Cross-modal Unsupervised Domain Adaptation for 3D Semantic Segmentation via Bidirectional Fusion-then-Distillation

Yao Wu*
School of Informatics
Xiamen University, Xiamen, China

Mingwei Xing*
Institute of Artificial Intelligence
Xiamen University, Xiamen, China

Yachao Zhang
Tsinghua University, Shenzhen, China

Yuan Xie†
East China Normal University,
Shanghai, China
Chongqing Institute of East China
Normal University, Chongqing, China

Jianping Fan
Zhongchao Shi
Lenovo Research, Beijing, China

Yanyun Qu†
Institute of Artificial Intelligence,
Xiamen University, Xiamen, China
yyqu@xmu.edu.cn

ABSTRACT

Cross-modal Unsupervised Domain Adaptation (UDA) becomes a research hotspot because it reduces the laborious annotation of target domain samples. Existing methods only mutually mimic the outputs of cross-modality in each domain, which enforces the class probability distribution agreeable in different domains. However, these methods ignore the complementarity brought by the modality fusion representation in cross-modal learning. In this paper, we propose a cross-modal UDA method for 3D semantic segmentation via Bidirectional Fusion-then-Distillation, named BFtD-xMUDA, which explores cross-modal fusion in UDA and realizes distribution consistency between outputs of two domains not only for 2D image and 3D point cloud but also for 2D/3D and fusion. Our method contains three significant components: Model-agnostic Feature Fusion Module (MFFM), Bidirectional Distillation (B-Distill), and Cross-modal Debiased Pseudo-Labeling (xDPL). MFFM is employed to generate cross-modal fusion features for establishing a latent space, which enforces maximum correlation and complementarity between two heterogeneous modalities. B-Distill is introduced to exploit bidirectional knowledge distillation which includes cross-modality and cross-domain fusion distillation, and well-achieving domain-modality alignment. xDPL is designed to model the uncertainty of pseudo-labels by a self-training scheme. Extensive experimental results demonstrate that our method outperforms state-of-the-art competitors in several adaptation scenarios.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**.

KEYWORDS

3D semantic segmentation, Unsupervised domain adaptation

*Equal contribution, †Corresponding authors

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612013

ACM Reference Format:
Yao Wu*, Mingwei Xing*, Yachao Zhang, Yuan Xie†, Jianping Fan, Zhongchao Shi, and Yanyun Qu†. 2023. Cross-modal Unsupervised Domain Adaptation for 3D Semantic Segmentation via Bidirectional Fusion-then-Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3581783.3612013

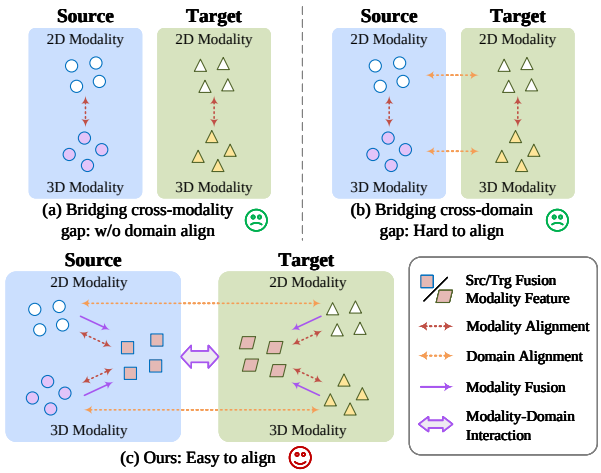


Figure 1: The comparison of three types of cross-modal UDA methods. Specifically, our approach explores whether and how Bidirectional Fusion-then-Distillation facilitates it.

1 INTRODUCTION

Point cloud which can be generated from LiDAR provides a powerful way to perceive, understand, and reconstruct the complex 3D visual world. In particular, 3D semantic segmentation is a critical task that enables applications like autonomous driving [4, 33, 44, 51]. Similar to other computer-vision tasks, 3D semantic segmentation faces the problem of domain shift which results in performance degradation on a new unlabeled dataset (target-domain) with a different distribution from the labeled training dataset (source-domain). For instance, the 3D segmentation model learned on point clouds collected in the USA usually performs terribly in the Singapore [20]. Annotating large-scale datasets for every new scenario

is a solution, but it is impracticable due to laborious and time-consuming human operations, especially for the tasks demanding point-wise annotations.

Unsupervised Domain Adaptation (UDA) aims at solving the aforementioned domain shift problem. However, existing UDA methods are mostly concerned with the uni-modality, *e.g.*, 2D UDA [16, 17, 46] and 3D UDA [32, 41]. Cross-modal UDA is under-explored, which is in its infancy. The latest cross-modal UDA methods [20, 27, 47] usually reinforce the contextual information in 2D images by leveraging the geometry of 3D point clouds, and vice versa, to mitigate the discrepancy in data distribution between the source domain and target domain.

The existing cross-modal UDA methods are divided into two types: (a) bridging the cross-modality gap and (b) bridging the cross-domain gap. The former tackles cross-modal UDA by mutually mimicking the output of each modality [20], disregarding the domain gap (see Fig. 1(a)). Based on the methods bridging the cross-modality gap, the latter further learns domain-invariant features [22, 24] and modality-specific information to complement each other [27, 47], reducing the domain gap (see Fig. 1(b)). However, in cross-modal learning, these methods only use one modality to mimic the other. The limitation is that it may bring negative mimicking due to *imbalanced modality adaptability* in different domains. For instance, a 2D image is beneficial to segment objects on days but loses plentiful context information on nights due to the dark light. In this condition, allowing 3D representations to mimic 2D representations potentially degrades the performance.

It is proved that the cross-modality fusion combining the exclusive advantages of 2D images and 3D point clouds is beneficial for 3D scene understanding. In practice, on the one hand, images provide dense, textured, and colored information, which is beneficial to 3D representations. On the other hand, point clouds with depth information can provide spatial geometric structure, which supplements the spatial perception for 2D representations. Therefore, we focus on leveraging the complementary advantage of cross-modality fusion to solve the negative mimicking in cross-modality UDA. There exist two challenging issues: 1) how to conduct cross-modality fusion of two heterogeneous modalities for 3D semantic segmentation and 2) how to make domain-modality alignment to bridge the cross-modality and cross-domain gaps based on the cross-modality fusion representation.

As for the first issue, considering that 2D images and 3D point clouds are heterogeneous, our goal is to achieve feature-level fusion, while discarding data-level and decision-level fusion. Based on feature-level fusion, we enforce maximum correlation and complementarity between two heterogeneous modalities for performing 3D semantic segmentation. As for the second issue, we consider alignment using bidirectional knowledge distillation to alleviate distribution discrepancy between different modalities and domains. Not only in the uni-modality space, such as images and point clouds in different domains, but also in the latent space formed by cross-modality fusion features, which establishes bidirectional knowledge distillation, that is, "2D/3D-to-Fusion" for each domain, and "source 2D/3D-to-stylized Fusion" for cross-domain. Note that, "stylized" denotes source-domain data in target style.

Accordingly, in this paper, we propose a cross-modal unsupervised domain adaptation approach for 3D semantic segmentation via Bidirectional Fusion-then-Distillation, named **BFtD-xMUDA**, with the goal of making the model more robust to the change of distributions from the source domain to the target domain. As shown in Fig. 1(c), our method enforces the class probability distribution consistency between outputs of two domains not only in the 2D image and 3D point cloud space, respectively, but also in the 2D/3D fusion space. Our method contains three significant components: Model-agnostic Feature Fusion Module (MFFM), Bidirectional Distillation (B-Distill), and Cross-modal Debiased Pseudo-Labeling (xDPL). To be specific, MFFM is designed as a plug-and-play feature fusion module, which fuses the features of images and point clouds. Considering large-scale images and point clouds, features are usually refined only by calculating the self-affinities within a sample and ignoring the pair-wise relationship. Therefore, a novel memorized modality attention block is designed to learn robust feature representations. MFFM not only can mine external affinities globally in the whole domain but also reduce computational complexity.

Furthermore, B-Distill is designed to alleviate cross-modality and cross-domain distribution discrepancy, which are the most crucial ingredients in our method. It contains two learning mechanisms: *i.e.*, Modality-Preserving Distillation (MPD), and Domain-Preserving Distillation (DPD). MPD bridges the cross-modality gap in each domain by leveraging the knowledge distillation to make the output of 2D/3D consistent with Fusion, relieving the class probability distribution discrepancy of cross-modality. DPD bridges the cross-domain gap from the source to the stylized domain by leveraging fusion knowledge distillation, relieving the class probability distribution discrepancy of cross-domain. Moreover, to take advantage of the unlabeled target-domain data for bridging the cross-domain gap, xDPL is designed to model the uncertainty of the pseudo-label via the multi-modal prediction variance, effectively exploiting the domain-specific information offered by pseudo-labels.

To summarize, the following are the main contributions:

- We present BFtD-xMUDA, a cross-modal UDA approach for 3D semantic segmentation, where a Bidirectional Fusion-then-Distillation mechanism is introduced. It utilizes knowledge distillation to explore the complementary advantage of cross-modality fusion representation in bridging the cross-modality and cross-domain gaps.
- MFFM in low computation complexity is designed, which conducts cross-modality fusion from two heterogeneous modalities for 3D semantic segmentation.
- B-Distill is proposed to conduct bidirectional fusion knowledge distillation. It makes the cross-modality and cross-domain alignments based on the cross-modality fusion representation.
- Extensive experimental results demonstrate that our method outperforms state-of-the-art (SOTA) competitors in several adaptation scenarios.

2 RELATED WORK

2.1 3D Semantic Segmentation

The mainstream methods to process 3D semantic segmentation are divided into four categories: Point-based methods, Projection-based methods, Voxel-based methods, and Multi-representation methods.

