MM-TTA: Multi-Modal Test-Time Adaptation for 3D Semantic Segmentation

Inkyu Shin[1] Yi-Hsuan Tsai[2] Bingbing Zhuang[3] Samuel Schulter[3]
Buyu Liu[3] Sparsh Garg[3] In So Kweon[1] Kuk-Jin Yoon[1]
[1]KAIST [2]Phiar Technologies [3]NEC Laboratories America

Abstract

*Test-time adaptation approaches have recently emerged as a practical solution for handling domain shift without access to the source domain data. In this paper, we propose and explore a new multi-modal extension of test-time adaptation for 3D semantic segmentation. We find that, directly applying existing methods usually results in performance instability at test time, because multi-modal input is not considered jointly. To design a framework that can take full advantage of multi-modality, where each modality provides regularized self-supervisory signals to other modalities, we propose two complementary modules within and across the modalities. First, Intra-modal Pseudo-label Generation (**Intra-PG**) is introduced to obtain reliable pseudo labels within each modality by aggregating information from two models that are both pre-trained on source data but updated with target data at different paces. Second, Inter-modal Pseudo-label Refinement (**Inter-PR**) adaptively selects more reliable pseudo labels from different modalities based on a proposed consistency scheme. Experiments demonstrate that our regularized pseudo labels produce stable self-learning signals in numerous multi-modal test-time adaptation scenarios for 3D semantic segmentation. Visit our project website at* `https://www.nec-labs.com/~mas/MM-TTA`

1. Introduction

3D semantic segmentation is a challenging task that requires both geometric and semantic reasoning about the input scene, but it can provide rich insights that enable applications like autonomous driving [32, 34], virtual reality and robotics [5, 27]. With the advancement of sensor technology, multi-modal sensors are considered as the key to effectively tackle this task [6, 16, 17]. In particular, to obtain more accurate 3D point-level semantic understanding, contextual information in 2D RGB images can be reinforced by the geometric property of 3D points from LiDAR sensors, and vice versa. Therefore, it is of great interest to develop multi-modal approaches for 3D semantic segmentation.

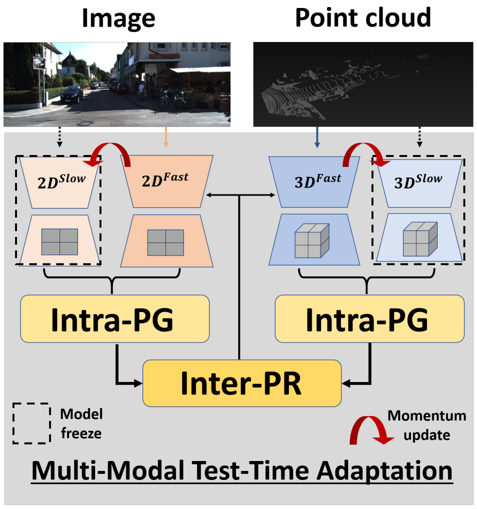

Figure 1. We propose a Multi-Modal Test-Time Adaptation (MM-TTA) framework that enables a model to be quickly adapted to multi-modal test data without access to the source domain training data. We introduce two modules: 1) Intra-PG to produce reliable pseudo labels within each modality via updating two models (batch norm statistics) in different paces, *i.e.*, slow and fast updating schemes with a momentum, and 2) Inter-PR to adaptively select pseudo-labels from the two modalities. These two modules seamlessly collaborate with each other and co-produce final cross-modal pseudo labels to help test-time adaptation.

However, multi-modal data is sensitive to a distribution shift at test time when a domain gap exists to the training data [1]. Therefore, it is critical for a model to quickly adapt to the new multi-modal data during testing for obtaining better performance, *i.e.*, through test-time adaptation (TTA) [19, 30]. This is different from the usual domain adaptive semantic segmentation setting [13, 28, 35] that can access both source and target data during training. In TTA, we only have access to model parameters pre-trained on the source data and the unlabeled test data for *quick* adaptation, which typically (and also in this work) refers to one epoch of training. This is practical for real-world scenarios, but it is also challenging because only the target data is available with a limited budget for adaptation.

