalities have very different properties and can resist different domain shifts. Therefore, it is a challenge to leverage their relationships to improve the performance

* Corresponding author.
E-mail address: zhiming.luo@xmu.edu.cn (Z. Luo).

https://doi.org/10.1016/j.isprsjprs.2021.04.012
Received 18 December 2020; Received in revised form 16 April 2021; Accepted 18 April 2021

of both modalities during the UDA. Jaritz et al. (2020) propose a cross-modal UDA framework named xMUDA, which achieves state-of-the-art performance for UDA-based 3D segmentation. The xMUDA adopts a KL-divergence to transfer the information of the two-modality among the matched 2D pixels and 3D points, and enforce the estimation of the 2D and 3D to be consistent. However, the matched 2D pixels are extremely sparse and cannot fully utilize the global structure information of the whole image in the source domain. In this study, we hypothesize that it could be beneficial for the UDA to use the whole 2D images instead of sampling sparse 2D points in the source domain. The primary motivation is that making the image features from the two domains with a similar distribution could narrow the domain gap and thus improve 3D semantic segmentation. To do so, we propose a domain adaptation framework for 3D semantic segmentation, combining adversarial learning and multi-modal learning.

On the other hand, real-world datasets have highly imbalanced data distributions. In general, *road* and *background* often constitute the dominant classes. Therefore, the trained segmentation model will likely be biased to these dominant classes, which will lead to a low recall rate for the rare classes (e.g., *pedestrian* and *bike*). To address this issue, we further incorporate a cost-sensitive loss function to balance the influence of different classes and propose an adaptive threshold-moving post-processing step for improving the recall rate for rare classes.

To summarize, the main contributions of this study are as follows:

(1) To make full use of the 2D features for 3D semantic segmentation, we proposes a UDA based segmentation framework integrating adversarial learning with multi-modal learning, which improves the predicted results of the 2D sub-network by a significant margin.
(2) To mitigate the harmful effects of imbalanced class distribution, we devise a simple but effective threshold-moving technique, which significantly improves the segmentation performance for rare classes while keeping the overall segmentation performance at a high level.
(3) We performed experiments on three unsupervised domain adaptation scenarios for 3D semantic segmentation. Experimental results demonstrate that our method obtains competitive performance to existing methods.

2. Related work

In this section, we discuss the related works regarding: unsupervised domain adaptation, adversarial domain adaptation, and multi-modality learning.

2.1. Unsupervised domain adaptation

The aim of domain adaptation (DA) is to mitigate the distribution discrepancy between the source domain and target domain (Wilson and Cook, 2020; Toldo et al., 2020). To bridge the domain gap, various UDA techniques (Liu and Li, 2014; Tzeng et al., 2017; Yan et al., 2019; Zhao et al., 2019; Zhao et al., 2019; Zhu et al., 2019; Iqbal and Ali, 2020; Zhang et al., 2020) have been proposed. These methods mainly approach UDA by minimizing the discrepancy between the labeled source and unlabeled target data at three different levels (Toldo et al., 2020), including input-level (Li et al., 2019; Hoffman et al., 2018), feature-level (Zhu et al., 2018; Volpi et al., 2018; Lee et al., 2019), and output-level (Luo et al., 2019; Vu et al., 2019; Pan et al., 2020).

To accomplish the input-level DA, many works address the statistical matching at the input level to achieve cross-domain uniformity of visual appearance of the input image samples (Toldo et al., 2020). These works mainly utilize style transfer techniques to close the source's marginal distributions and target images from original image-level sets. The typical approach is to design a function that maps source samples into a domain invariant space.

The feature-level DA approaches seek a distribution alignment of feature embeddings. All these works share the same core idea of forcing the feature extractor to extract domain-invariant features by adjusting the distribution of latent representations from source and target domains.

To avoid the high complexity, the output-level adaptation performs the cross-domain distribution alignment across the segmentation output space. A domain discriminator is provided with prediction maps from source and target inputs, and is optimized to recognize the input domain. Conversely, the segmentation network has to fool it by aligning the distribution of predicted dense labels across domains.

Although many UDA works have been proposed for segmentation, most these studies deal with the 2D image segmentation (Zou et al., 2018; Chen et al., 2019; Vu et al., 2019; Vu et al., 2019; Pan et al., 2020; Iqbal and Ali, 2020). Only a few have been proposed for the 3D segmentation (Qin et al., 2019; Wu et al., 2019; Luo et al., 2020). Compared with the 2D image modality, the 3D point clouds are usually unstructured and unordered, making its corresponding UDA more challenging.

2.2. Adversarial domain adaptation

In the field of 2D image segmentation (Liu et al., 2018; Mi and Chen, 2020), various UDA approaches have been proposed to minimize cross-domain discrepancy by employing adversarial training (Vu et al., 2019; Tzeng et al., 2017; Tsai et al., 2018; Vu et al., 2019; Pan et al., 2020; Michieli et al., 2020). These adversarial schemes mainly consist of two networks, a generator paired with a discriminator. The generator is designed to learn to produce data with the same statistical distribution of training samples. The goal of the discriminator is to discern the input domain of the input data. To make statistics of generated data match that of the training set, the generator is optimized to fool the discriminator by producing samples that resemble the original ones. Adversarial learning-based UDA methods have demonstrated the efficiency in aligning feature distributions of the two domains at the image feature level (Hoffman et al., 2018; Murez et al., 2018) or the output level (Tsai et al., 2018; Tsai et al., 2019).Hoffman et al. (2016) has been the first to address domain adaptation in semantic segmentation. In particular, they devise a global domain adversarial alignment based on a domain discriminator taking as input the feature representations from the segmentation network's intermediate activations. Following a similar approach to (Hoffman et al., 2016), a line of works (Chen et al., 2018; Zhang et al., 2018; Li et al., 2019) seeks for alignment of latent network embeddings. There are some researches (Hoffman et al., 2018; Chen et al., 2019) combines a generative approach with the adversarial feature alignment. To accomplish category-wise adaptation, some works (Chen et al., 2017; Du et al., 2019) exploit multiple feature discriminators.

To avoid the complexity of high-dimensional feature space, there are many works (Tsai et al., 2018; Chang et al., 2019; Luo et al., 2019) accomplish adversarial adaptation on the low-dimensional output space spanned by the segmentation network. These schemes typically consist of a segmentation network and a domain discriminator. In particular, the segmentation network has to fool the domain discriminator by aligning the distribution of predicted labels across domains. Conversely, the domain discriminator takes as input predicted segmentation maps from source and target.

2.3. Multi-Modality Learning

In practical application, 3D datasets usually consist of data from different modalities, typically 3D point clouds (Liu et al., 2019) and 2D images. Therefore, it is essential to explore the relationship between different modalities. Previous studies (Guo et al., 2019; Vu et al., 2019; Jaritz et al., 2020; Feng et al., 2020) have shown the benefits of exploiting the complementarity between 2D and 3D modalities for 3D semantic segmentation.

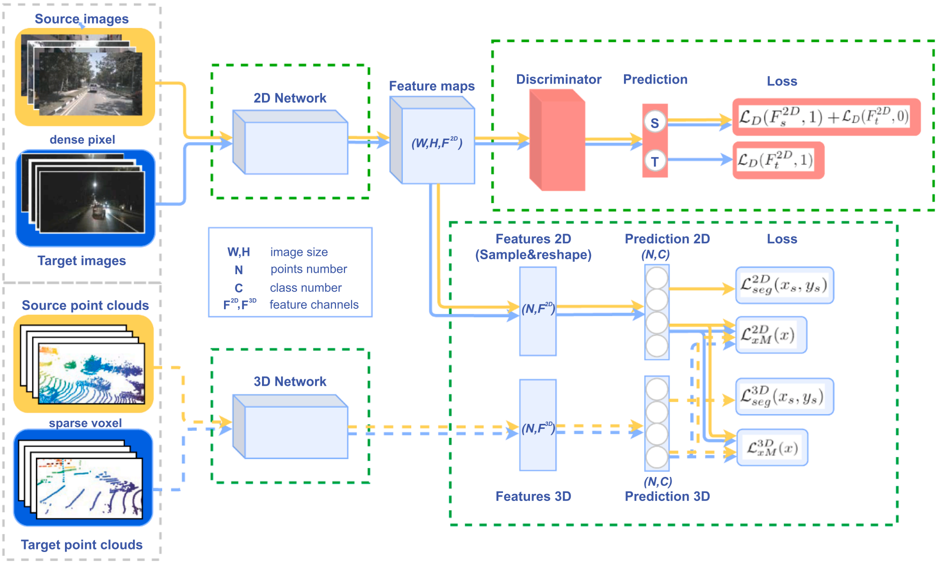


Fig. 1. Illustration of the proposed 3D semantic segmentation framework. The proposed framework consists of two main modules. One module is a two-stream network which predicts the semantic labels for the input point cloud and the corresponding front image, which could be from the source or target domain. The other module acts as a domain discriminator that takes the feature maps from the 2D segmentation sub-network and tries to predict the domain of the input. (Best view in color). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