In the point-based methods, PointNet [28, 29] is a pioneering work that directly uses multi-layer perceptron (MLPs) to learn the features from the unordered sequences of point clouds. Later on, based on PointNet, convolution operations are implemented on the point-wise features output by MLPs [19, 35, 40] that perform well on the synthetic point cloud [1, 7] rather than the sparse LiDAR point cloud [2, 3]. Projection-based methods project 3D point cloud to the 2D image space in the bird's-eye-view or range-view, achieving efficient 3D segmentation with 2D CNNs, such as SqueezeSeg series [38, 39, 42] and others [6, 26, 48]. Voxel-based methods [5, 11] adopt sparse voxels considering the balance between efficiency and effectiveness. Multi-representation methods [4, 15, 34, 43] use point-based and projection-based representations to potentially facilitate voxel features. However, these methods severely suffer from inferior segmentation on distant objects due to sparse laser information.

2.2 Multi-modality Fusion for 3D Semantic Segmentation

Image and point cloud are two heterogeneous modalities popularly used in 3D semantic segmentation for autonomous driving. Recently, many studies leverage multi-modality fusion to improve the performance [8, 9, 21, 44, 51]. Early works [8, 21] warp suboptimal RGB values or separately learned CNN features into point cloud range image as additional inputs. PMF [51] projects the point cloud onto the image plane for fusion with image data through perspective projection. Besides, 2D3DNet [9] trains a 3D model from pseudo-labels derived from 2D semantic segmentation by using multi-view fusion. 2DPASS [44] distills the knowledge of multi-modality representation to the feature representation of point cloud to enhance the 3D feature learning because multi-modality provides rich texture information and structural regularization. Although these methods have achieved astonishing performance on the condition that the distribution of the training and testing datasets are consistent, they may degrade significantly when the testing scenario is different from the training scenario. Moreover, such multi-modal training approaches lack joint 2D-3D optimization.

2.3 Unsupervised Domain Adaptation for 3D Semantic Segmentation

UDA methods for 3D semantic segmentation can use uni-modality cases [32, 41, 45, 49] and cross-modality cases [20, 22, 24, 27, 47]. For uni-modality, in [49], real dropout noise is simulated on synthetic data through a generative adversarial network. Similarly, in [41], domain shift is disentangled into appearance difference and sparsity difference, and a generative network is applied to mitigate each difference. Complete&Label [45] resolves the domain adaptation from the perspective of the 3D surface completion task. CosMix [32] crops counterpart patches by using the semantic information in the source point clouds and mixes it with the target point clouds to form the mixed domain for training, and vice versa. For cross-modality, xMUDA [20] first provides a cross-modal learning method for 3D semantic segmentation in UDA. SSE-xMUDA [47] presents a self-supervised exclusive learning mechanism to exploit the exclusive information of different modalities to complement each other. Dual-Cross [22] designs a multi-modal stylized transfer module to alleviate the domain shift problem. [24, 27] exploit adversarial learning method to learn the domain-invariant representations. Differently, we focus on how to utilize the complementary advantage of cross-modality fusion to alleviate the imbalanced modality adaptability in cross-modal learning and benefit 3D scene segmentation in UDA.

3 PROPOSED METHOD

3.1 Problem Definition

We define input source samples $\{X_S^{2D}, X_S^{3D}, Y_S^{3D}\} \in \mathcal{S}$ and target samples $\{X_T^{2D}, X_T^{3D}\} \in \mathcal{T}$, where X^{2D} represents the image and X^{3D} represents the corresponding point cloud, with 3D points in the camera reference frame. Only the source domain has annotations Y_S^{3D} for each 3D point. Note that X^{3D} contains only points visible from the RGB camera, assuming that the calibration of the LiDAR and Camera is available for both domains and does not change over time. Given the samples of source domain $\mathcal{S}$ and target domain $\mathcal{T}$, we intend to learn a function $f : \mathcal{S} \cup \mathcal{T} \mapsto Y_T^{3D}$ for 3D semantic segmentation in the target domain.

3.2 Overview

For cross-modal UDA tasks, we consider transferring exclusive modality information from dense images and sparse point clouds to the cross-modality fusion representations. The overall framework of BFtD-xMUDA is illustrated in Fig. 2, which contains three streamlines: 2D, 3D, and Fusion streamlines. The two learning mechanisms, MPD and DPD, form Bidirectional Distillation, which bridges the cross-modality and cross-domain gaps.

Concretely, before MPD, images X_S^{2D}, X_T^{2D} and point clouds X_S^{3D}, X_T^{3D} are fed into parallel 2D and 3D streamlines respectively, which output image features and voxel features. Like xMUDA [20], only the points falling into the intersected field of view are geometrically associated with multi-modal data (*i.e.*, 2D-to-3D projection). Accordingly, in the source and target domains, we can obtain point-wise 2D features $F_S^{2D}, F_T^{2D} \in \mathbb{R}^{N \times D_1}$ and point-wise 3D features $F_S^{3D}, F_T^{3D} \in \mathbb{R}^{N \times D_2}$. Both of them are used for cross-modality feature fusion between images and point clouds in MFFM, where N is the number of points projected onto the image plane, D_1 and D_2 denote the number of channels of the feature maps output from 2D and 3D backbone, respectively. Before DPD, regarding the stylized representation as domain-related representation, we utilize MMST [22] to generate stylized samples X_{ST}^{2D} and X_{ST}^{3D}. Afterward, these stylized samples are fed into the parallel 2D and 3D streamlines. Similar to MPD, the output from 2D and 3D streamlines are fused in MFFM and the cross-modality fusion features of source samples in the target style are generated. Of note, Student and Teacher share the same 2D and 3D backbones, and the parameters of 2D and 3D backbones in Teacher are updated to the counterparts in Student via Exponential Moving Average (EMA), to obtain the stylized 2D and 3D features, *i.e.*, F_{ST}^{2D} and F_{ST}^{3D}.

After extracting point-wise 2D, 3D, and fusion features, both modality-specific features and fusion features are learned by the B-Distill mechanism to achieve cross-modality and cross-domain alignments. Besides, to take advantage of the unlabeled target-domain data, we introduce a simple yet effective self-training scheme,

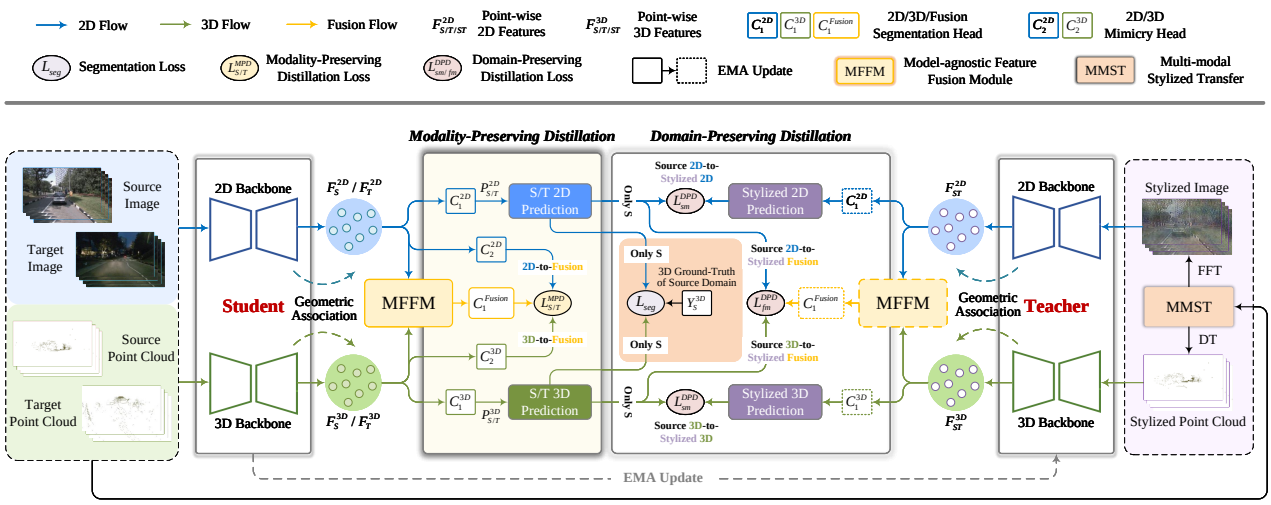


Figure 2: Overview framework of BFtD-xMUDA, which contains Modality-Preserving Distillation (MPD) for cross-modal learning and Domain-Preserving Distillation (DPD) for cross-domain learning.

named Cross-modal Debiased Pseudo-Labeling (xDPL), training again from scratch using the produced pseudo-labels for an additional segmentation loss on the target-domain training set.

3.3 Model-agnostic Feature Fusion Module

A key challenge in cross-modality fusion is how to effectively maximize the correlation and complementarity of 2D and 3D features. To solve this problem, a Model-agnostic Feature Fusion Module (MFFM) is introduced to generate cross-modality fusion features via a self-attention mechanism.

As shown in Fig. 2, there are two inputs of MFFM: 2D features $F_S^{2D}, F_T^{2D} \in \mathbb{R}^{N\times D_1}$ and 3D features $F_S^{3D}, F_T^{3D} \in \mathbb{R}^{N\times D_2}$. If global self-attention is adopted in cross-modality fusion, only concentrates on the self-affinities between different locations within a single sample, whose computational cost is intensive with the complexity of $O(DN^2)$, D is the feature dimension. Thus, it is infeasible to directly apply self-attention to point-wise features.

Therefore, a novel memorized modality attention module is meticulously designed to mine external affinities of multi-modality from whole domain samples. Global self-attention ignores the relationship between samples in different batches. Unlike global self-attention whose key and value matrices are generated from a linear projection of input, inspired by [12], our module uses the learnable key and value units $(M_k^{(\cdot)}, M_v^{(\cdot)})$ and then save them in the memory bank. If multiplying $M_k^{(\cdot)}$, it can learn the potential affinity between N query elements and K memorized key elements. If multiplying $M_v^{(\cdot)}$, it updates the input features from $M_v^{(\cdot)}$ by the similarities in attention map. This mechanism not only mines external affinities from whole domain samples but also reduces computational complexity from $O(DN^2)$ to $O(DNK)$ due to the small neighbors but the wider scope of attention map $A^{(\cdot)} \in \mathbb{R}^{N\times K}$.