In this paper, we study multi-modal 3D semantic segmentation in the setting of test-time adaptation, using both image and point cloud as input. Prior works on general test-time adaptation like TENT [30] propose entropy minimization as a self-training loss to update batch norm parameters. While TENT [30] is not designed for multi-modality, we show a simple extension that updates parameters in individual branches for each modality (2D image and 3D point cloud). However, we find that this extension causes instability during training. One reason is that, since entropy minimization tends to generate sharp output distributions, using it separately for 2D and 3D branches may increase the cross-modal discrepancy. This would further lead to a sub-optimal model ensemble for 2D and 3D outputs, which is the common scheme for multi-modal semantic segmentation. One way to alleviate this cross-modal discrepancy is to utilize a consistency loss [13] between predictions of 2D and 3D branches, via KL divergence. However, since the test data during adaptation is unlabeled, enforcing the consistency across modalities may even worsen predictions if the output of one branch is inaccurate.

To tackle the aforementioned issues and design better test-time self-supervisory signals, we propose a cross-modal regularized self-training framework that aims to generate reliable and adaptive pseudo labels (see Fig. 1). Our method mainly consists of two modules: 1) Intra-modal Pseudo-label Generation (**Intra-PG**), and 2) Inter-modal Pseudo-label Refinement (**Inter-PR**). For the intra-modal module, we aim to produce reliable pseudo labels in each modality that alleviate the instability issue in test-time adaptation, *i.e.*, only updating batch norm parameters by seeing the test data once. To this end, we design a slow-fast modeling strategy. Specifically, to maintain the model stability, we initialize one batch norm statistics from the pre-trained source model, and *slowly* update it with a momentum from another fast-updated batch norm parameter, while this *fast*-updated model is directly updated by the test data, which is more aggressive but also provides up-to-date statistics. Our model is thus able to fuse predictions from the slow-/fast-updated statistics to enjoy their complementary benefits.

For the inter-modal module, we propose to adaptively select reliable pseudo labels from the individual 2D and 3D branches, because each modality brings its own advantage for 3D semantic segmentation. To this end, we first leverage the Intra-PG module to measure the prediction consistencies of each modality separately, and then provide a fused prediction from slow-fast models to the Inter-PR module (Fig. 1). Based on these consistencies, our model adaptively selects reliable pseudo labels from two modalities to form a final cross-modal pseudo label as the self-training signal to update 2D/3D batch norm parameters.

The proposed two modules collaborate with each other for multi-modal test-time adaptation, and thus we name our framework as *MM-TTA*. We conduct extensive experiments to include several TTA state-of-the-art baselines and show that our MM-TTA framework achieves favorable performance over different benchmark settings, including cross-dataset with different sensors, synthetic-to-real, and day-to-night scenarios. Moreover, we provide comprehensive analysis to demonstrate the benefits of our two proposed modules (Intra-PG and Inter-PR) and the stability comparisons with existing methods. Here are our main contributions:

1. We explore a new task, test-time adaptation for multi-modal 3D semantic segmentation, and propose a framework that effectively produces cross-modal pseudo labels as self-training signals.

2. We introduce two modules that seamlessly work together: The Intra-PG module produces pseudo labels for each modality separately and the Inter-PR module adaptively selects pseudo labels across modalities.

3. We demonstrate our framework under different adaptation settings with extensive ablation studies and experimental comparisons against strong baselines and state-of-the-art methods.

2. Related Work

Test-Time Adaptation (TTA) aims to enable quick adaptation of an existing model to new target data without having access to the source domain data the model was trained on. As an important challenge for dealing with dynamic domain shift in real-world, TTA is attracting more and more attention in several tasks [4, 15, 19, 25, 30]. Among them, Test-time Training (TTT) [25] updates model parameters in an online manner by applying a self-supervised proxy task on the test data. Since this proxy task is also required for training samples, finding an optimal proxy task that works well in both training and testing is challenging.

From that point of view, TENT [30], the first Test-Time Adaptation (TTA) approach, proposes a simple yet effective entropy minimization method to optimize for test-time batch norm parameters without requiring any proxy task during training, which is demonstrated for image classification and 2D semantic segmentation. However, entropy minimization tends to encourage the model to increase confidence despite false predictions. To design a regularized self-learning signal at test-time, a concurrent work, S4T [19], proposes a selective self-training scheme for 2D semantic segmentation by regularizing pseudo labels with aligned predictive view generation. Nonetheless, this design is considered to be specific to an image-level task where spatial augmentation can be performed. Compared to the aforementioned work, we study a similar TTA setting but in the different context of using multi-modality for 3D semantic segmentation, *i.e.*, Multi-Modal Test-Time Adaptation

(MM-TTA), in which we develop intra-modal and inter-modal modules that seamlessly work with each other to obtain more reliable self-learning signals.