There are two common types of multi-modal 2D-and-3D data. One is RGB-D data (Valada et al., 2019; Vu et al., 2019), consisting of a RGB image and a depth image where are pixel-to-pixel corresponded. Vu et al. (2019) built a unified depth-aware UDA framework, DADA, improving the semantic segmentation by leveraging the depth prediction as an auxiliary task. Valada et al. (2019) proposed a UDA architecture, which involves two modality-specific encoders and a unified decoder. The unified decoder fuses the intermediate representations from the two encoders as input and is trained in a self-supervised adaptation fusion manner.

The other type is a 3D point cloud with a 2D image. Compared to RGB-D data, it is harder to fuse the 3D point cloud with a 2D image since the 3D point cloud data is unstructured and unordered. In the task of object detection, the common strategy is projecting both 2D and 3D features into a 'LiDAR front view' (Meyer et al., 2019) or 'bird-eye view' (Liang et al., 2018; Liang et al., 2019). For the semantic segmentation task, a solution for multi-modality learning is to interpolate 2D features from single-view or multi-view images into the 3D point cloud, thus enabling joint 2D-3D processing (Su et al., 2018; Chiang et al., 2019; Jaritz et al., 2019; Jaritz et al., 2020). Similar toJaritz et al. (2020), we only use single view images and their corresponding point clouds for the multi-modality learning in this study.

3. Problem definition

Suppose we have a labeled source dataset $S = \{x_s^j, y_s^j\}_{j=1}^{N_s}$ and an unlabeled target dataset $T = \{x_t^j\}_{j=1}^{N_t}$, where N_s and N_t are the number of the observed source point clouds and target point clouds, respectively. The domain S and domain T share the same predefined semantic classes $\{1,\ldots,C\}$, where C is the number of classes. For the source dataset S, each sample x_s consists of an unlabeled 2D front image x_s^{2D} and a labeled 3D point cloud x_s^{3D} with point-wise 3D semantic segmentation labels y_s. The image data labels are obtained by the projection of the corresponding 3D labeled point clouds on the image plane. Hence, only a small portion of image pixels have labels. For each sample x_t in T, there is no available annotation in the training stage.

Given S and T, the main goal for the 3D semantic segmentation task is to learn an unsupervised adaption model that can correctly predict the labels of every 3D point for the target domain. Notice that, only the 3D points visible in the corresponding image are used for training and evaluation.

4. The proposed framework

Fig. 1 illustrates the architecture of our proposed adversarial training-based segmentation framework for bridging the domain gap. The proposed framework mainly involves two modules. One module is a two-stream segmentation network, which predicts the segmentation maps for the point cloud and image pair from either the source or target domain. The other module acts as a domain discriminator that takes the feature maps from the 2D segmentation sub-network and tries to predict the input domain. By leveraging the adversarial training strategy, we enforce the image features from the two domains have a similar distribution by fooling the discriminator.

4.1. *Supervised learning on the labeled source domain*

On the labeled source domain, we train our model with a supervised segmentation loss, i.e., cross-entropy loss function. In detail, for each sample $x_s \in S$, it consists of a 2D front image x_s^{2D}, a 3D point cloud x_s^{3D}, and point-wise semantic segmentation labels y_s. Let N be the number of points in x_s^{3D}, the supervised segmentation loss for the 2D front image x_s^{2D} can be formulated as:

$$\mathscr{L}_{seg}^{2D}(x_s, y_s) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_s^{(n,c)} \log P_{2D}^{(n,c)}, \tag{1}$$

where $y_s^{(n,c)}$ and $P_{2D}^{(n,c)}$ is the ground-truth label and prediction of the point n for the class c, respectively.

Similarly, the supervised segmentation loss for the point cloud x_s^{3D} is as follows:

$$\mathscr{L}_{seg}^{3D}(x_s, y_s) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_s^{(n,c)} \log P_{3D}^{(n,c)}. \tag{2}$$

Therefore, the overall objective for the 2D and the 3D sub-networks on the labeled source domain is:

$$\min_{\theta_{2D},\theta_{3D}} \frac{1}{N_s}\sum_{x_s \in S} \mathscr{L}_{seg}^{2D}(x_s, y_s) + \mathscr{L}_{seg}^{3D}(x_s, y_s), \tag{3}$$

where θ_{2D} and θ_{3D} are the parameters of the 2D sub-network and the 3D sub-network, respectively.

4.2. Adversarial learning scheme

To reduce the domain gap between the source domain and the target domain, we adopt adversarial learning to align the feature distribution between these two domains. The main intuition of adversarial learning is that if we can learn a feature space that minimizes the distance between the source and target distributions, then we could apply the model trained on the labeled source domain directly to the target domain. In this study, the proposed adversarial learning is implemented in an adversarial training procedure with the assistance of a discriminator. The discriminator network takes the 2D representations from the two domains as inputs and classifies them either from the source or the target domain. By contrast, the 2D feature networks will then optimize into a feature distribution space in which the discriminator can not distinguish the input domains.

As shown in the top-right of Fig. 1, given a front image x from the source or target domain, the 2D segmentation sub-network extract the feature map F^{2D} of the image x. After that, the discriminator predicts whether the feature map F^{2D} is from the source or the target domain. During the training, the 2D segmentation sub-network will try to fool the discriminator by making feature maps from the two domains with a similar distribution.

To this end, we construct a 4-layer fully-convolutional discriminator network D with parameters θ_{Dis}. It takes the image feature F^{2D} as input and is trained to distinguish the source images from the target ones. We label the source domain and the target domain as '1' and '0', respectively. Let $\mathscr{L}_D$ represent the cross-entropy domain classification loss of the discriminator. The training objective of the discriminator is:

$$\min_{\theta_{Dis}} \frac{1}{N_s}\sum_{x_s^{2D}} \mathscr{L}_D(F_s^{2D}, 1) + \frac{1}{N_t}\sum_{x_t^{2D}} \mathscr{L}_D(F_t^{2D}, 0). \tag{4}$$

As mentioned above, the 2D segmentation sub-network is trained to fool the discriminator. Thus, the adversarial objective to update the 2D segmentation sub-network is:

$$\min_{\theta_{2D}} \frac{1}{N_t}\sum_{x_t^{2D}} \mathscr{L}_D(F_t^{2D}, 1) \tag{5}$$

4.3. Multi-modal learning

In general, the 2D image depicts the object's appearance while the 3D image indicates the object structure from the depth that makes these two modalities provide complementary information for each other. To learn a shared space that both 2D and 3D features are projected to, we use the KL divergence (Kullback and Leibler, 1951; Jaritz et al., 2020) to increase the similarity of output distribution from the 2D and 3D modalities. Given a sample x from the source domain or the target domain, the cross-modal loss for the 2D segmentation sub-network is defined as follows:

$$\mathscr{L}_{xM}^{2D}(x) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} P_{2D}^{(n,c)} \log \frac{P_{2D}^{(n,c)}}{P_{3D}^{(n,c)}}, \tag{6}$$

where $P_{2D}^{(n,c)}$ and $P_{3D}^{(n,c)}$ are the predicted probabilities of the 2D sub-network and the 3D sub-network for the point n about the class c, respectively.