Take the cross-modality fusion feature in the source domain **S** as an example, we introduce the process of MFFM. As shown in Fig. 3, given a 3D feature F_S^{3D}, the attention map between F_S^{3D} and a 3D memorized key unit $M_k^{3D} \in \mathbb{R}^{K\times D_2}$ can be calculated via:

$$A^{3D} = \text{Norm}(F_S^{3D}(M_k^{3D})^T), \tag{1}$$

where $A^{3D} \in \mathbb{R}^{N\times K}$ is the attention matrix, K is the channel of 3D memorized key unit, $\text{Norm}(\cdot)$ is the normalization operation. To increase the capability of the network, we calculate the pairwise affinity between A^{3D} and a 3D memorized value unit $M_v^{3D} \in \mathbb{R}^{K\times D_2}$, and attach a skip-connect via:

$$\tilde{F}_S^{3D} = A^{3D}M_v^{3D} + F_S^{3D}, \tag{2}$$

where the refined 3D feature $\tilde{F}_S^{3D} \in \mathbb{R}^{N\times D_2}$ is strengthened. Of note, both of M_k^{3D} and M_v^{3D} are learnable.

Similar to $\tilde{F}_S^{3D}$, we can obtain the refined 2D feature $\tilde{F}_S^{2D} \in \mathbb{R}^{N\times D_1}$. As shown in Fig. 3, MFFM concatenates $\tilde{F}_S^{2D}$ and $\tilde{F}_S^{3D}$, and feeds them into the submodule "fusion learner" $\Phi(\cdot)$, which contains a layer of MLP, struggling to narrow the gap between 2D and 3D modalities. The fusion feature $F_S^{Fusion} \in \mathbb{R}^{N\times D_f}$ is obtained by:

$$F_S^{Fusion} = \Phi(\text{Concat}(\tilde{F}_S^{2D}, \tilde{F}_S^{3D})), \tag{3}$$

where $\text{Concat}(\cdot,\cdot)$ concatenates two inputs along the feature dimension. External attention has linear complexity and implicitly considers the correlation between all data samples. However, it overlooks the correlation between different modalities. Therefore, we aim to learn exclusive and discriminative features from each modality, capturing the most informative parts and excluding interference from other samples to the fusion modality. Specifically, after calculating F_S^{Fusion}, the fusion attention module calculates external attention between the fusion representation and the 2D and 3D memory units separately. Ultimately, the output refined fusion feature $\tilde{F}_S^{Fusion}$ is obtained by:

$$A^{Fusion} = \text{Norm}(F_S^{Fusion}(M_k^{2D})^T + F_S^{Fusion}(M_k^{3D})^T), \tag{4}$$

$$\tilde{F}_S^{Fusion} = A^{Fusion}M_v^{2D} + A^{Fusion}M_v^{3D} + F_S^{Fusion}. \tag{5}$$

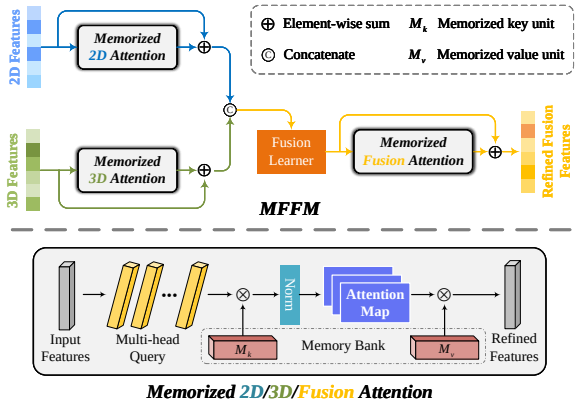


Figure 3: The diagram of MFFM with memorized modality attention module.

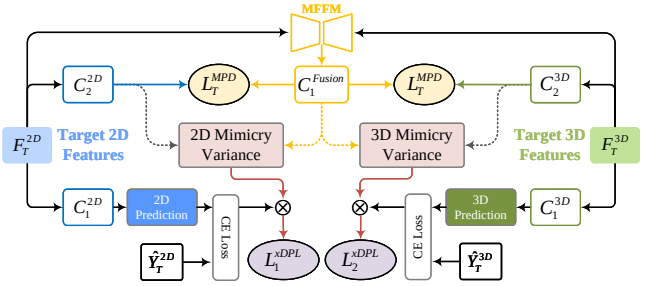


Figure 4: The diagram of xDPL for the target domain in the self-training stage.

In the same way, we can generate refined fusion feature $\tilde{F}_T^{Fusion}$ of the target domain according to MFFM.

3.4 Bidirectional Distillation

In this subsection, we introduce Bidirectional Distillation, which contains two learning mechanisms: Modality-Preserving Distillation and Domain-Preserving Distillation. Both of them aim at improving the domain-modality alignment using the complementary advantage of cross-modality fusion through establishing bidirectional knowledge distillation, which we present below.

Modality-Preserving Distillation (MPD). According to MFFM, for cross-modal learning, we establish a mimicking game between the 2D/3D and fusion streamlines, that is, the output of the mimicry head of the 2D/3D streamline should be consistent with that of the fusion streamline, toward an agreement in latent space.

To be specific, as depicted in Fig. 2, the 2D and 3D streamlines both have two segmentation heads: the main head C_1^{2D}/C_1^{3D} for the best 2D/3D predictions, and the mimicry head C_2^{2D}/C_2^{3D} to estimate the output of MFFM with the main head C_1^{Fusion}. Note that, the mimicry head has the same structure as the main head, but the parameters are updated differently. Then, we conduct class probability distribution alignment between the 2D/3D predictions and cross-modality fusion predictions, *i.e.*, "2D-to-Fusion" and "3D-to-Fusion". We choose the Kullback-Leibler divergence $D_{KL}(\cdot||\cdot)$ for the MPD loss and define it as follows:

$$\mathcal{L}_S^{MPD} = D_{KL}(P_S^{Fusion}||P_S^{2D,m}) + D_{KL}(P_S^{Fusion}||P_S^{3D,m}), \quad (6)$$

$$\mathcal{L}_T^{MPD} = D_{KL}(P_T^{Fusion}||P_T^{2D,m}) + D_{KL}(P_T^{Fusion}||P_T^{3D,m}), \quad (7)$$

where the superscript "m" refers to the mimicry head, P_S^{Fusion} and P_T^{Fusion} denote the prediction from C_1^{Fusion} in the source and target domain, respectively.

Domain-Preserving Distillation (DPD). As depicted in Fig. 2, DPD is designed to conduct cross-domain alignment by knowledge distillation to relieve the class probability distribution discrepancy between the source domain and the target domain.

Firstly, we adopt the multi-modal style transfer module (MMST) to make Teacher be aware of the target style. To be specific, the stylized image combines the low frequency of the target sample and the high frequency of the source sample, while the stylized point cloud converts the density of the source point cloud to that of the target point cloud (*i.e.*, X_{ST}^{2D} and X_{ST}^{3D}, source data in target style), which can be defined as:

$$X_{ST}^{2D} = \mathcal{F}^{-1}(\mathcal{F}^{low}(X_T^{2D}) + \mathcal{F}^{high}(X_S^{2D})), \quad (8)$$

$$X_{ST}^{3D} = \sqrt[3]{\frac{\mathcal{D}(X_T^{3D})}{\mathcal{D}(X_S^{3D})}} X_S^{3D}, \quad (9)$$

where $\mathcal{F}^{low}(\cdot)$ and $\mathcal{F}^{high}(\cdot)$ denote the low-frequency and high-frequency components of the Fourier Transform $\mathcal{F}$ of an RGB image, respectively, and $\mathcal{F}^{-1}$ is the inverse of $\mathcal{F}$. $\mathcal{D}(\cdot) = (X_{max} - X_{min})(Y_{max} - Y_{min})(Z_{max} - Z_{min})/N$ denotes the density of one point cloud. $X/Y/Z_{max}$ is the maximum coordinate value while $X/Y/Z_{min}$ is the minimum.

After that, we directly enforce the probability distributions obtained by Student and Teacher to be consistent under the "Source-to-Stylized" alignment. We define the DPD loss in the single 2D/3D modality as the KL divergence:

$$\mathcal{L}_{sm}^{DPD} = D_{KL}(P_{ST}^{2D}||P_S^{2D}) + D_{KL}(P_{ST}^{3D}||P_S^{3D}), \quad (10)$$

where P_S^{2D} and P_S^{3D} are the prediction outputs of source data. P_{ST}^{2D} and P_{ST}^{3D} are the prediction outputs of source data in the target style. In the source domain, Teacher transfers the target-style information to Student. In this way, we can make Student perceive the target domain when it is trained with source data, improving its perception of the target domain.

Different from [22], which uses hybrid-modal prediction by taking the mean value of Teacher predictions of 2D and 3D modalities, we utilize the EMA-MFFM (*i.e.*, MFFM in DPD, whose parameters are transferred from the MFFM in MPD via EMA) to generate hybrid-modal prediction output P_{ST}^{Fusion}. Similar to $\mathcal{L}_{sm}^{DPD}$, we conduct "Source 2D/3D to Stylized Fusion" alignment through DPD loss in the fusion modality and define it as the KL divergence:

$$\mathcal{L}_{fm}^{DPD} = D_{KL}(P_{ST}^{Fusion}||P_S^{2D}) + D_{KL}(P_{ST}^{Fusion}||P_S^{3D}). \quad (11)$$

In this way, BFtD-xMUDA not only preserves the complete information from multi-modal data but also combines the semantic content of the source domain with the style of the target domain, well-conducting domain-modality alignment.

3.5 Cross-modal Debiased Pseudo-Labeling

Pseudo-label learning is a frequently used technique in UDA, which leverages the pseudo-label to learn from the unlabeled target data. To address the noise of pseudo-labels, Cross-modal Debiased Pseudo-Labeling (xDPL) is introduced to substitute the general pseudo-label scheme, which models the uncertainty of the pseudo-label via the multi-modal prediction variance.

Following [50], we use variance regularization term to rectify the learning from noisy labels, preventing that the model predicts the large variance all the time. Diversely, as shown in Fig. 4, we utilize the mimicry head of the 2D/3D streamline and the main head of the fusion streamline to construct multi-modal prediction variance, which we call 2D/3D mimicry variance. If two heads provide two different class predictions, the approximated variance will obtain a large value. It reflects the uncertainty of the model on the target prediction.