3D Semantic Segmentation has been recognized as an important 3D scene understanding task aimed at classifying each LiDAR point into semantic categories. Therefore, point clouds from LiDAR are deemed to be the dominant modality to solve this task [5, 27, 31, 32, 34]. Range-based methods [31, 32] adopt spherical projection to project 3D points onto the 2D image plane and then pass this through a 2D-based backbone [11]. Another effort is to utilize the raw 3D point cloud by designing 3D segmentation models [5, 27]. KPConv [27] operates on point clouds without any intermediate representation, while MinkowskiNet [5] voxelizes the point cloud and utilizes SparseConvNet [9] for processing. Despite these efforts, the LiDAR point cloud itself lacks 2D contextual information that is essential to understand the complex semantics of a scene.

To address this weakness, recent work [6, 17] explores the use of multi-modal inputs (RGB images and LiDAR point clouds) for 3D segmentation. These methods commonly separate the backbones for 2D and 3D modalities and propose fusion techniques between the two outputs. Considering both contextual and geometric information from each modality is shown to boost the performance in 3D semantic segmentation. However, since each modality has different dataset biases (*e.g.*, style distribution in 2D and point distribution in 3D), multi-modality based models are harder to adapt to new data. In this work, different from supervised training, we tackle multi-modal 3D semantic segmentation in the test-time adaptation setting, which is practical as it incorporates test data statistics during inference, thus improving results from multi-modal baselines.

Unsupervised Domain Adaptation (UDA) aims at bridging the gap between labeled source data and unlabeled target data. Methods for both 2D [14, 24, 28, 29, 36, 37] and 3D [20, 23, 33, 35] data have been proposed. Recently, few works [13, 18] also introduced UDA approaches for 2D/3D multi-modal data. Specifically, xMUDA [13] executes consistency learning at training time between two modalities both in source and target domains, while DsCML [18] further utilizes adversarial learning with dynamic sparse-to-dense cross-modal learning between modalities. All UDA methods are allowed to access source data during adaptation, while we tackle test-time adaptation, in which only the source pre-trained model is available and a limited budget is given to update test time statistics of the model.

3. Proposed Method

We start by introducing the preliminaries for test-time adaptation in Sec. 3.1. Then, we explore several baselines for multi-modal test-time adaptation in Sec. 3.2, using the image (2D modality) and point cloud (3D modality) as input for 3D semantic segmentation. Finally, we propose our Multi-Modal Test-Time-Adaptation (MM-TTA) framework, as shown in Fig. 2, with two newly designed modules: 1) Intra-PG to generate pseudo-labels within each modality (Sec. 3.3), and 2) Inter-PR to adaptively select reliable pseudo-labels across modalities (Sec. 3.4).

Setup and notation. We follow the setting of test-time adaptation [30], where we are not able to access the source data but only the source pre-trained multi-modal segmentation model. This model consists of 2D and 3D branches, F^{2D} and F^{3D}, each of which including the feature extractor $G^{\text{2D}}/G^{\text{3D}}$ and a classifier. Here, we also denote the multi-modal test-time target data (see Fig. 2), images $x_t^{\text{2D}} \in \mathbb{R}^{H\times W\times 3}$ and point clouds $x_t^{\text{3D}} \in \mathbb{R}^{N\times 3}$ (3D points in the camera field of view). Note that the feature extracted from the 2D branch, $G^{\text{2D}}(x_t^{\text{2D}}) \in \mathbb{R}^{H\times W\times f}$, is sampled at the N projected 3D points resulting in a feature shape of $N \times f$. Individual network predictions from 2D/3D are denoted as $p(x_t^M) = F^M(x_t^M) \in \mathbb{R}^{N\times|K|}$, where $|K|$ is the number of categories and $M \in \{\text{2D}, \text{3D}\}$.