Likewise, the KL divergence based cross-modal loss for the 3D segmentation sub-network is:

$$\mathscr{L}_{xM}^{3D}(x) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} P_{3D}^{(n,c)} \log \frac{P_{3D}^{(n,c)}}{P_{2D}^{(n,c)}}. \tag{7}$$

4.4. Complete objectives for the proposed learning scheme

The complete objective for the 2D segmentation sub-network is the combination of the supervised segmentation loss $\mathscr{L}_{seg}^{2D}$ on the source domain, the adversarial objective on the target domain, and the cross-modal loss $\mathscr{L}_{xM}^{2D}$ on both of the two domains. We formulate the complete objective for the 2D segmentation sub-network as:

$$\min_{\theta_{2D}} \frac{1}{N_s}\sum_{x_s\in S}\left\{\mathscr{L}_{seg}^{2D}(x_s, y_s) + \lambda_s \mathscr{L}_{xM}^{2D}(x_s)\right\} + \frac{1}{N_t}\sum_{x_t\in T}\{\lambda_{adv}\mathscr{L}_D(F_t^{2D}, 1) + \lambda_t \mathscr{L}_{xM}^{2D}(x_t)\}, \tag{8}$$

where hyperparameters λ_s, λ_t, and λ_{adv} are the weights of the corresponding loss functions.

The complete objective for the 3D segmentation sub-network is the combination of the supervised segmentation loss $\mathscr{L}_{seg}^{3D}$ on the source domain, and the cross-modal loss on both of the two domains $\mathscr{L}_{xM}^{3D}$. Thus, the objective loss function of 3D segmentation sub-network can be formulated as:

$$\min_{\theta_{3D}} \frac{1}{N_s}\sum_{x_s\in S}\left\{\mathscr{L}_{seg}^{3D}(x_s, y_s) + \lambda_s \mathscr{L}_{xM}^{3D}(x_s)\right\} + \frac{1}{N_t}\sum_{x_t\in T}\lambda_t \mathscr{L}_{xM}^{3D}(x_t). \tag{9}$$

The values of the three hyper-parameters are selected on the validation set. For each epoch during training on source and target, we alternatively optimize the segmentation network and the discriminator. Specifically, the segmenation network is optimized by the objective loss function (8) and (9). The discriminator is optimized by the objective loss function (4).

4.5. Techniques to deal with class imbalance

In practice, many real-world datasets have an imbalanced class distribution, ie., *pedestrian*, *bike* and *truck* often constitute a minority of the data set in contrast with *road* and *background* (Luo et al., 2018). The imbalanced class distribution could bias the segmentation model to the dominant classes. In this study, we use two techniques to deal with this class-imbalance issue, including incorporating class ratio priors for training and a simple threshold-moving technique.

To tackle the class imbalance issue, we first incorporate class ratio priors computed from the source labels into the segmentation framework during the training phase. Similar to Vu et al. (2019) and Jaritz et al. (2020), we assign different weights for different classes. Concretely, let n_i be the number of points labeled as class i in the source domain, the log-smoothed class weight w_c for class c can be computed as:

$$w_c = \frac{\log\left(\frac{\alpha\sum_{k=1}^{C} n_k}{n_c}\right)}{\min_{1\leqslant j\leqslant C}\log\left(\frac{\alpha\sum_{k=1}^{C} n_k}{n_j}\right)}, \tag{10}$$

where α is a hyper-parameter controlling the smoothness of the weights and is set to 5. Notice that, $w_c = 1$ for the category with the most points in the source domain, $w_c > 1$ for the others. By injecting the weights into the loss function, we then have cost-sensitive loss functions of the Eqs. (1) and (2) are:

$$\mathscr{L}_{seg}^{2D}(x_s, y_s) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c y_s^{(n,c)} \log P_{2D}^{(n,c)}, \tag{11}$$

and

$$\mathscr{L}_{seg}^{3D}(x_s, y_s) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c y_s^{(n,c)} \log P_{3D}^{(n,c)}, \tag{12}$$

respectively.

Apart from incorporating the cost-sensitive functions in the training phase, we also devise a simple but effective threshold-moving technique

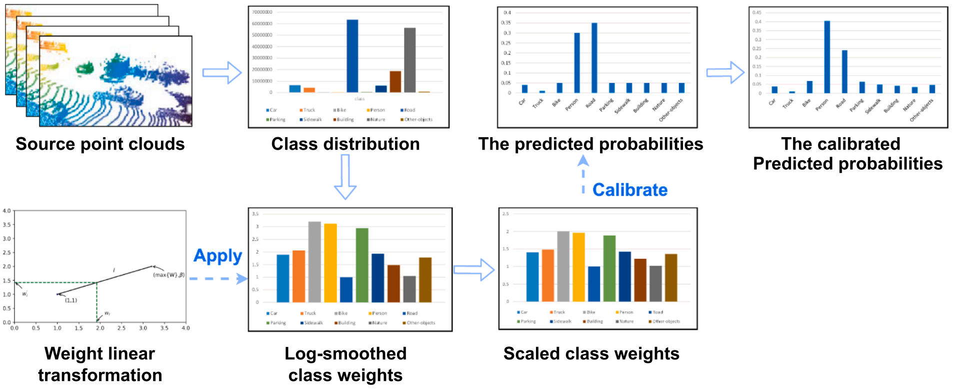


Fig. 2. Illustration of the proposed threshold-moving technique. For the source domain with an imbalanced distribution, we first compute log-smoothed classes weights, and scale it to the range $[1,\beta]$. Then we calibrate the predicted probabilities on the target domain using the scaled class weights. (Best view in color and zooming in for more details). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to increase the recall-rate of the minor categories during the testing phase. The overall computation process of this threshold-moving technique is presented in Fig. 2. As shown in the block at the bottom left of the Fig. 2, our goal is to scale the class weights $\mathscr{W} = \{w_i\}_{i=1}^{C}$ to the range $[1,\beta]$. Specifically, we obtain a line l, according the two points $(1,1)$ and $(max\{w_i|w_i \in \mathscr{W}\},\beta)$. For $w_i \in \mathscr{W}$, we get the scaled value $w_i^{'}$ from the line l taking w_i as the abscissa. Let $m = max\{w_i|w_i \in \mathscr{W}\}$, then the scaled weight $w_i^{'}$ can be computed as:

$$w_i^{'} = \frac{(\beta - 1) * w_i + m - \beta}{m - 1} \tag{13}$$

Given the predicted probabilities $\mathbf{p} = (p_1, p_2, \cdots, p_C)$ with respect to sample x, we calibrate the predict $\mathbf{p}$ using the scaled log-smoothed class weights. The calibrated probabilities can be represented as:

$$\mathbf{p}^{'} = \frac{1}{\sum_{i}^{C} w_i^{'} p_i}(w_1^{'}p_1, w_2^{'}p_2, \cdots, w_C^{'}p_C). \tag{14}$$

In this paper, we set the value of hyperparameter β to 1.5 for all of the three scenarios in the experiment section.

5. Experiments

In this section, we report the datasets, the evaluation metrics, the implementation details, the comparison with other methods, and the ablation study of our proposed method.

5.1. Datasets and evaluation metrics

We evaluate the performance of the proposed method and the baselines in three real-to-real adaption scenarios, including country-to-country, day-to-night, and dataset-to-dataset, following the setup of Jaritz et al. (2020). Three recently published autonomous driving datasets, *i.e.*, nuScenes dataset(v1.0) (Caesar et al., 2020), A2D2 (Geyer et al., 2020), and SemanticKITTI (Behley et al., 2019) are adopted to build the scenarios.

For each dataset, the LiDAR and the RGB-camera are synchronized and calibrated. Consequently, we can directly obtain the projection between a 3D point and its corresponding 2D image pixel. In this study, we only use the front camera image and the corresponding projected LiDAR points for training and testing.