To be specific, according to Eq. (7), KL divergence of two heads predictions also can be regarded as the variance. Therefore, we introduce the variance regularization term to rectify the learning from noisy labels. Then, we combine the prediction variance with the cross-entropy loss on pseudo supervision to obtain the rectified unsupervised loss of 2D and 3D terms, which can be formulated as:

$$\mathcal{L}_1^{xDPL} = exp\big(-D_{KL}(P_T^{Fusion}||P_T^{2D,m})\big) \cdot \mathcal{L}_{CE}(P_T^{2D}, \hat{Y}_T^{2D}), \quad (12)$$

$$\mathcal{L}_2^{xDPL} = exp\big(-D_{KL}(P_T^{Fusion}||P_T^{3D,m})\big) \cdot \mathcal{L}_{CE}(P_T^{3D}, \hat{Y}_T^{3D}), \quad (13)$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss, $\hat{Y}_T^{2D}$ and $\hat{Y}_T^{3D}$ are the pseudo-labels of target sample generated by the first training stage. For the region of ambiguous predictions, the 2D/3D mimicry variance makes the model neglect pseudo-labels, thus with fewer outliers. In other words, xDPL could provide dynamic and more accurate thresholds, concentrating the segmentation model on learning reliable pixels and points.

3.6 Overall Loss

The point-wise supervised segmentation loss of the source domain is formulated as follows:

$$\mathcal{L}_{seg} = -\frac{1}{N \times C} \sum_{n=1}^{N} \sum_{c=1}^{C} Y_S^{(n,c)} \log P_S^{(n,c)}, \quad (14)$$

where Y_S is the 3D ground-truth label of source domain, *i.e.*, Y_S^{3D}, P_S is either P_S^{2D} or P_S^{3D} or P_S^{Fusion}, C is the number of categories.

Finally, the overall loss function is the sum of all aforementioned loss terms, which can be defined as:

$$\begin{aligned}\mathcal{L}_{all} =& \mathcal{L}_{seg} + \lambda_1 \mathcal{L}_S^{MPD} + \lambda_2 \mathcal{L}_T^{MPD} + \lambda_3 \mathcal{L}_{sm}^{DPD} + \lambda_4 \mathcal{L}_{fm}^{DPD} + \\ & \lambda_5 (\mathcal{L}_1^{xDPL} + \mathcal{L}_2^{xDPL}),\end{aligned} \quad (15)$$

where $\{\lambda_i\}_{i=1}^5$ are weights trading off in the losses of MPD, DPD, and xDPL, respectively.

4 EXPERIMENT

4.1 Datasets

For evaluation, we use three public autonomous driving datasets, including nuScenes [3], A2D2 [10], and SemanticKITTI [2]. Following [20], we utilize the accessible 3D bounding boxes annotations to obtain the 3D point-wise labels. More specifically, we assign point-wise labels based on whether the points are located in a specific 3D bounding box and the points outside the box are labeled as background. To make a more convincing evaluation, following [27], we also experiment on nuScenes-Lidarseg [3] ("Seg" for short) which contains the point-wise annotation. Of note, for all datasets, LiDAR and RGB cameras are synchronized and calibrated, allowing 2D-to-3D projection. Furthermore, following [20], we only use the front camera image and the corresponding LiDAR points.

Our experimental scenarios cover typical real-to-real domain adaptation challenges like changes in lighting (*i.e.*, Day→Night of nuScenes), scene layout (*i.e.*, USA→Singapore of nuScenes, driving between right and left-hand-side), and sensor setups (*i.e.*, A2D2→SemanticKITTI, different resolution or FoV). For the first two scenarios, we choose 4 merged classes with background, while for the last scenario, we select 10 shared classes between two datasets. Of note, the source and target classes are coincident.

4.2 Implementation Details

For the 2D backbone, we use a modified version of U-Net [30], which consists of ResNet-34 [13] pre-trained on ImageNet [31] as the encoder and transposed convolutions and skip connections as the decoder. For the 3D backbone, we use the official SparseConvNet [11] implementation. The voxel size is set to 5cm which is small enough to only have one 3D point per voxel.

We employ standard 2D/3D data augmentation and log-smoothed class weights on point-wise supervised segmentation loss to address the class imbalance. The batch size is set to 8 and Adam optimizer with β_1=0.9, β_2=0.999 is employed. The decay rate of the moving average is set to 0.999. Our model is trained with 100k iterations for all scenarios. We utilize an iteration-based learning schedule where the initial learning rate is 0.001 and then it is divided by 10 at 80k and 90k iterations. Furthermore, we implement parameter-sensitive analysis experiments, λ_1 and λ_2 in MPD are set to 1.0 and 0.1, while λ_3 and λ_4 in DPD are set to 0.1 and 0.001, respectively. Note that we initially train the network without using xDPL loss, *i.e.*, $\lambda_5 = 0$, and in the self-training stage $\lambda_5 = 0.1$.

4.3 Quantitative and Qualitative Comparison

We compare our method with five typical uni-modal UDA methods, which can be easily extended to cross-modal UDA. Moreover, we compare our method with five cross-modal UDA methods. The comparison results for 3D semantic segmentation in mean Intersection over Union (mIoU) on the target testing data are shown in Tab. 1. Overall, BFtD-xMUDA performs best on almost all UDA scenarios against the compared methods, *w.r.t.* ensemble result "2D+3D".

The source-only model is the lower bound, which is not domain adaptation as it is only trained on the source-domain dataset of each scenario. It is observed that BFtD-xMUDA brings a significant adaptation effect on all scenarios compared to the source-only model, with the gains of +11.6%, +12.9%, +2.8%, +5.2%, and +9.3% in mIoU, respectively. Compared with the baseline (xMUDA), BFtD-xMUDA exceeds it by large margins with the gains of +9.2%, +10.3%, +1.3%, +3.5%, and +8.6% in mIoU. Compared with Dual-Cross [22], it is observed that BFtD-xMUDA achieves superior performance and achieves +1.7%, +5.9%, +0.7%, +3.0%, and +2.4% mIoU gains over all

Table 1: Quantitative results (mIoU, %) with both Uni-modal (Uni.) and Cross-modal (Cross.) adaptation methods for 3D semantic segmentation in different scenarios. The best value of "2D+3D" is marked in red, and the second best value is marked in blue. DsCML-ALL presents DsCML training with inter-domain cross-modal learning and fine-tuning with the pseudo labels (PL).

Modality	Method	Day→Night			Day→Night (Seg)			USA→Sing.			USA→Sing. (Seg)			A2D2→Sem.KITTI		
		2D	3D	**2D+3D**	2D	3D	**2D+3D**	2D	3D	**2D+3D**	2D	3D	**2D+3D**	2D	3D	**2D+3D**
	Source-only	41.8	41.4	47.6	41.8	43.8	48.0	53.2	46.8	61.2	53.3	48.0	61.6	36.4	37.3	42.2
Uni.	MinEnt [37]	44.9	43.5	51.3	44.9	44.3	51.8	53.4	47.0	59.7	53.6	48.6	61.9	38.8	38.0	42.7
	PL [23]	43.7	45.1	48.6	43.9	47.6	50.9	55.5	51.8	61.5	55.4	52.7	62.8	37.4	44.8	47.7
	CyCADA [14]	45.7	45.2	49.7	45.5	47.8	49.6	54.9	48.7	61.4	54.9	51.3	62.6	38.2	43.9	43.9
	AdaptSegNet [36]	45.3	44.6	49.6	45.5	45.3	49.3	56.3	47.7	61.8	56.5	49.0	62.0	38.8	44.3	44.2
	CLAN [25]	45.6	43.7	49.2	45.6	45.1	50.1	57.8	51.2	62.5	57.7	52.1	63.1	39.2	44.7	44.5
Cross.	xMUDA [20]	46.2	44.2	50.0	47.3	46.0	50.6	59.3	52.0	62.7	61.7	52.6	63.3	36.8	43.3	42.9
	AUDA [24]	49.0	47.6	54.2	-	-	-	59.8	52.0	63.1	-	-	-	43.0	43.6	46.8
	DsCML [27]	48.0	45.7	51.0	49.8	47.2	51.7	61.3	53.3	63.6	63.3	54.0	64.2	39.6	45.1	44.5
	Dual-Cross [22]	52.3	46.2	57.5	51.7	46.6	55.0	59.4	52.2	63.3	62.0	53.1	63.8	43.0	47.1	49.1
	SSE-xMUDA [47]	52.2	46.3	56.5	51.3	47.8	54.6	61.3	52.6	65.1	64.0	53.4	64.6	44.5	46.6	48.4
	BFtD-xMUDA	52.1	44.9	59.2	51.2	46.3	60.9	59.0	51.8	64.0	61.4	55.2	66.8	46.7	49.6	51.5
	xMUDA+PL [20]	47.1	46.7	50.8	48.4	47.5	51.2	61.1	54.1	63.2	63.0	54.3	64.2	43.7	48.5	49.1
	AUDA+PL [24]	48.7	46.2	55.7	-	-	-	59.7	51.7	63.0	-	-	-	43.3	43.3	47.3
	DsCML-ALL [27]	50.1	48.7	53.0	51.4	49.8	53.8	63.9	56.3	65.1	65.6	57.5	66.9	46.8	51.8	52.4
	Dual-Cross+PL [22]	53.6	46.8	58.2	52.1	48.2	56.2	61.5	54.7	63.8	64.1	55.7	65.6	46.3	52.1	54.1
	SSE-xMUDA+PL [47]	52.6	47.0	56.7	52.2	48.5	55.9	64.2	54.6	67.2	65.8	56.4	67.9	45.1	50.7	52.1
	BFtD-xMUDA+PL	51.8	48.7	60.1	51.5	49.4	61.4	60.1	54.0	65.3	63.5	57.0	68.0	47.0	51.4	54.4

scenarios. Furthermore, cross-modal learning and self-training with pseudo-labels (PL) are complementary in their combination. When re-trained with PL, our model achieves superior performance on all scenarios except for "USA→Sing." scenario. Of note, the results of cross-modal UDA methods evaluated on point-wise annotations with high accuracy are more convincing than those on 3D bounding boxes annotations.