3.1. Preliminaries

Batch Normalization (BN) [12] has been widely used in current DNNs for both 2D and 3D models. It generally includes normalization statistics and transformation parameters in the j-th BN layer given the target mini-batch input x_t^M with $M \in \{\text{2D}, \text{3D}\}$:

$$\hat{x}_{t_j}^M = \frac{x_{t_j}^M - \mu_{t_j}^M}{\sigma_{t_j}^M} \quad \text{and} \quad y_{t_j}^M = \gamma_{t_j}^M \hat{x}_{t_j}^M + \beta_{t_j}^M, \tag{1}$$

where $\mu_{t_j}^M = \mathbb{E}[x_{t_j}^M]$ and $(\sigma_{t_j}^M)^2 = \mathbb{E}[(\mu_{t_j}^M - x_{t_j}^M)^2]$ are normalization statistics, and $\gamma_{t_j}^M, \beta_{t_j}^M$ are learnable transformation parameters. To simplify notation, we use $\Omega_t^{\text{2D}} = (\mu, \sigma, \gamma, \beta)_t^{\text{2D}}$ for 2D and $\Omega_t^{\text{3D}} = (\mu, \sigma, \gamma, \beta)_t^{\text{3D}}$ for 3D.

The number of parameters updated in Test-time Adaptation is constrained to be small for reasons of efficiency and stability. Following TENT [30], we only estimate and optimize $(\mu, \sigma, \gamma, \beta)_t$ occupying <1% of parameters both in 2D and 3D branches.

3.2. Baselines for MM-TTA

Since we propose the first attempt of multi-modal test-time adaptation for 3D semantic segmentation, we first study several self-learning baselines based on existing methods that we extend to our MM-TTA setting.

Self-learning with Entropy is originally proposed by TENT [30]. Its test-time objective $L(x_t)$ is to minimize the entropy of model predictions $p(x_t^M) = F^M(x_t^M)$, where F^M is either the 2D or 3D branch (recall that $M \in$