(1) **nuScenes**. The original dataset(v1.0) consists of 1,000 driving scenes collected from the USA and Singapore using LiDAR and RGB camera. Two domain adaptation scenarios are designed in this dataset, including Country → Country and Day → Night. Since there are no semantic labels available of the official testing set, therefore, we treat the official validation set as the testing set and divide the official training set into train/val, as the same as (Jaritz et al., 2020). A total of 750 driving scenes out of 1000 are used in this paper. The point-wise 3D semantic labels are obtained from 3D boxes like in (Wu et al., 2018), and the objects are merged into 5 categories, *i.e.*, *vehicle*, *pedestrian*, *bike*, *traffic boundary* and *background*.
(2) **A2D2**. The point cloud comes from 3 LiDARs with 16 layers. The point clouds in this dataset are rather sparse. It provides semantic segmentation labels for 2D images. 3D labels were obtained by projection of the point cloud into the labeled 2D image.
(3) **SemanticKITTI**. The dataset contains point clouds from 10 different scenes captured by one high-resolution LiDAR with 64 layers. The 3D point clouds are labeled into 28 different classes. In this study, the point clouds from Scene $\{0,1,2,3,4,5,6,9,10\}$ are used as the training set, Scene 7 as the validation set, and the Scene 8 as the testing set.

The details of the three cross-domain scenarios are presented as follows:

- **Country-to-Country** (USA → Singapore). The domain shift can be large for LiDAR or camera: for some classes the 3D shape might change more than the visual appearance or vice versa. The train set for the source has 15,695 frames. The train set, validation set and the test set of the target domain have 9,665 frames, 2,770 frames, and 2,929 frames respectively.
- **Day-to-Night**. LiDAR is an active sensor sending out laser beams which are mostly invariant to lighting conditions. In contrast, cameras suffer from lack of light sources, leading to drastic changes in object appearance. Compared to LiDAR, camera has a large domain gap. The train set for the source has 24,745 frames. The train set, validation set and the test set of the target domain have 2,779 frames, 606 frames, and 602 frames respectively.
- **Dataset-to-Dataset** (A2D2 → SemanticKITTI). Point clouds in SemanticKITTI are denser than point clouds in A2D2. The scenario contains 10 shared classes, including *car*, *truck*, *bike*, *pedestrian*, *road*, *parking*, *sidewalk*, *building*, *nature*, and *other-objects*. The train set for the source domain has 27,695 frames. The train set, validation set, and the test set of the target domain have 18,029 frames, 1,101 frames, and 4,071 frames, respectively.

Two commonly used semantic segmentation evaluation metrics, including mean Intersection-over-Union (mIoU) and Average Recall (AR), are used to evaluate the proposed method's performance.

Table 1
Comparisons of different models in three different UDA scenarios (%) in terms of mIoU. For brevity, we use SG for Singapore and SK for SemanticKITTI. *Avg* takes the mean of the predicted probabilities of the 2D and 3D sub-networks after softmax.

Method	USA → SG			Day → Night			A2D2 → SK		
	2D	3D	Avg	2D	3D	Avg	2D	3D	Avg
Source only	53.4	46.5	61.3	42.2	41.2	47.8	36.0	36.6	41.8
DL (Morerio et al., 2017)	52.6	47.1	59.1	41.4	42.8	51.8	35.8	39.3	40.3
MinEnt (Vu et al., 2019)	53.4	47.0	59.7	44.9	43.5	51.3	38.8	38.0	42.7
PL (Li et al., 2019)	55.5	51.8	61.5	43.7	45.1	48.6	37.4	44.8	**47.7**
xMUDA (Jaritz et al., 2020)	59.3	**52.0**	62.7	46.2	44.2	50.0	38.3	**46.0**	44.0
AUDA (ours)	**59.8**	**52.0**	**63.1**	**49.0**	**47.6**	54.2	**43.0**	43.6	46.8
AUDA$_{TM}$ (ours)	59.7	51.7	63.0	48.7	46.2	**55.7**	43.3	43.3	47.3

Table 2
Comparisons of different models in three different UDA scenarios (%) in terms of mIoU. For brevity, we use SG for Singapore and SK for SemanticKITTI. *Avg* takes the mean of the predicted 2D and 3D probabilities after softmax.

Method	USA → SG			Day → Night			A2D2 → SK		
	2D	3D	Avg	2D	3D	Avg	2D	3D	Avg
xMUDA (Jaritz et al., 2020)	59.3	52.0	62.7	46.2	44.2	50.0	38.3	46.0	44.0
xMUDA$_{PL}$ (Jaritz et al., 2020)	61.1	54.1	63.2	47.1	46.7	50.8	41.2	**49.8**	47.5
AUDA (ours)	59.8	52.0	63.1	49.0	47.6	**54.2**	43.0	43.6	46.8
AUDA$_{PL}$ (ours)	**61.9**	**54.8**	**65.6**	**50.3**	**49.7**	52.6	**46.8**	48.1	**50.6**

5.2. Implementation details

In this study, we implement our model by using the PyTorch toolbox, and adopt the SparseConvNet (Graham et al., 2018) for the 3D network and a modified version of U-Net (Ronneberger et al., 2015) with ResNet34 (He et al., 2016) for the 2D network. For the **3D network**, the official PyTorch implementation and a U-Net with six times downsampling are utilized. The 3D voxel size is set to 5 cm. For the **2D network**, we use the ImageNet pre-trained weights to initialize the encoder. Each layer consists of a transposed convolutional layer, a skip connection layer, and another convolution layer to mix the features from the previous two layers. The **discriminator network** after the 2D network has four convolutional layers with the leaky-ReLU activation. The kernel size, stride, and padding size of these convolutional layers are all set to 4, 2, and 1, respectively. The hyper-parameters λ_s, λ_{adv} and β are set to 1, 0.001, and 1.5, respectively. The weight λ_t is set as 0.1 for the USA → Singapore and Day → Night scenarios, and 0.01 for the A2D2 → SemanticKITTI scenario.

During the training phase, we optimize the parameters of the model using the Adam optimizer (Kingma and Ba, 2014). The initial learning rate is 10^{-4} for the discriminator and 10^{-3} for the 2D network and 3D network. The batch size is set as 8 for the USA → Singapore and Day → Night scenarios, and 6 for the A2D2 → SemanticKITTI scenario, respectively. We train the model for 150,000 iterations on each scenario, and the whole optimization process takes around 48 h on a single 12 GB TITAN X Pascal GPU.

5.3. The comparison with other different methods

In this section, we compare our proposed method with other methods on the three adaption scenarios. The comparing methods can be categorized into (1)“Source only”, the model is only trained on the source domain, and directly test on the target domain; (2) uni-modal UDA methods, *i.e.,* Deep logCORAL (DL) (Morerio et al., 2017), entropy minimization (MinEnt) (Vu et al., 2019), pseudo-labeling (PL) (Li et al., 2019); (3) cross-modal UDA method, *i.e.,* xMUDA (Jaritz et al., 2020). The comparison results are reported in Table 1. For the scenario of A2D2 → SemanticKITTI scenario of xMUDA (Jaritz et al., 2020), we use the newest result on their GitHub project page[1].

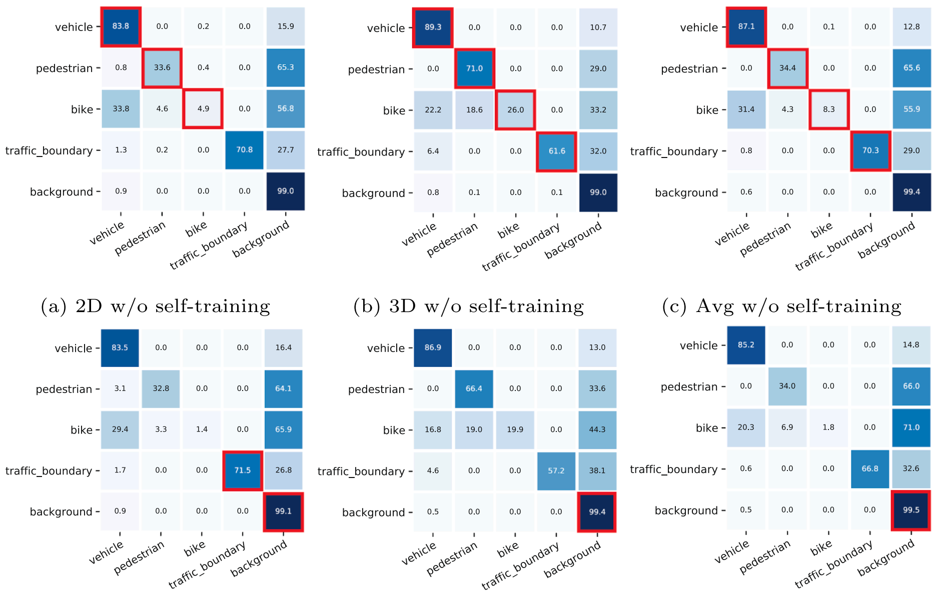
As shown in Table 1, using only the 2D sub-network for 3D segmentation during the testing phase, our method AUDA achieves the best results in all three scenarios in terms of mIoU, in particular in Day → Night and A2D2 → SemanticKITTI. Specifically, the proposed method AUDA achieves 59.8%, 49.0% and 43.0% mIoU. It shows that our algorithm could well exploit the 2D modality. When combining the 3D segmentation results of the 2D and 3D sub-networks, the proposed method achieves comparable or better results to the baselines. Especially compared with the-state-of-art method xMUDA, AUDA achieves improvements of 0.4%, 4.2%, and 2.8% mIoU on USA → Singapore, Day → Night, and A2D2 → SemanticKITTI. The thresh-moving technique is devised to increase recall rates of rare classes. With the proposed thresh-moving technique, our method AUDA$_{TM}$ has comparable performance with AUDA, in terms of mIoU.