Some qualitative segmentation results are presented in Fig. 6, showing the versatility of BFtD-xMUDA across all proposed UDA scenarios. For instance, we highlight the target "Bike" with green bounding boxes, where the riders are on the bicycles or motorcycles. It is observed that our method can predict them correctly from far and near, while the baseline (xMUDA) and Dual-Cross incorrectly predict them as "pedestrian" or "person".

4.4 Ablation Studies

In this subsection, we perform an in-depth analysis of BFtD-xMUDA with ablation studies on each component to highlight its strengths.

Effectiveness of MFFM. We conduct an ablation study on different modality attention modules of MFFM. As shown in Tab. 2, for a fair comparison, all the ablations are under three adaptation scenarios. First of all, we simply concatenate point-wise 2D features and 3D features and feed them into two-layer MLPs that consist of two successive linear layers, with a ReLU layer between them (#1, without any multi-modal self-attention mechanism). We then combine the MMST module to implement cross-modal learning as a baseline. "Channel att." means using the SENet [18] that is attached to the front of MLPs, which can learn the correlation between channels and screen out the attention to them (#2). This fusion strategy can bring a slight improvement on Day→Night (Seg) with the gain of +1.7%, but decrease the performance on the latter two adaptation scenarios. "Single att." (#3) is the MFFM without the memorized fusion attention block, which directly outputs cross-modality fusion features from the fusion learner. It outperforms the baseline with the gain of +4.7%, +0.5%, and +2.6% in mIoU. "Fusion att." (#4) is MFFM without memorized 2D and 3D attention blocks, which directly input concatenated 2D and 3D features into a fusion learner. It achieves the significant gains of +5.1%, +2.0%, and +2.5%. Finally, MFFM consists of #3 and #4, achieving the best results.

Table 2: Ablation study on MFFM. #x denotes the x-th row.

	MFFM	Day→Night (Seg)			USA→Sing. (Seg)			A2D2→Sem.KITTI		
		2D	3D	2D+3D	2D	3D	2D+3D	2D	3D	2D+3D
#1	w/o att.	49.7	44.1	54.9	60.2	52.2	63.5	43.6	46.8	48.1
#2	Channel att.	50.3	44.6	56.5	59.8	51.4	63.0	42.8	46.8	47.6
#3	Single att.	50.7	45.4	59.6	61.2	53.8	64.0	46.1	49.2	50.7
#4	Fusion att.	50.8	46.0	60.0	61.0	54.5	65.5	46.2	49.3	50.6
#5	#3 & #4	51.2	46.3	60.9	61.4	55.2	66.8	46.7	49.6	51.5

Effectiveness of B-Distill. Based on MFFM, we conduct an ablation study on B-Distill. As shown in Tab. 3, for a fair comparison, all the ablations are under three adaptation scenarios. Above all, we use xMUDA to train a model as our baseline, implementing cross-modal learning with naive KD (#1). Simply adding the baseline with MPD improves the baseline by +7.2%, +1.8%, and +4.6% in mIoU, which indicates that transferring exclusive modality information from dense images and sparse point clouds to the cross-modality fusion

Table 3: Ablation study on B-Distill.

	MPD	DPD	Day→Night (Seg)			USA→Sing. (Seg)			A2D2→Sem.KITTI		
			2D	3D	2D+3D	2D	3D	2D+3D	2D	3D	2D+3D
#1	xMUDA		47.3	46.0	50.6	61.7	52.6	63.3	36.8	43.3	42.9
#2	✓		50.4	45.7	57.8	61.3	54.5	65.1	45.2	46.8	47.5
#3		✓	50.1	44.6	57.4	61.1	55.0	64.5	45.8	47.0	49.3
#4	✓	✓	51.2	46.3	60.9	61.4	55.2	66.8	46.7	49.6	51.5

representations is beneficial to cross-modal learning. Meanwhile, adding DPD brings an improvement by +6.8%, +1.2%, and +3.5% in mIoU against the baseline. MPD along with DPD significantly improves segmentation performance, achieving SOTA results with mIoU of 60.9%, 66.8%, and 51.5%.

Fusion for source-domain adaptation. Fig. 5 shows that cross-modality fusion can benefit both Day→Night and USA→Sing. adaptation. Specifically, at the beginning of the training step, fusion predictions are much worse than "Avg", even than "2D". Until 5~10k iterations, "Fusion" starts to move up and diverges from other outcomes. Compared with "Avg" which achieves limited gains, our method can achieve increasing results progressively, up to the peak with 87.5%/90.6% mIoU. It shows that when existing *imbalanced modality adaptability* (*i.e.*, 3D is much weaker than 2D), simply taking the mean of the predicted 2D and 3D probabilities after SoftMax for cross-modal self-distillation also results in negative mimicking, potentially degrading the performance of ensemble results.

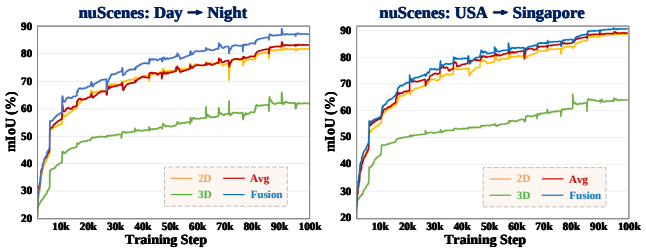


Figure 5: Segmentation results for 2D/3D/Avg/Fusion modality of source-domain samples at each training step.

Average and Fusion Ensemble for inference. Interestingly, it is observed that our method improves on difficult classes as in the case of "pedestrian", "bike", and "traffic boundary", whose performance is low before credible adaptation (see Tab. 4). Compared with xMUDA, Dual-Cross lifts limited performance for these distant objects, which are easily disturbed by the background sundries. In BFtD-xMUDA w/ Avg, we apply MFFM to the training stage but remove it away to take the mean of the predicted 2D and 3D probabilities after SoftMax in the inference stage. It still outperforms Dual-Cross, demonstrating that B-Distill can drive maximum correlation and complementarity between heterogeneous modalities. In BFtD-xMUDA w/ Fusion, we further leverage the predicted probabilities output from MFFM in the inference stage, the experimental results on per-class proving that the powerful performance of BFtD-xMUDA (*e.g.*, 30.4% vs 10.4% on "bike" and 54.5% vs 47.6% on "traffic boundary").

Table 4: Per-class IoU (%) performance on the test dataset of Day→Night (Seg), where † denotes the reproduced results.

Method	Day→Night (Seg)					
	back.	vehicle	pede.	bike	traffic.	mIoU
xMUDA† [20]	99.3	83.5	22.4	6.2	46.2	51.5
Dual-Cross† [22]	99.4	88.6	28.5	10.4	47.6	54.9
Improvement	+0.1	+5.1	+6.1	+4.2	+1.4	+3.4
BFtD-xMUDA w/ Avg	99.4	88.3	27.3	12.1	54.4	56.3
Improvement	+0.1	+4.8	+4.9	+5.9	+8.2	+4.8
BFtD-xMUDA w/ Fusion	**99.4**	**88.6**	**31.4**	**30.4**	**54.5**	**60.9**
Improvement	+0.1	+5.1	+9.0	+24.2	+8.3	+9.4

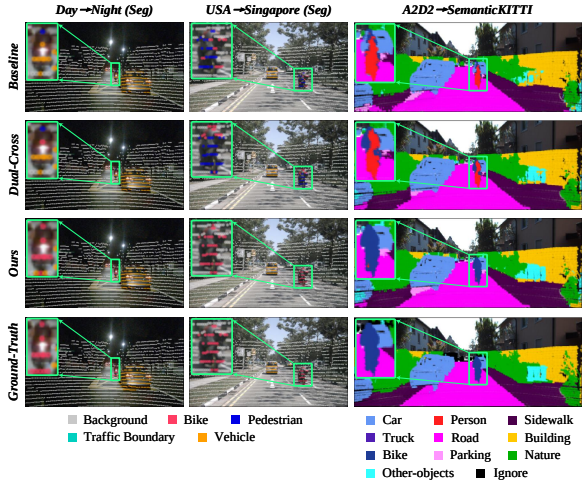


Figure 6: Qualitative results on three adaptation scenarios.

5 CONCLUSION

In this paper, we present BFtD-xMUDA, a cross-modal UDA method for 3D semantic segmentation. Compared with existing methods, BFtD-xMUDA leverages the complementary advantage of cross-modality fusion to solve the negative mimicking in cross-modal UDA. Firstly, MFFM in low computation complexity is designed, which conducts cross-modality fusion from two heterogeneous modalities. B-Distill is proposed to conduct bidirectional fusion knowledge distillation. It makes the cross-modality and cross-domain alignments based on the cross-modality fusion representation. The proposed xDPL alleviates the negative impact of noisy pseudo-labels to some extent. Technically, our method can also be generalized to domain generalization task, which will be explored in future work.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China No.2020AAA0108301; National Natural Science Foundation of China under Grant No.62176224, No.62222602; Natural Science Foundation of Chongqing under No.CSTB2023NSCQ-JQX0007; China Postdoctoral Science Foundation under No.2023M7 31957; CCF-Lenovo Blue Ocean Research Fund. Yuan Xie was supported by CAAI-Huawei MindSpore Open Fund.

REFERENCES

[1] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 2016. 3D semantic parsing of large-scale indoor spaces. In *CVPR*. 1534–1543.

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. 2019. SemanticKITTI: A dataset for semantic scene understanding of lidar sequences. In *ICCV*. 9297–9307.

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*. 11621–11631.

[4] Ran Cheng, Ryan Razani, Ehsan Taghavi, Enxu Li, and Bingbing Liu. 2021. (AF)2-S3Net: Attentive feature fusion with adaptive feature selection for sparse semantic segmentation network. In *CVPR*. 12547–12556.

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*. 3075–3084.

[6] Tiago Cortinhal, George Tzelepis, and Eren Erdal Aksoy. 2020. SalsaNext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In *ISVC*. Springer, 207–222.