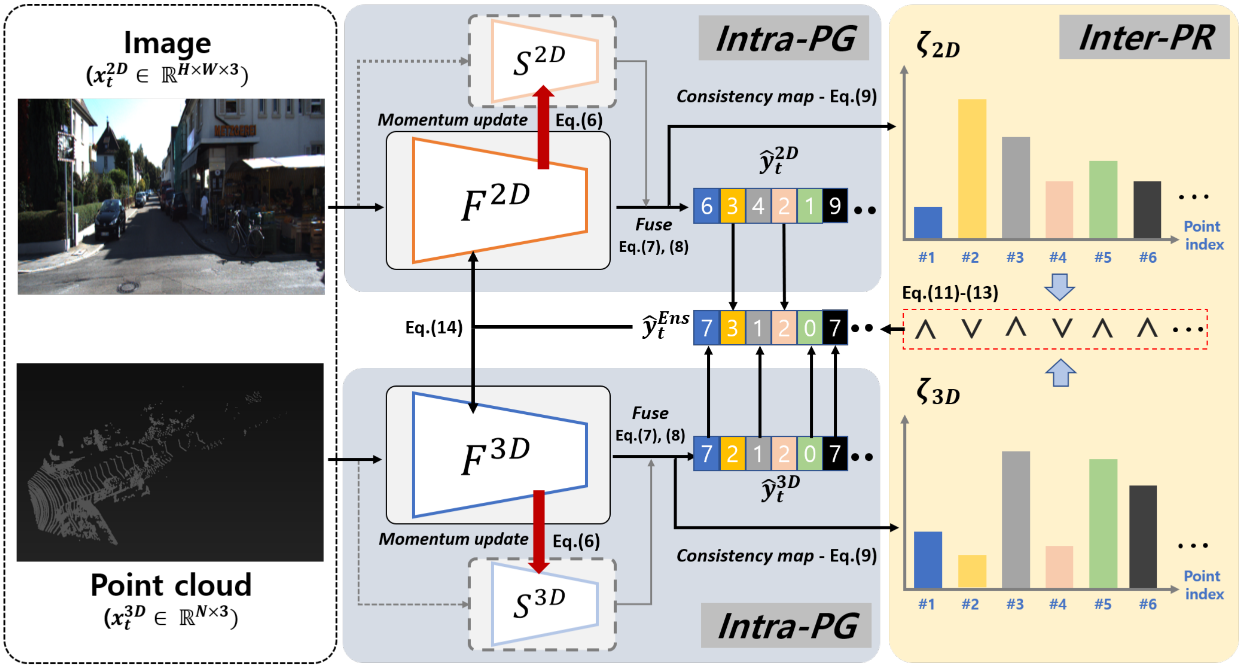

Figure 2. **Overview of the proposed Multi-Modal Test-Time Adaptation (MM-TTA) framework.** Our MM-TTA consists of two modules: Intra-modal Pseudo-label Generation (**intra-PG**) and Inter-modal Pseudo-label Refinement (**inter-PR**). For Intra-PG, we adopt a slowly-updated model S that is gradually updated by a fast-updated model S with a momentum. Note that, statistics in the fast-updated model S are directly updated by the data, which is more aggressive but up-to-date, while the model S slowly moves towards the target data statistics and thus is more stable. By aggregating slow-fast models, each modality can generate robust pseudo labels ($\hat{y}_t^{2D}$ and $\hat{y}_t^{3D}$). For Inter-PR, we measure the consistency map between slow-fast models and enable an adaptive selection process for finding confident pseudo labels based on calculated ζ_{2D} and ζ_{3D}. After obtaining the cross-modal regularized pseudo label ($\hat{y}_t^{Ens}$) by jointly considering 2D and 3D confidences, we update the batch norm parameters for F in both modalities.

$\{2\mathrm{D}, 3\mathrm{D}\}$). The overall objective of entropy minimization for this MM-TTA baseline is expressed as:

$$\begin{aligned} L_{\text{ent}}(x_t) = &-\sum_k p(x_t^{\text{2D}})^{(k)} \log p(x_t^{\text{2D}})^{(k)} \\ &-\sum_k p(x_t^{\text{3D}})^{(k)} \log p(x_t^{\text{3D}})^{(k)}, \end{aligned} \tag{2}$$

where k denotes the class. Despite its simplicity, this objective only encourages sharp output distributions, which may reinforce wrong predictions, and may not lead to cross-modal consistency.

Self-learning with Consistency aims to achieve multi-modal test-time adaptation via a consistency loss between predictions of 2D and 3D modalities:

$$\begin{aligned} L_{\text{cons}}(x_t) = &D_{\text{KL}}(p(x_t^{\text{2D}})||p(x_t^{\text{3D}})) \\ &+ D_{\text{KL}}(p(x_t^{\text{3D}})||p(x_t^{\text{2D}})), \end{aligned} \tag{3}$$

where D_{KL} is the KL divergence. Different from xMUDA [13], which operates in the standard domain adaptation setting with access to the source data, our MM-TTA is not regularized by the source task loss and thereby this objective may fail to capture the correct consistency when one of the branches provides a wrong prediction.

Self-learning with Pseudo-labels is another common approach for test-time adaptation. Typically, pseudo-labels $\hat{y}_t$ can be obtained by:

$$\hat{y}_t = \operatorname*{arg\,max}_{k\in K} \mathbb{1}[\, p(x_t)^{(k)} > \theta^{(k)} \,]\, p(x_t)^{(k)}, \tag{4}$$

where $\mathbb{1}[\cdot]$ is an indicator function returning true if the condition is satisfied, *i.e.*, if prediction $p(x_t)^{(k)}$ for class k is larger than the threshold $\theta^{(k)}$. Note that the pseudo-label $\hat{y}_t$ can be obtained similarly for both 2D and 3D branches, *i.e.*, $\hat{y}_t^{\text{2D}}$ and $\hat{y}_t^{\text{3D}}$. The objective for pseudo-labeling uses the standard cross-entropy loss L_{seg} for semantic segmentation:

$$\begin{aligned} L_{\text{pseudo}}(x_t) = &L_{\text{seg}}(p(x_t^{\text{2D}}), \hat{y}_t^{\text{2D}}) \\ &+ L_{\text{seg}}(p(x_t^{\text{3D}}), \hat{y}_t^{\text{3D}}). \end{aligned} \tag{5}$$

Although pseudo-labels provide supervisory signals to update models, there are potential issues when it is applied to our MM-TTA setting. First, only the batch norm statistics are updated to replace the original source statistics during adaptation, but the model to generate pseudo labels for target data still mainly consists of fixed parameters pre-trained on source data, which can lead to low-quality pseudo labels. Second, the model still lacks information exchange across modalities to refine pseudo labels, which can also result in sub-optimal performance. In contrast, our proposed MM-TTA framework provides simple yet effective solutions to these limitations with the following two modules.