5.3.1. Effectiveness of adversarial learning

One main difference between the proposed method and xMUDA is that the proposed method uses adversarial training to bridge the domain gap. We further evaluate the effectiveness of the proposed adversarial learning module, compared with self-training (Vu et al., 2019; Pan et al., 2020; Jaritz et al., 2020), which involves generating pseudo-labels for target samples during the training stage. The compared methods are (1) xMUDA (Jaritz et al., 2020); (2)xMUDA$_{PL}$(Jaritz et al., 2020), *i.e.,* combining xMUDA with self-training; (3) AUDA; (4)AUDA$_{PL}$, *i.e.,* combining AUDA with self-training.

From Table 2, we can have the following observations. (1) The self-training technique can increase accuracy of 2D and 3D segmentation for both xMUDA and our AUDA. For example, on the USA→SG, the 2D and 3D accuracy of our AUDA will increase from 59.8% to 61.9% and 52.0% to 54.8%, respectively. (2) We also notice that combining the results of 2D and 3D may decrease the mIoU, as shown in Avg of Day→Night by our AUDA (54.2% vs. 52.6%). The main reason is that the quality of pseudo-labels has a great influence on self-training. The overall confidence scores for the rare classes are relatively low compared with the dominant classes. In the self-training step, a high threshold is used to

[1] https://github.com/valeoai/xMUDA



(a) 2D w/o self-training (b) 3D w/o self-training (c) Avg w/o self-training

(d) 2D with self-training (e) 3D with self-training (f) Avg with self-training

Fig. 3. Normalized confusion matrices of the proposed method in the Day → Night scenario with five categories. The first row of the sub-figures corresponds to the results of the proposed method AUDA, and the second row corresponds to the results of AUDA$_{PL}$. The rows of confusion matrices correspond to truth classes, and columns correspond to predicted classes. The diagonal elements represent the recall values. (Best view in color and zooming in for more details.).

Table 3
The ablation study on hyperparameter λ_{adv} for the weight of the adversarial loss function.

	Day → Night					
λ_{adv}	0	0.0001	0.001	0.01	0.1	1
mIoU	50.0	53.0	54.2	52.7	50.4	51.5

Table 4
Statistics for the Day → Night scenario. The second row shows the number of points for each category on the source domain. The third row shows the log-smoothed class weights computed using the Eq. 10. The forth row shows the scaled weights computed using Eq. 13, based on the log-smoothed weights.

	vehicle	pedestrian	bike	traffic boundary	background
# of points	4582123	275688	42299	608214	77672950
W	2.7	4.4	5.5	3.9	1.0
$W^{'}$	1.2	1.4	1.5	1.3	1.0

select samples and then may ignore the information of rare classes. Thus, it will cause a performance decrease.

To further illustrate the impact of self-training on rare categories, we visualize the normalized confusion matrices of the proposed method in the Day → Night scenario. As shown in Fig. 3, the proposed model doesn't perform well for the rare classes, *car* and *bike*, in term of AR. After the use of self-training, the performance of the proposed model regarding rare *pedestrian* and *bicycles* is further reduced, with only the AR of *background* increased. In the Day → Night scene, *background* is the largest class.

5.3.2. *Analysis of hyper-parameter λ_{adv}*

Similar to the one in (Pan et al., 2020), we also conduct a hyper-parameter sensitivity analysis on the λ_{adv} of the final segmentation result in our experiment of the Day → Night scenario. As presented in Table 3, the best performance of the model is achieved at $\lambda_{adv} = 0.001$. Notice that our method degenerates to xMUDA, with $\lambda_{adv} = 0.001$. When the value of λ_{adv} is in the range of $\{0.0001, 0.001, 0.01, 0.1, 1\}$, the proposed perform better than xMUDA. These results indicate that the effectiveness of the proposed adversarial scheme and is robust to the value of the hyper-parameter λ_{adv}.

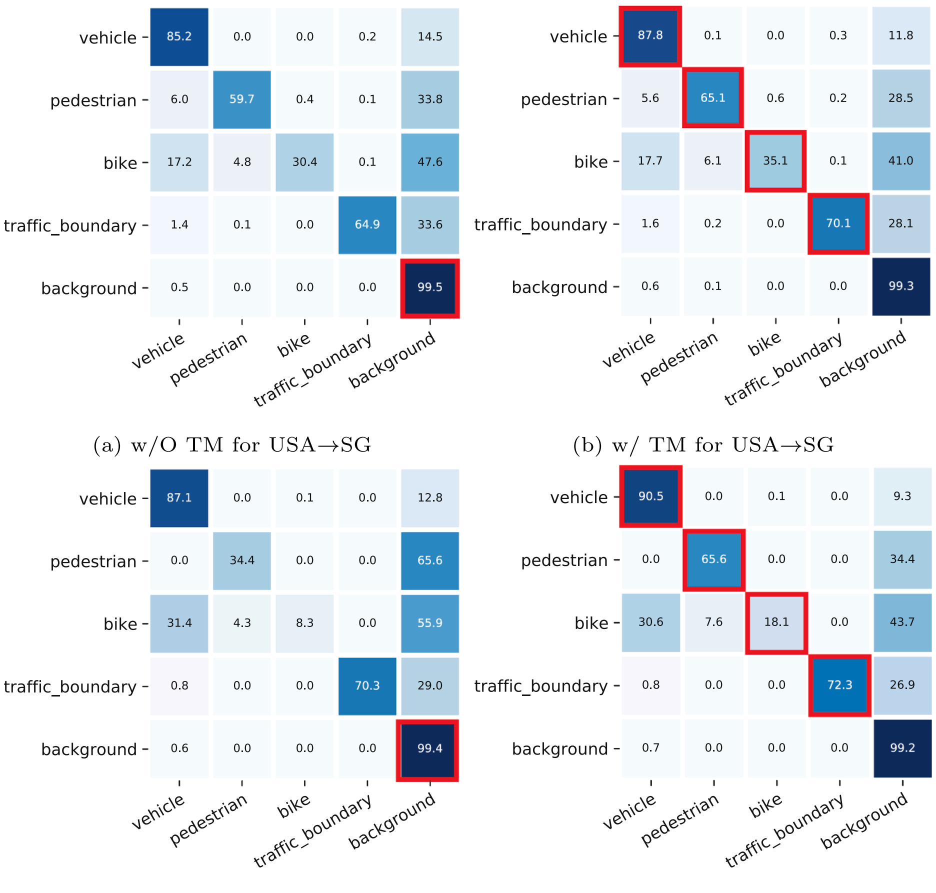
Table 5
Effects of the proposed threshold-moving technique in 3 different UDA scenarios (%). For brevity, we use SG for Singapore, SK for SemanticKITTI, AR for average recall, and TM for threshold-moving.