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*. 5828–5839.

[8] Khaled El Madawi, Hazem Rashed, Ahmad El Sallab, Omar Nasr, Hanan Kamel, and Senthil Yogamani. 2019. Rgb and lidar fusion based 3d semantic segmentation for autonomous driving. In *ITSC*. IEEE, 7–12.

[9] Kyle Genova, Xiaoqi Yin, Abhijit Kundu, Caroline Pantofaru, Forrester Cole, Avneesh Sud, Brian Brewington, Brian Shucker, and Thomas Funkhouser. 2021. Learning 3D semantic segmentation with only 2D image supervision. In *3DV*. IEEE, 361–372.

[10] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. 2020. A2D2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320* (2020).

[11] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*. 9224–9232.

[12] Meng-Hao Guo, Zheng-Ning Liu, Tai-Jiang Mu, and Shi-Min Hu. 2022. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 5 (2022), 5436–5447.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[14] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. CyCADA: Cycle-consistent adversarial domain adaptation. In *ICML*. Pmlr, 1989–1998.

[15] Yuenan Hou, Xinge Zhu, Yuexin Ma, Chen Change Loy, and Yikang Li. 2022. Point-to-Voxel Knowledge Distillation for LiDAR Semantic Segmentation. In *CVPR*. 8479–8488.

[16] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 2022. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *CVPR*. 9924–9935.

[17] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. 2022. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*. Springer, 372–391.

[18] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *CVPR*. 7132–7141.

[19] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. 2020. RandLA-Net: Efficient semantic segmentation of large-scale point clouds. In *CVPR*. 11108–11117.

[20] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. 2020. xMUDA: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*. 12605–12614.

[21] Georg Krispel, Michael Opitz, Georg Waltner, Horst Possegger, and Horst Bischof. 2020. FuseSeg: Lidar point cloud segmentation fusing multi-modal data. In *WACV*. 1874–1883.

[22] Miaoyu Li, Yachao Zhang, Yuan Xie, Zuodong Gao, Cuihua Li, Zhizhong Zhang, and Yanyun Qu. 2022. Cross-Domain and Cross-Modal Knowledge Distillation in Domain Adaptation for 3D Semantic Segmentation. In *ACMMM*. 3829–3837.

[23] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *CVPR*. 6936–6945.

[24] Wei Liu, Zhiming Luo, Yuanzheng Cai, Ying Yu, Yang Ke, José Marcato Junior, Wesley Nunes Gonçalves, and Jonathan Li. 2021. Adversarial unsupervised domain adaptation for 3D semantic segmentation with multi-modal learning. *ISPRS* 176 (2021), 211–221.

[25] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*. 2507–2516.

[26] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. 2019. RangeNet++: Fast and accurate lidar semantic segmentation. In *IROS*. IEEE, 4213–4220.

[27] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. 2021. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *ICCV*. 7108–7117.

[28] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. PointNet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*. 652–660.

[29] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. 2017. PointNet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS* (2017), 5099–5108.

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*. Springer, 234–241.

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. ImageNet large scale visual recognition challenge. *IJCV* 115, 3 (2015), 211–252.

[32] Cristiano Saltori, Fabio Galasso, Giuseppe Fiameni, Nicu Sebe, Elisa Ricci, and Fabio Poiesi. 2022. CosMix: Compositional semantic mix for domain adaptation in 3d lidar segmentation. In *ECCV*. Springer, 586–602.

[33] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In *ECCV*. Springer, 685–702.

[34] Haotian Tang, Zhijian Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In *ECCV*. Springer, 685–702.

[35] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. 2019. KPConv: Flexible and deformable convolution for point clouds. In *ICCV*. 6411–6420.

[36] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. 2018. Learning to adapt structured output space for semantic segmentation. In *CVPR*. 7472–7481.

[37] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. AdvEnt: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*. 2517–2526.

[38] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. 2018. SqueezeSeg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *ICRA*. IEEE, 1887–1893.

[39] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. 2019. SqueezeSegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *ICRA*. IEEE, 4376–4382.

[40] Wenxuan Wu, Zhongang Qi, and Li Fuxin. 2019. PointConv: Deep convolutional networks on 3d point clouds. In *CVPR*. 9621–9630.

[41] Aoran Xiao, Jiaxing Huang, Dayan Guan, Fangneng Zhan, and Shijian Lu. 2022. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In *AAAI*, Vol. 36. 2795–2803.

[42] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. 2020. SqueezeSegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *ECCV*. Springer, 1–19.

[43] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. 2021. RPVNet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *ICCV*. 16024–16033.

[44] Xu Yan, Jiantao Gao, Chaoda Zheng, Chao Zheng, Ruimao Zhang, Shuguang Cui, and Zhen Li. 2022. 2DPASS: 2d priors assisted semantic segmentation on lidar point clouds. In *ECCV*. Springer, 677–695.

[45] Li Yi, Boqing Gong, and Thomas Funkhouser. 2021. Complete & Label: A domain adaptation approach to semantic segmentation of lidar point clouds. In *CVPR*. 15363–15373.

[46] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. 2021. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *CVPR*. 12414–12424.

[47] Yachao Zhang, Miaoyu Li, Yuan Xie, Cuihua Li, Cong Wang, Zhizhong Zhang, and Yanyun Qu. 2022. Self-supervised Exclusive Learning for 3D Segmentation with Cross-Modal Unsupervised Domain Adaptation. In *ACMMM*. 3338–3346.

[48] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. 2020. PolarNet: An improved grid representation for online lidar point clouds semantic segmentation. In *CVPR*. 9601–9610.

[49] Sicheng Zhao, Yezhen Wang, Bo Li, Bichen Wu, Yang Gao, Pengfei Xu, Trevor Darrell, and Kurt Keutzer. 2021. ePointDA: An end-to-end simulation-to-real domain adaptation framework for lidar point cloud segmentation. In *AAAI*, Vol. 35. 3500–3509.

[50] Zhedong Zheng and Yi Yang. 2021. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *IJCV* 129, 4 (2021), 1106–1120.

[51] Zhuangwei Zhuang, Rong Li, Kui Jia, Qicheng Wang, Yuanqing Li, and Mingkui Tan. 2021. Perception-aware multi-sensor fusion for 3d lidar semantic segmentation. In *ICCV*. 16280–16290.

Table 5: Size of the splits in frames for all proposed UDA scenarios.

Scenarios	Source	Target		Categories
	Train	Train	Val/Test	
nuScenes: Day→Night	24745	2779	606/602	**Background**: [background]; **Vehicle**: [bus, car, construction_vehicle, trailer, truck]; **Pedestrian**: [pedestrian]; **Bike**: [bicycle, motorcycle]; **Traffic_boundary**: [barrier, traffic_cone]
nuScenes: USA→Singapore	15695	9665	2770/2929	
A2D2→SemanticKITTI	27695	18029	1101/4071	**Car**; **Truck**; **Bike**; **Person**; **Road**; **Parking**; **Sidewalk**; **Building**; **Nature**; **Other-objects**

Table 6: Fusion without stylized input.

Method	2D	3D	2D+3D
xMUDA	47.3	46.0	50.6
BFtD-xMUDA w/o MMST and Teacher	50.3	46.5	56.0
BFtD-xMUDA w/o MMST	50.7	46.0	57.6
BFtD-xMUDA (Ours)	51.2	46.3	60.9

Table 7: Computation complexity of attention module.

Method	Params	FLOPs	GPU Mem.	mIoU (%)
self-attention	0.69M	13.65G	14849MB	52.3
channel attention	0.05M	0.95G	1193MB	56.5
memorized attention (Ours)	0.12M	3.88G	2795MB	60.9

A SUPPLEMENTARY MATERIAL

This material starts with more details of implementations, including dataset split, additional analysis, and qualitative results. Code is available at https://github.com/Barcaaaa/BFtD-xMUDA.

A.1 Dataset Split

To compose our domain adaptation scenarios, following [20], we exploit public datasets, including nuScenes [3], A2D2 [10], and SemanticKITTI [2]. The split details are tabulated in Tab. 5.

nuScenes contains 1,000 scenes, each of 20 seconds, which corresponds to 40k annotated keyframes taken at 2Hz. The original scenes are split into 28,130 training frames and 6,019 validation frames. Each frame contains a 32-beam LiDAR point cloud provided with point-wise annotations and six RGB images captured by six cameras from different views of LiDAR. For nuScenes: Day→Night, we choose 602 night scenes for testing data, while for nuScenes: USA→Singapore, we choose 2,929 Singapore scenes for testing data. Both of them merge the objects into the background and 4 categories: **Vehicle**, **Pedestrain**, **Bike**, and **Traffic_boundary**.

A2D2 consists of 20 drives, which corresponds to 28,637 frames. The point cloud comes from three 16-layer front LiDARs (center, left, and right), where the left and right LiDARs are inclined. By projecting 3D point clouds onto 2D images, corresponding 2D semantic labels are regarded as 3D point-wise labels, which contain 38 categories.

SemanticKITTI is a large-scale dataset based on the KITTI Odometry Benchmark captured in Germany. The original scenes are split into 19,130 training scans and 4,071 validation scans. Unlike nuScenes, SemanticKITTI only provides the front-view images and a 64-layer front LiDAR. 19 categories are used for segmentation.

Note that, we select 10 shared classes between the A2D2 and SemanticKITTI, including **Car**, **Truck**, **Bike**, **Person**, **Road**, **Parking**, **Sidewalk**, **Building**, **Nature**, and **Other-objects**.