3.3. Intra-modal Pseudo-label Generation

We propose Intra-PG to generate reliable online pseudo labels within each modality by having two models, S^M and F^M, with different updating paces (see Fig. 2). First, we define a fast-updated model F^M that replaces and updates batch norm statistics directly from the test data, which is identical to baseline models in Section 3.2. Second, we introduce an additional slowly-updated model S^M that is initially source pre-trained and has a momentum update scheme from the fast-updated model F^M. In short, we denote these two models as slow/fast model as S^M/F^M. That is, the statistics in the fast model are updated more aggressively by the test data, while the slow model's statistics gradually move towards the target statistics, and thus provide a stable and complementary supervisory signal. Note that only the slow model S^M is used at inference time. Here, we present the batch norm statistics for the slow model S^M as:

$$\begin{aligned} \Omega^S_{t_i} &= (1-\lambda)\Omega^F_{t_i} + \lambda\Omega^S_{t_{i-1}}, \\ \Omega^S_{t_0} &= \Omega_s, \end{aligned} \tag{6}$$

where $\Omega^S_{t_i} = (\mu, \sigma, \gamma, \beta)^S_{t_i}$ is the moving averaged statistics at iteration i with a momentum factor λ to aggregate fast model's statistics $\Omega^F_{t_i}$ and slow model's statistics $\Omega^S_{t_{i-1}}$. The initial statistics $\Omega^S_{t_0}$ are from the source pre-trained model denoted as Ω_s. Note that, when we set a large value for λ (0.99 in the paper), it will move slower towards the target statistics, and otherwise it moves faster. To further leverage both the slow-fast statistics in each modality, we fuse their predictions as:

$$p(x^M_t) = \frac{(S^M(x^M_t) + F^M(x^M_t))}{2}. \tag{7}$$

Then, we can obtain aggregated pseudo labels from slow-fast models for each modality $M \in \{\text{2D}, \text{3D}\}$:

$$\hat{y}^M_t = \arg\max_{k\in K} p(x^M_t)^{(k)}. \tag{8}$$

3.4. Inter-modal Pseudo-label Refinement

After obtaining initial aggregated pseudo labels for each modality in (8), we propose the Inter-PR module to improve pseudo labels via cross-modal fusion. To realize this, we first calculate a consistency measure (ζ_M) between slow and fast models of Intra-PG for each modality separately:

$$\zeta_{\text{M}} = Sim(S^M(x^M_t), F^M(x^M_t)), \tag{9}$$

where we define $Sim(\cdot)$ as the inverse of KL divergences to express the similarity between two probabilities:

$$Sim(x,y) = \left(\frac{1}{D_{\text{KL}}(x||y)+\epsilon} + \frac{1}{D_{\text{KL}}(y||x)+\epsilon}\right)/2. \tag{10}$$

Here, ϵ is a small scalar constant to prevent division-by-zero. This consistency measure helps us to fuse the per-modality predictions and estimate more reliable pseudo labels. We propose two variants: *Hard Select* and *Soft Select*. The former takes each pseudo label exclusively from one of the modalities, while the latter conducts a weighted sum of pseudo labels from the two modalities using the consistency measure. We define *Hard Select* as

$$\hat{y}^H_t = \begin{cases} \hat{y}^{\text{2D}}_t, & \text{if} \zeta_{\text{2D}} \geq \zeta_{\text{3D}}, \\ \hat{y}^{\text{3D}}_t, & \text{otherwise.} \end{cases} \tag{11}$$

and *Soft Select* as

$$\hat{y}^S_t = \arg\max_{k\in K} p^{W(k)}_t, \tag{12}$$

with $p^{W(k)}_t = \zeta^*_{\text{2D}}\, p(x^{\text{2D}}_t)^{(k)} + \zeta^*_{\text{3D}}\, p(x^{\text{3D}}_t)^{(k)}$ and where $\zeta^*_{\text{2D}} = \zeta_{\text{2D}}/(\zeta_{\text{2D}} + \zeta_{\text{3D}})$, and $\zeta^*_{\text{3D}} = 1 - \zeta^*_{\text{2D}}$ are normalized consistency measures. In addition, we ignore pseudo labels whose maximum consistency measure over the two modalities, *i.e.*, $\max(\zeta_{\text{2D}}, \zeta_{\text{3D}})$, is below a threshold $\theta^{(k)}$. Formally, our MM-TTA objective to use the generated pseudo label $\hat{y}^{\text{Ens}}_t$ ($\hat{y}^H_t$ or $\hat{y}^S_t$) for updating batch norm statistics is:

$$L_{\text{mm-tta}}(x_t) = L_{\text{seg}}(p(x^{\text{2D}}_t), \hat{y}^{\text{Ens}}_t) + L_{\text{seg}}(p(x^{\text{3D}}_t), \hat{y}^{\text{Ens}}_t). \tag{13}$$