Method	USA → SG		Day → Night		A2D2 → SK	
	AR	mIoU	AR	mIoU	AR	mIoU
w/o TM	67.9	63.1	59.9	54.2	61.9	46.8
w/ W	79.0	56.2	77.0	45.2	65.9	46.2
w/ $W^{'}$	71.5	63.0	69.2	55.7	63.7	47.3

5.3.3. *Effect of the proposed threshold-moving technique*

We also design experiments to evaluate the effectiveness of the proposed threshold-moving technique for imbalanced segmentation. As shown in Table 4, the source domain has a highly imbalanced class distribution for the source domain in the scenario of Day → Night. Rare classes including *pedestrian* and *bike* only constitute a minority of the data set, in contrast with *vehicle* and *background*. The log-smoothed class weights W computed using the Eq. 10 have a wide range of values, in the range 1 to 5.5. Further, using Eq. 13 for a linear transformation, the log-smoothed weights are mapped to the ranges 1 to 1.5. Table 4 indicates that the proposed threshold-moving technique allows rarer categories to have higher weights, and keep the range of the weights not too large.

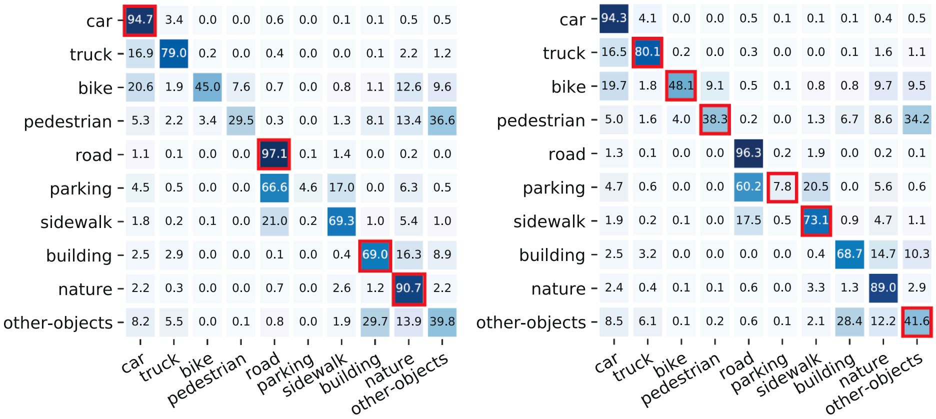
As presented in Table 5, using the proposed threshold-moving technique with the scaled weights W, the proposed method attains AR rates of 71.5%, 69.2%, and 63.7% over USA → Singapore, Day → Night, and A2D2 → SemanticKITTI, respectively. Without the proposed threshold-moving technique, the proposed method attains AR of 67.9%, 59.9%, and 61.9% over USA → Singapore, Day → Night, and A2D2 → SemanticKITTI, respectively. The threshold-moving technique increase AR by 3.4%, 9.3%, and 1.8%, respectively. Decreasing false negatives may increase false positives. Therefore, it is difficult to greatly improve AR and mIoU at the same time. In terms of mIoU, these two methods are



(a) w/O TM for USA→SG (b) w/ TM for USA→SG

(c) w/O TM for Day→Night (d) w/ TM for Day→Night

Fig. 4. Normalized confusion matrices of the proposed method in the USA → Singapore and Day → Night scenario with five categories. The rows of confusion matrices correspond to truth classes, and columns correspond to predicted classes. The diagonal elements represent the recall values. The diagonal elements represent the recall values. (Best view in color and zooming in for more details.).



(a) w/O TM (b) w/ TM

Fig. 5. Normalized confusion matrices of the proposed method in the A2D2 → SemanticKITTI scenario with ten categories. The rows of confusion matrices correspond to truth classes, and columns correspond to predicted classes. (Best view in color and zooming in for more details.).

comparable on the three scenarios. Using the log-smoothed weights W, the proposed method can greatly improve AR, with the risk of a sharp decline in mIoU. The results in Table 5 indicate that the proposed threshold-moving technique with the scaled weights significantly boost the performance in terms of AR while keeping mIoU within a small range of changes on all of the three scenarios.

Moreover, we visualize normalized confusion matrices of the proposed method as presented in Fig. 4. For the scenario of Day → Night, as shown in Fig. 4c and d, the proposed method achieves recall of 87.1%, 34.4%, 8.3%, 70.3% and 99.4% for *vehicle*, *pedestrian*, *bike*, *traffic boundary* and *background*, respectively. With the proposed threshold-moving technique, we obtain 90.5%, 65.6%, 18.1%, 72.3% and 99.2% for *vehicle*, *pedestrian*, *bike*, *traffic boundary* and *background*, respectively. For the rare classes *pedestrian* and *bike*, the proposed threshold-moving

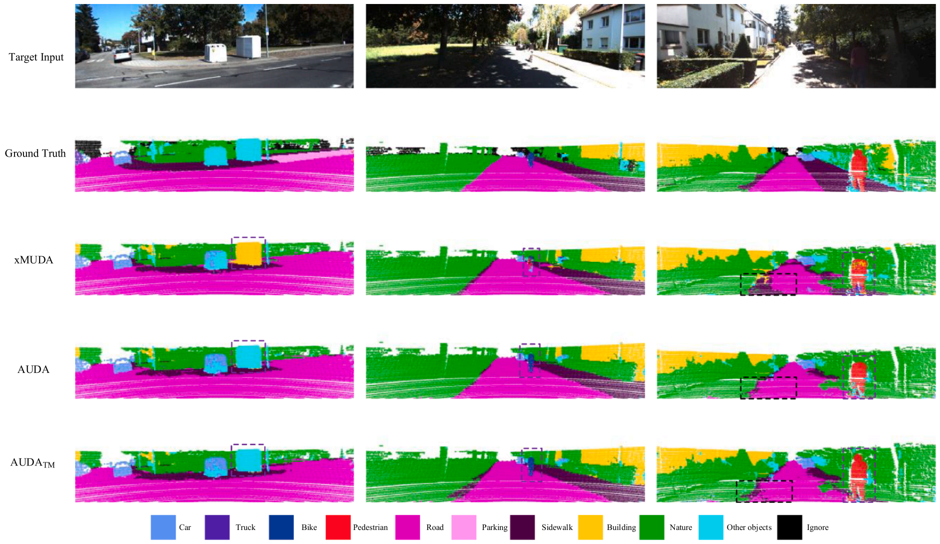


Fig. 6. The example results of evaluation for the A2D2→ SemanticKITTI set-up. Dash bounding boxes are added to show the main differences in the outputs of the three methods. (Best view in color and zooming in for more details.).

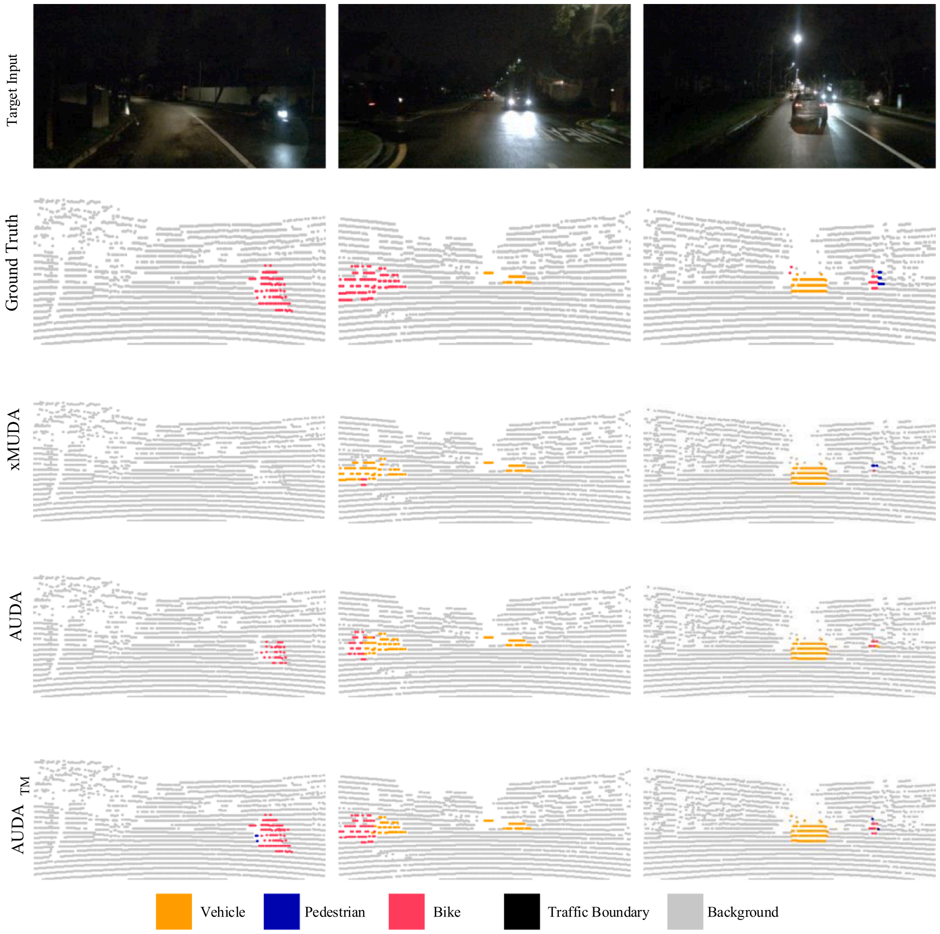


Fig. 7. The example results of evaluation for the Day→ Night set-up. (Best view in color and zooming in for more details.).