A.2 More Analysis

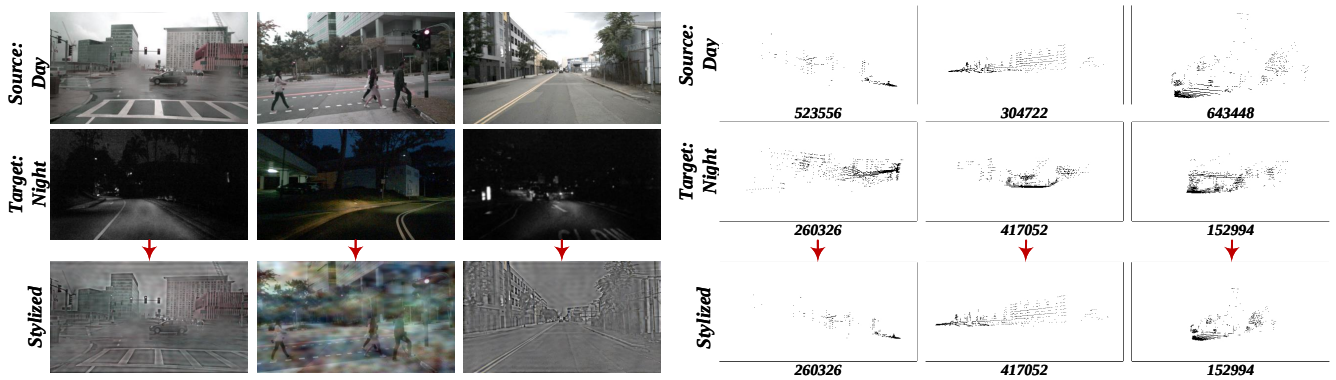
Multi-modal Style Transfer. To evaluate whether the multi-modal style transfer (MMST) module can benefit the fusion streamline in our framework, we visualize the 2D and 3D style transfer on all adaptation scenarios. As shown in Fig. 7a, Fig. 8a, and Fig. 9a, the source images in the first row are quite different from the target images in the second row. The last row shows the stylized images of the source domain in the target style. It is observed that stylized images are similar to the source images in terms of semantic content but relatively dark. Similarly, in Fig. 7b, Fig. 8b, and Fig. 9b, point clouds in stylized domain contain the same semantic content with source point clouds but sparser or denser according to the range of the target domain, proving the effectiveness of MMST.

Fusion without Stylized Input. As shown in Tab. 6, we supplement two additional ablation studies without MMST in the scenario: "nuScenes: Day→Night(Seg)", where"w/o MMST" means that removing the style transfer generated by MMST, and only the samples of source domain are fed into the Teacher. Our method still achieves a gain of +7.0% mIoU. Furthermore, "w/o MMST and Teacher" means that the MMST and Teacher are simultaneously removed and only separately performing cross-modal learning on the source domain and target domain (*i.e.*, only Modality-Preserving Distillation). It is observed that our cross-modal fusion representation is superior to the xMUDA [20] with a gain of +5.4% mIoU.

Computation Complexity of attention module. To evaluate our fusion module, in Tab. 7, we show the comparison of computational complexity between self-attention, channel attention, and our proposed multi-modal memorized attention in the scenarios: "nuScenes: Day→Night(Seg)". It is observed that our method achieves better results while keeping the floating point per second (FLOPs) and parameters relatively low.

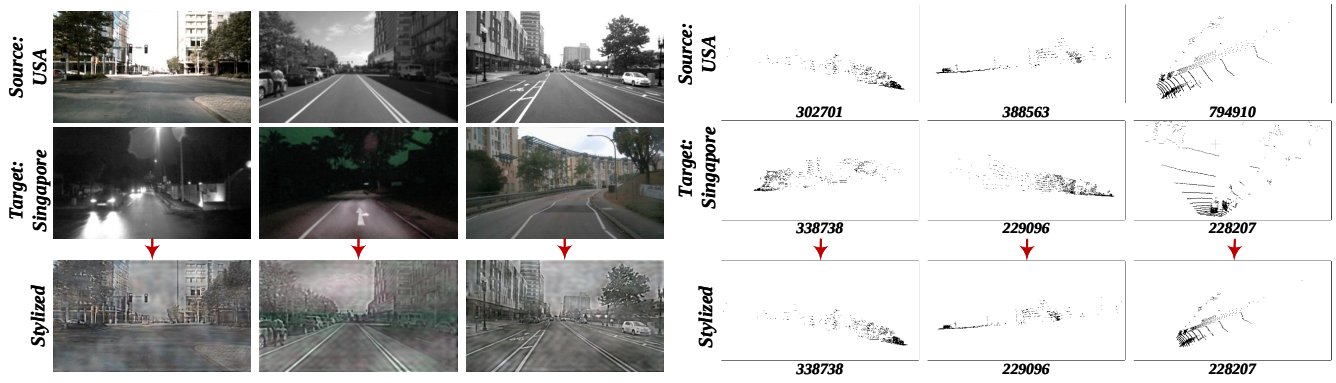
Pseudo-Labeling Strategy. Pseudo-label learning is a frequently used technique in UDA, which leverages the pseudo-label to learn from the unlabeled target data. However, after trial and error, it is observed that our model easily degenerates and achieves poor performance when following the general pseudo-label in cross-modal UDA. The main reason lies in one inherent problem the target pseudo label inevitably contains noise in the UDA task, which compromises the re-training of MFFM.

Therefore, the proposed cross-modal debiased pseudo-labeling (xDPL) can alleviate this problem. As shown in Tab. 8, we tabulate different pseudo-labeling strategies to prove the effectiveness of xDPL. "PL" means the 2D(3D) pseudo-label with a hand-crafted



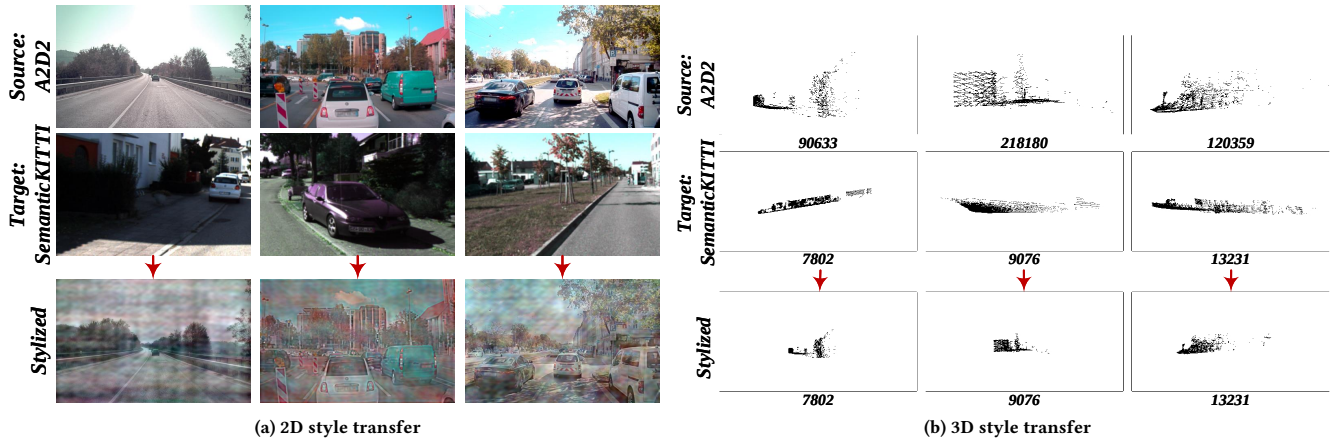
(a) 2D style transfer (b) 3D style transfer

Figure 7: Style transfer on nuScenes: Day→Night.



(a) 2D style transfer (b) 3D style transfer

Figure 8: Style transfer on nuScenes: USA→Singapore.



(a) 2D style transfer (b) 3D style transfer

Figure 9: Style transfer on A2D2→SemanticKITTI.

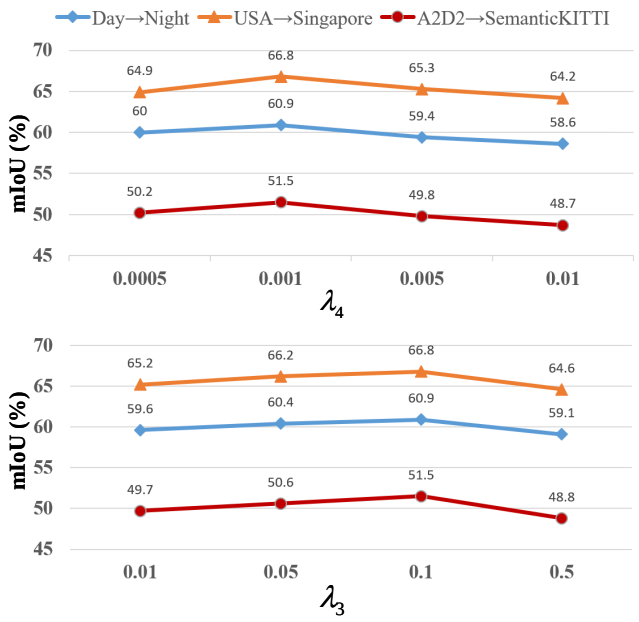


Figure 10: Ablation study on weight λ_3 and λ_4 of DPD loss.

Table 8: Ablation study on different pseudo-labeling strategies.

Strategy	λ_5	mIoU (%)
PL	0.1	59.2
	1.0	58.9
PL_{fusion}	0.1	58.8
	1.0	57.6
xDPL	0.1	60.1
	1.0	59.5
xDPL_{fusion}	0.1	58.4
	1.0	57.3

threshold output from the 2D(3D) streamline and then exploits it to supervise the output from the 2D(3D) streamline of the target domain. "PL_{fusion}" means the pseudo-label with a hand-crafted threshold output from fusion streamline and then exploits fusion pseudo-labels to supervise the output from 2D and 3D streamlines of the target domain. Of note, the hand-crafted threshold γ is set to $\gamma = min\,[mod, 0.9]$, where "mod" denotes the median of the prediction probability distribution. During re-training, adding pseudo-labels is relatively unstable, so we run 3 times for each setting and take the best performance. It is observed that the proposed xDPL arrives the superior performance to the general pseudo-labeling strategy with a hand-crafted threshold.