4. Experimental Results

4.1. Datasets and Settings

We evaluate our proposed MM-TTA on several scenarios where test-time adaptation is necessary. First, sensor setups of camera and LiDAR are different between training and test data in real-world, where we adopt the benchmark ***A2D2-to-SemanticKITTI***. In particular, A2D2 [7] provides a 2.3 MegaPixels (MP) camera and 16 channels of LiDAR, while SemanticKITTI [2] uses a 0.7MP camera and 64 channels of LiDAR. This difference in hardware specification can cause unpredictable domain shift in the real-world so that the pre-trained model on source needs to be quickly adapted to the incoming test data. Second, another real-world case is ***nuScenes Day-to-Night***, where we use nuScenes [3] for this adaptation scenario. LiDAR is an active sensor that emits laser beams that are mostly invariant to lighting conditions. However, images captured by day and night are obviously different in color distribution, leading to a performance degradation without any adaptation.

		A2D2 → SemanticKITTI			Synthia → SemanticKITTI			nuScenes Day → Night		
Method	Adapt	2D	3D	Softmax avg	2D	3D	Softmax avg	2D	3D	Softmax avg
Source-only	-	37.4	35.3	41.5	21.1	25.9	28.2	42.2	41.2	47.8
xMUDA [13]	UDA	36.8	43.3	42.9	25.6	30.3	33.4	46.2	44.2	50.0
xMUDA$_{PL}$ offline [13]		43.7	48.5	49.1	25.4	33.9	35.3	47.1	46.7	50.8
TENT [30] - Eq.(2)	TTA	39.2	36.6	40.8	25.3	23.8	27.8	39.0	43.6	43.0
TENT$_{Ens}$ - Eq.(2)		39.6	36.6	41.1	27.7	23.8	29.7	39.5	43.7	43.5
xMUDA - Eq.(3)		37.5	38.0	40.2	24.0	24.1	28.0	41.7	43.9	47.0
xMUDA+TENT - Eq.(2),(3)		38.1	37.5	40.5	24.4	24.0	28.0	41.8	**44.0**	43.5
xMUDA+TENT$_{Ens}$ - Eq.(2),(3)		37.5	38.0	40.2	24.1	24.1	28.0	40.9	43.9	43.0
xMUDA$_{PL}$ - Eq.(3),(5)		36.5	39.5	42.9	24.2	25.0	29.0	40.8	43.6	45.2
xMUDA$_{PL}$+TENT$_{Ens}$ - Eq.(2),(3),(5)		37.0	40.0	43.0	24.3	25.0	29.0	41.3	43.6	46.0
MM-TTA (*Hard Select* - Eq.(11))	TTA	43.3	42.4	47.0	31.4	29.9	**35.2**	42.6	43.6	51.1
MM-TTA (*Soft Select* - Eq.(12))		**43.7**	**42.5**	**47.1**	**31.5**	**30.0**	35.1	**44.2**	43.7	**51.8**

Table 1. Quantitative comparisons with UDA methods and TTA baselines for multi-modal 3D semantic segmentation.

Finally, we evaluate test-time adaptation between synthetic and real data using ***Synthia-to-SemanticKITTI***, which is a challenging benchmark that needs to handle a significant domain shift not only in camera (style gap due to the lack of photorealism in synthetic data) but also in LiDAR (point distribution and depth accuracy).

For A2D2-to-SemanticKITTI and nuScenes Day-to-Night, we follow the dataset setting in xMUDA [13]. For Synthia-to-SemanticKITTI, we newly organize Synthia [22] by constructing point clouds with provided image and depth ground truth. Since depth maps are dense, we randomly sample pixels to obtain corresponding point clouds. Details are provided in the supplementary material.

4.2. Implementation Details

Multi-modal model: we follow xMUDA [13] to construct the two-stream multi-modal framework. For the 2D branch, we adopt U-Net [21] with a ResNet34 [10] encoder. For the 3D branch, we use a U-Net (downsampling 6-times) that utilizes sparse convolution [9] on the voxelized point cloud input, where we use either SparseConvNet [8] or MinkowskiNet [5] for our settings[1]. For each setting, all the baseline comparisons are evaluated using the same framework and backbone models.

Pre-training with source data: we directly utilize the source pre-trained model from the xMUDA official code for fair comparisons when we use the SparseConvnet. On the other hand, we train the MinkowskiNet on source data from scratch. To reproduce similar performance only using the source as in xMUDA, we use the Adam optimizer with learning rate of 1x10^{-3} for the 2D model, and SGD momentum with learning rate of 2.4x10^{-1} for the 3D model.