technique increases the recall by 31.2% and 9.8%, respectively. It is worth noting that the recall of the rare classes has been significantly improved, while the recall of the dominant class *background* has not declined significantly. Similar results can be seen from Figs. 4a, b and 5.

Fig. 6 shows some qualitative results for the A2D2 → SemanticKITTI scenario. We observe that overall, the proposed method with threshold-moving (AUDA$_{TM}$) performs best. Especially for the rare classes such as *bike* and *pedestrian*, the proposed method AUDA$_{TM}$ attains the best results. We can also draw consistent conclusions from Figs. 7 and 8.

Therefore, the proposed thresh-moving technique can improve the segmentation performance for rare classes while keeping the dominant classes' segmentation performance at a high level.

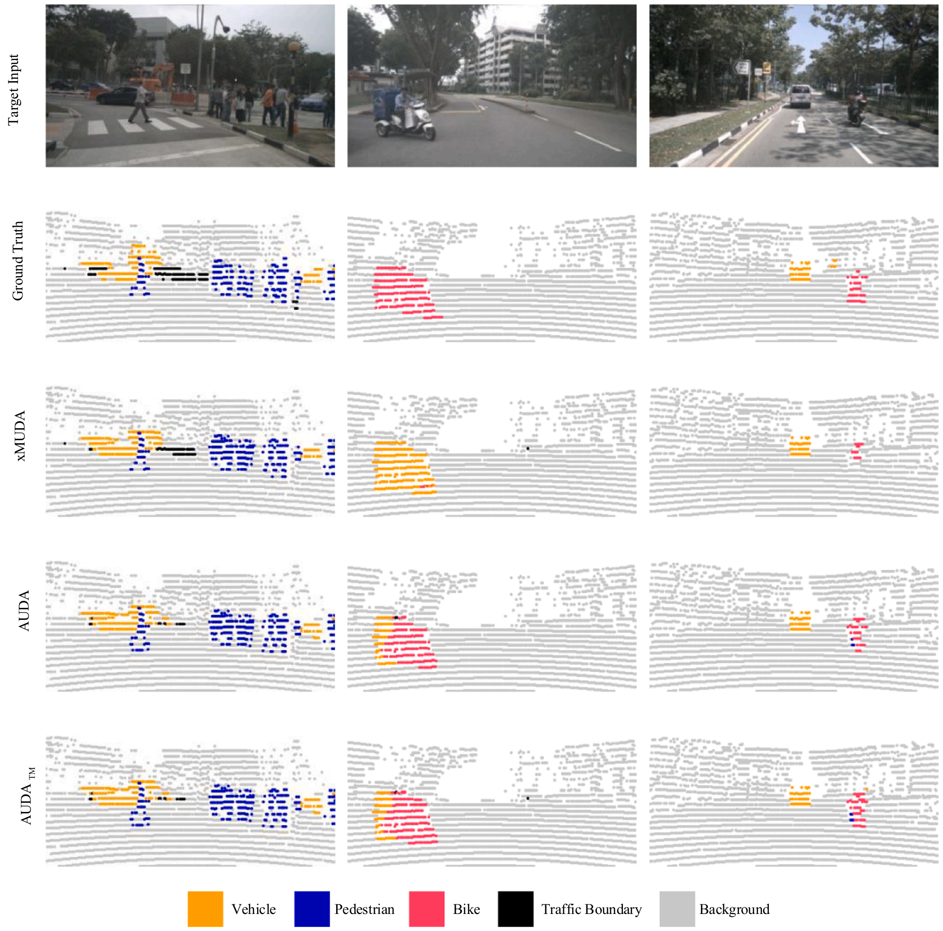


Fig. 8. The example results of evaluation for the USA→ Singapore set-up. (Best view in color and zooming in for more details.).

6. Conclusion

In this paper, we propose an Adversarial Unsupervised Domain Adaptation (AUDA) for 3D semantic segmentation framework to mitigate the harmful effects of domain shift without requiring any training annotations on the target domain. By combining adversarial learning and multi-modal learning, the proposed framework fully exploit the complementary information between the 2D and 3D modalities. We incorporate class ratio priors over the source labels into the segmentation framework to address the class imbalance and improve segmentation performance for rare classes, such as *pedestrian* and *bike*. We also devise a threshold-moving technique to refrain the model biasing to dominant classes and improve the framework's average recall. The experimental results on three domain adaptation scenarios demonstrate the effectiveness of proposed methods that can significantly improve segmentation performance for rare classes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61662024, 61806172, 61866012, 41871380), the Natural Science Foundation of Jiangxi Province (20202BAB202021), the Natural Science Foundation of Fujian Province (2020J01828), the Open Fund of Fujian Province Key Laboratory (KLCCIIP2019201), and the Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control(Minjiang University)(Grant No. MJUKF-IPIC202006).

References

Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J., 2019. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In: IEEE International Conference on Computer Vision, pp. 9297–9307.

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

Chang, W.L., Wang, H.P., Peng, W.H., Chiu, W.C., 2019. All about structure: Adapting structural information across domains for boosting semantic segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1900–1909.

Chen, M., Xue, H., Cai, D., 2019. Domain adaptation for semantic segmentation with maximum squares loss. In: IEEE International Conference on Computer Vision, pp. 2090–2099.

Chen, Y., Li, W., Van Gool, L., 2018. Road: Reality oriented adaptation for semantic segmentation of urban scenes, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7892–7901.

Chen, Y.C., Lin, Y.Y., Yang, M.H., Huang, J.B., 2019. Crdoco: Pixel-level domain transfer with cross-domain consistency, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1791–1800.

Chen, Y.H., Chen, W.Y., Chen, Y.T., Tsai, B.C., Frank Wang, Y.C., Sun, M., 2017. No more discrimination: Cross city adaptation of road scene segmenters. In: IEEE International Conference on Computer Vision, pp. 1992–2001.

Chiang, H.Y., Lin, Y.L., Liu, Y.C., Hsu, W.H., 2019. A unified point-based framework for 3d segmentation, in. In: IEEE International Conference on 3D Vision, pp. 155–163.

Dai, W., Yang, B., Dong, Z., Shaker, A., 2018. A new method for 3d individual tree extraction using multispectral airborne lidar point clouds. ISPRS journal of photogrammetry and remote sensing 144, 400–411.

Du, L., Tan, J., Yang, H., Feng, J., Xue, X., Zheng, Q., Ye, X., Zhang, X., 2019. Ssf-dan: Separated semantic feature based domain adaptation network for semantic segmentation, in. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 982–991.

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., Wiesbeck, W., Dietmayer, K., 2020. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. IEEE Trans. Intell. Transp. Syst.

Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., Fernandez, T., Jänicke, M., Mirashi, S., Savani, C., Sturm, M., Vorobiov, O., Oelker, M., Garreis, S., Schuberth, P., 2020. A2D2: Audi Autonomous Driving Dataset URL: https://www.a2d2.audi, arXiv:2004.06320.

Graham, B., Engelcke, M., Van Der Maaten, L., 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 9224–9232.

Grilli, E., Menna, F., Remondino, F., 2017. A review of point clouds segmentation and classification algorithms. The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 42, 339.