Hyperparameters of Overall Loss. For the weight of modality-preserving distillation loss in each domain, we follow xMUDA [20] and then set λ_1 and λ_2 to 1.0 and 0.1, respectively. For the weight of domain-preserving distillation loss, we conduct ablation study for λ_3 and λ_4. As shown in Fig. 10, we first fix $\lambda_3 = 0.1$ and analyse the impact of λ_4, it is observed that when $\lambda_4 = 0.001$, the results can achieve the best performance. After that, we further fix $\lambda_4 = 0.001$ and analyse the impact of λ_3, it is observed that when $\lambda_3 = 0.1$, the results can achieve the best performance. Of note, although the fluctuation of λ_3 and λ_4 has a litter influence on bridging the domain gap, bidirectional fusion-then-distillation is insensitive to hyperparameters compared to adversarial-based methods, indicating that our method is stable.

A.3 More Qualitative Results

We provide additional qualitative results for cross-modal UDA in all scenarios. Firstly, we show the output of ensemble results to illustrate our strengths. As shown in Fig. 11, the visual appearance is quite different during the day (motorcycle visible) than during the night (only the headlight visible). In Col.2 and 5, xMUDA and Dual-Cross mistakenly identify *motorcycle* as a *vehicle*, while our method can accurately classify points located on *motorcycle*. Moreover, in Col.4, many *motorcycles* are stacked on top of each other, partially occluded. Our method still correctly predicted most *motorcycles* compared to xMUDA and Dual-Cross. It is observed that our method can adapt and easily recognize distant objects.

Moreover, as shown in Col. 5 of Fig. 12, a motorcycle is in oncoming traffic, and in Col. 1 and 3, a motorcycle is parked on the side of the drivable surface. xMUDA and Dual-Cross mistakenly identified the motorcycle as a pedestrian. In particular, Col.2 shows a special case. When a pedestrian appears in front of an oncoming vehicle, xMUDA and Dual-Cross mix pedestrians and the vehicle together for prediction due to the occlusion relationship, while our method distinguishes and conducts accurate classification.

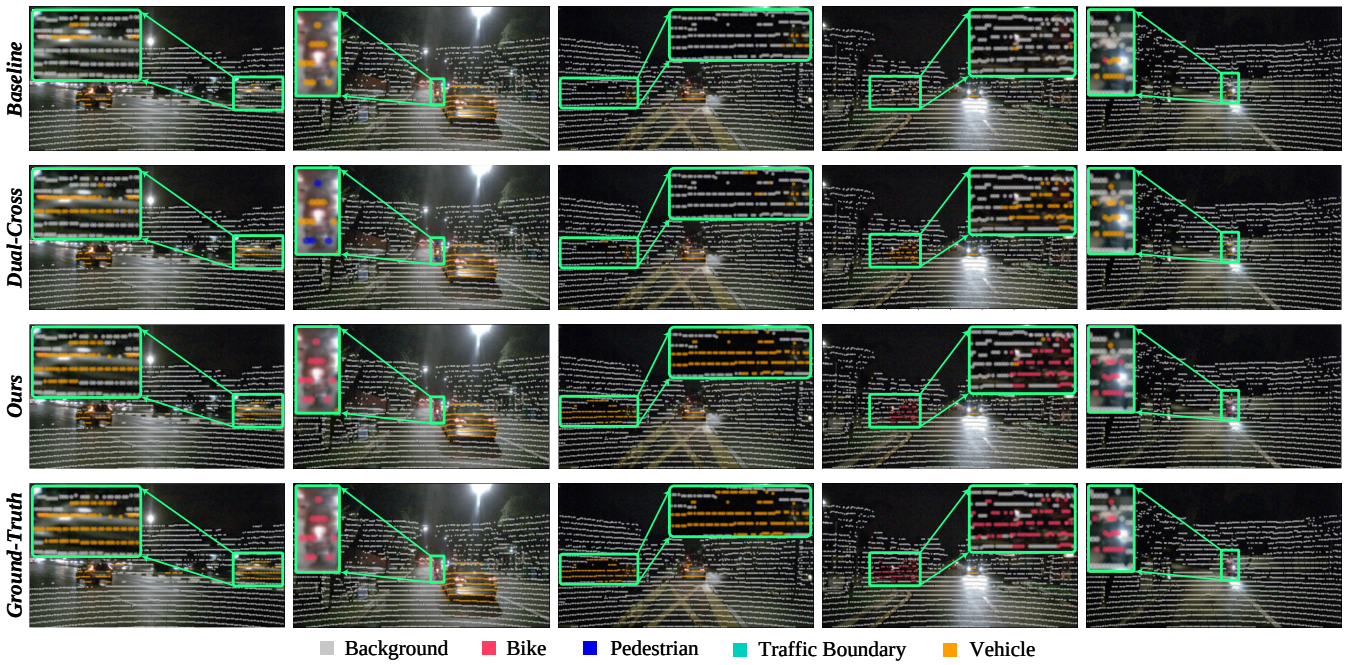


Figure 11: Qualitative results on nuScenes: Day→Night.

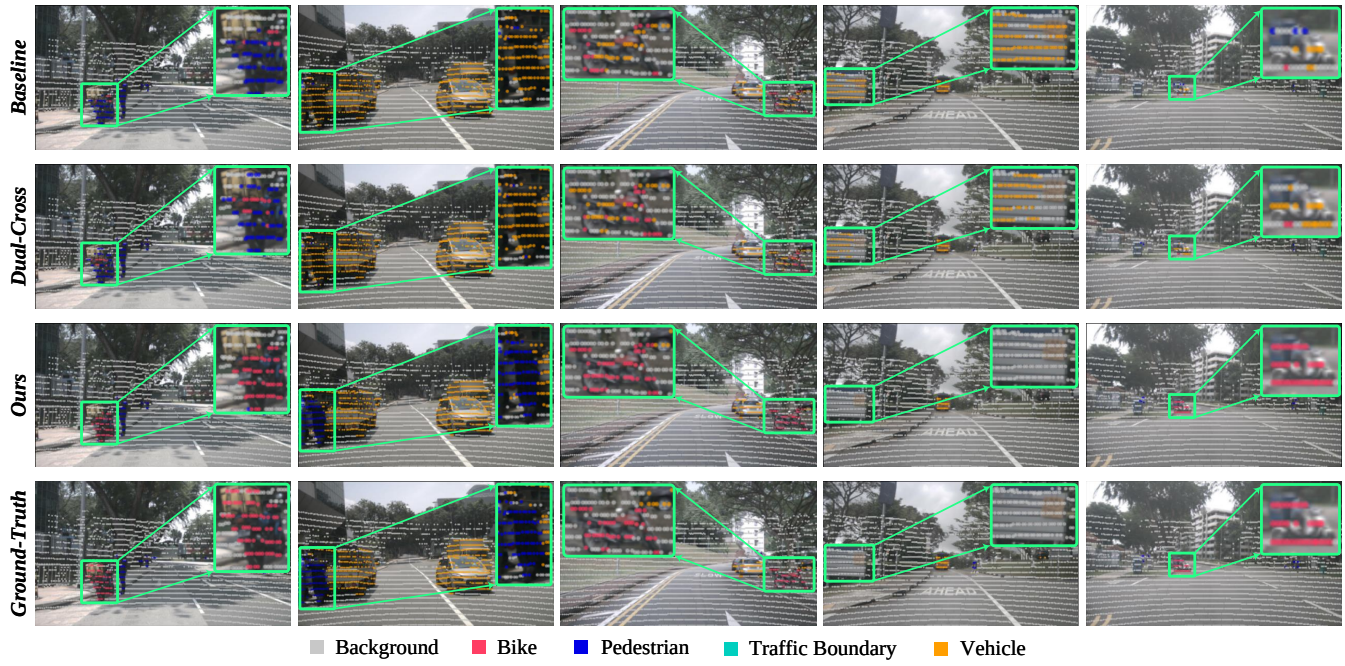


Figure 12: Qualitative results on nuScenes: USA→Singapore.

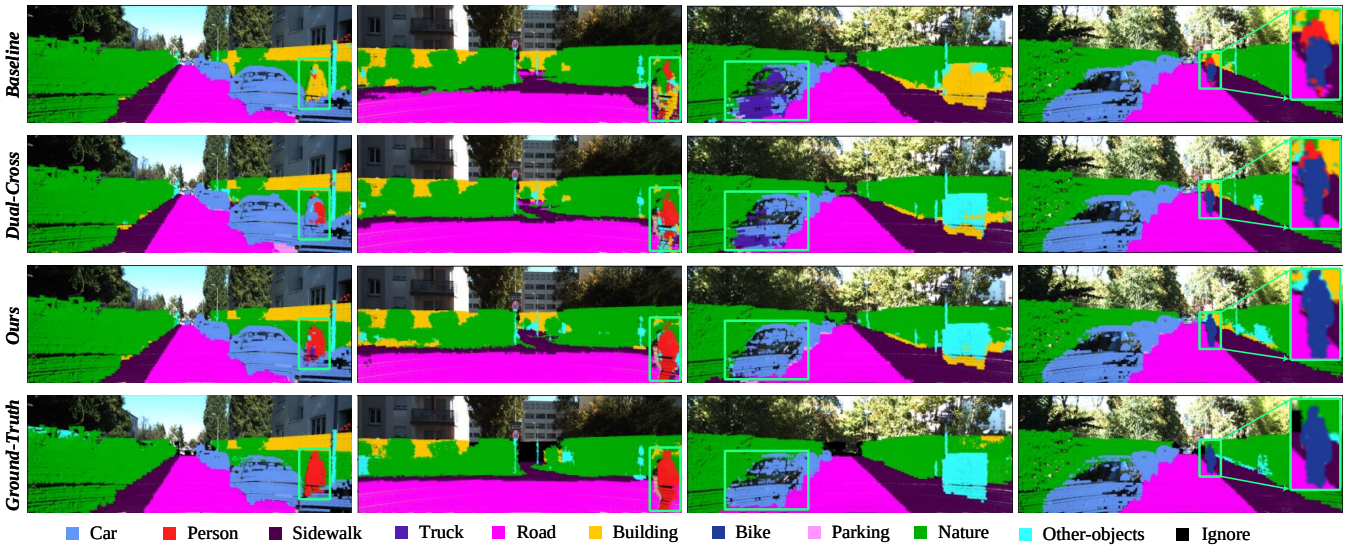


Figure 13: Qualitative results on A2D2→SemanticKITTI.