Guo, Z., Li, X., Huang, H., Guo, N., Li, Q., 2019. Deep learning-based image segmentation on multimodal medical imaging. IEEE Transactions on Radiation and Plasma Medical Sciences 3, 162–169.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. Cycada: Cycle-consistent adversarial domain adaptation. In: International Conference on Machine Learning, pp. 1989–1998.

Hoffman, J., Wang, D., Yu, F., Darrell, T., 2016. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. arXiv preprint arXiv:1612.02649.

Iqbal, J., Ali, M., 2020. Weakly-supervised domain adaptation for built-up region segmentation in aerial and satellite imagery. ISPRS Journal of Photogrammetry and Remote Sensing 167, 263–275.

Jaritz, M., Gu, J., Su, H., 2019. Multi-view pointnet for 3d scene understanding. In: IEEE International Conference on Computer Vision Workshops.

Jaritz, M., Vu, T.H., Charette, R.d., Wirbel, E., Pérez, P., 2020. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition, pp. 12605–12614.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. In: International Conference on Learning Representations.

Kullback, S., Leibler, R.A., 1951. On information and sufficiency. The annals of mathematical statistics 22, 79–86.

Lee, S., Kim, D., Kim, N., Jeong, S.G., 2019. Drop to adapt: Learning discriminative features for unsupervised domain adaptation. In: IEEE International Conference on Computer Vision, pp. 91–100.

Li, Y., Yuan, L., Vasconcelos, N., 2019. Bidirectional learning for domain adaptation of semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 6936–6945.

Liang, M., Yang, B., Chen, Y., Hu, R., Urtasun, R., 2019. Multi-task multi-sensor fusion for 3d object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7345–7353.

Liang, M., Yang, B., Wang, S., Urtasun, R., 2018. Deep continuous fusion for multi-sensor 3d object detection. In: European Conference on Computer Vision, pp. 641–656.

Liu, W., Li, S., Cao, D., Su, S., Ji, R., 2016. Detection based object labeling of 3d point cloud for indoor scenes. Neurocomputing 174, 1101–1106.

Liu, Y., Fan, B., Wang, L., Bai, J., Xiang, S., Pan, C., 2018. Semantic labeling in very high resolution images via a self-cascaded convolutional neural network. ISPRS Journal of Photogrammetry and Remote Sensing 145, 78–95.

Liu, Y., Fan, B., Xiang, S., Pan, C., 2019. Relation-shape convolutional neural network for point cloud analysis, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Liu, Y., Li, X., 2014. Domain adaptation for land use classification: A spatio-temporal knowledge reusing method. ISPRS journal of photogrammetry and remote sensing 98, 133–144.

Luo, H., Khoshelham, K., Fang, L., Chen, C., 2020. Unsupervised scene adaptation for semantic segmentation of urban mobile laser scanning point clouds. ISPRS Journal of Photogrammetry and Remote Sensing 169, 253–267.

Luo, Y., Zheng, L., Guan, T., Yu, J., Yang, Y., 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2507–2516.

Luo, Z., Branchaud-Charron, F., Lemaire, C., Konrad, J., Li, S., Mishra, A., Achkar, A., Eichel, J., Jodoin, P.M., 2018. Mio-tcd: A new benchmark dataset for vehicle classification and localization. IEEE Trans. Image Process. 27, 5129–5141.

Meyer, G.P., Charland, J., Hegde, D., Laddha, A., Vallespi-Gonzalez, C., 2019. Sensor fusion for joint 3d object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops.

Mi, L., Chen, Z., 2020. Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation. ISPRS Journal of Photogrammetry and Remote Sensing 159, 140–152.

Michieli, U., Biasetton, M., Agresti, G., Zanuttigh, P., 2020. Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation. IEEE Transactions on Intelligent Vehicles 5, 508–518.

Morerio, P., Cavazza, J., Murino, V., 2017. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. arXiv preprint arXiv:1711.10288.

Murez, Z., Kolouri, S., Kriegman, D., Ramamoorthi, R., Kim, K., 2018. Image to image translation for domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4500–4509.

Pan, F., Shin, I., Rameau, F., Lee, S., Kweon, I.S., 2020. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3764–3773.

Qin, C., You, H., Wang, L., Kuo, C.C.J., Fu, Y., 2019. Pointdan: A multi-scale 3d domain adaption network for point cloud representation. In: Advances in Neural Information Processing Systems, pp. 7192–7203.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241.

Serna, A., Marcotegui, B., 2014. Detection, segmentation and classification of 3d urban objects using mathematical morphology and supervised learning. ISPRS Journal of Photogrammetry and Remote Sensing 93, 243–255.

Su, H., Jampani, V., Sun, D., Maji, S., Kalogerakis, E., Yang, M.H., Kautz, J., 2018. Splatnet: Sparse lattice networks for point cloud processing. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2530–2539.

Toldo, M., Maracani, A., Michieli, U., Zanuttigh, P., 2020. Unsupervised domain adaptation in semantic segmentation: a review. Technologies 8, 35.

Tsai, Y.H., Hung, W.C., Schulter, S., Sohn, K., Yang, M.H., Chandraker, M., 2018. Learning to adapt structured output space for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7472–7481.

Tsai, Y.H., Sohn, K., Schulter, S., Chandraker, M., 2019. Domain adaptation for structured output via discriminative patch representations. In: IEEE International Conference on Computer Vision, pp. 1456–1465.

Tzeng, E., Hoffman, J., Saenko, K., Darrell, T., 2017. Adversarial discriminative domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 7167–7176.

Valada, A., Mohan, R., Burgard, W., 2019. Self-supervised model adaptation for multimodal semantic segmentation. Int. J. Comput. Vision 1–47.

Volpi, R., Morerio, P., Savarese, S., Murino, V., 2018. Adversarial feature augmentation for unsupervised domain adaptation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 5495–5504.

Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P., 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2517–2526.

Vu, T.H., Jain, H., Bucher, M., Cord, M., Pérez, P., 2019. Dada: Depth-aware domain adaptation in semantic segmentation. In: IEEE International Conference on Computer Vision, pp. 7364–7373.

Wang, L., Huang, Y., Hou, Y., Zhang, S., Shan, J., 2019. Graph attention convolution for point cloud semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 10296–10305.

Wilson, G., Cook, D.J., 2020. A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology (TIST) 11, 1–46.

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

Wu, B., Zhou, X., Zhao, S., Yue, X., Keutzer, K., 2019. Squeezesegv 2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In: IEEE International Conference on Robotics and Automation, pp. 4376–4382.

Yan, L., Fan, B., Liu, H., Huo, C., Xiang, S., Pan, C., 2019. Triplet adversarial domain adaptation for pixel-level classification of vhr remote sensing images. IEEE Trans. Geosci. Remote Sens. 58, 3558–3573.

Zhang, Y., Qiu, Z., Yao, T., Liu, D., Mei, T., 2018. Fully convolutional adaptation networks for semantic segmentation, in. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6810–6818.

Zhang, Z., Doi, K., Iwasaki, A., Xu, G., 2020. Unsupervised domain adaptation of high-resolution aerial images via correlation alignment and self training. IEEE Geosci. Remote Sens. Lett.

Zhao, S., Li, B., Yue, X., Gu, Y., Xu, P., Hu, R., Chai, H., Keutzer, K., 2019. Multi-source domain adaptation for semantic segmentation. In: Advances in Neural Information Processing Systems, pp. 7287–7300.

Zhao, S., Wang, G., Zhang, S., Gu, Y., Li, Y., Song, Z., Xu, P., Hu, R., Chai, H., Keutzer, K., 2019b. Multi-source distilling domain adaptation. arXiv preprint arXiv:1911.11554.

Zhu, R., Yan, L., Mo, N., Liu, Y., 2019. Semi-supervised center-based discriminative adversarial learning for cross-domain scene-level land-cover classification of aerial images. ISPRS Journal of Photogrammetry and Remote Sensing 155, 72–89.

Zhu, X., Zhou, H., Yang, C., Shi, J., Lin, D., 2018. Penalizing top performers: Conservative loss for semantic segmentation adaptation, in. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 568–583.

Zou, Y., Yu, Z., Vijaya Kumar, B., Wang, J., 2018. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: European Conference on Computer Vision, pp. 289–305.