Test-time adaptation on target data: TTA [30] only optimizes for batch norm affine parameters during training and then reports performance after 1 epoch of adaptation. We adopt the same setting for all the baselines and our method, where we use batch statistics to compute the normalization parameters at test time. To implement our slow-fast modeling strategy in Intra-PG, we first copy the source pre-trained model and then gradually update batch norm statistics during adaptation with a momentum from the fast model.

4.3. Main Results

In this section, we show quantitative evaluations on the aforementioned three benchmark settings by reporting mIoU on predictions of 2D, 3D and ensembling between their predicted probabilities (see Table 1). For each benchmark setting, we mainly compare our method with Test-Time Adaptation (TTA) baselines, while also reporting results for xMUDA that uses Unsupervised Domain Adaptation (UDA) as references, which can access both source and target data without the budget constraint during training.

Baselines. For UDA, we compare with the multi-modal xMUDA framework that utilizes consistency loss (**xMUDA**) and self-training using offline pseudo-labels (**xMUDA$_{PL}$ offline**). For TTA baselines, we evaluate **TENT**, **xMUDA**, **xMUDA$_{PL}$**, as introduced in Sec. 3.2. Then, we extend TENT to multiple modalities (**TENT$_{Ens}$**), where we do entropy minimization on the ensemble of 2D and 3D logits. We also include the combinations of these methods, xMUDA+TENT, xMUDA+TENT$_{Ens}$, and xMUDA$_{PL}$+TENT$_{Ens}$. For all methods, we do a hyperparameter search and report best results.

Results. In Table 1, we show that our MM-TTA methods (both *Hard Select* and *Soft Select*) perform favorably against all the TTA baselines in three benchmark settings. For TTA baselines on A2D2-to-SemanticKITTI and Synthia-to-SemanticKITTI, we find that entropy and pseudo-labeling

[1] For the A2D2-to-SemanticKITTI and Synthia-to-SemanticKITTI settings, we find that there is a reported implementation issue of SparseConvnet in xMUDA, and thus we use MinkowskiNet in other 3D segmentation repostitory [26] for stability. For nuScenes Day-to-Night, we use SparseConvNet as in xMUDA [13].

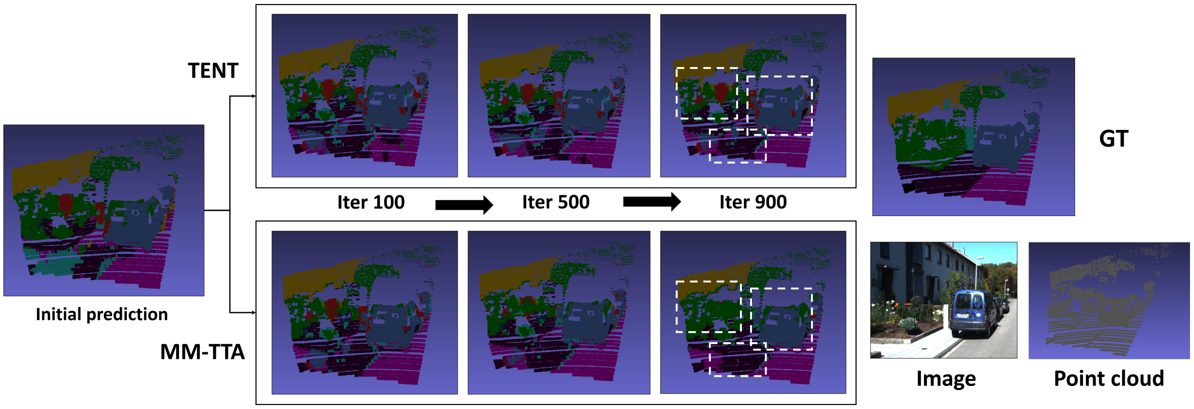

Figure 3. **Example results of our MM-TTA during test-time adaptation for gradual improvement.** While TENT [30] shows little improvements during adaptation, our method can effectively suppress the noise and achieve visually similar results to the ground truth, especially within the area of dotted white boxes.

	Method	Intra-PG		Inter-PR		Thres. on pseudo-label	2D	3D	Softmax avg
		Fast	Slow	Fusion	Select				
Pseudo	(1)	✓					39.2	36.7	40.8
	(2)	✓	✓				40.1	37.6	41.9
	(3)	✓	✓	Consensus			40.8	39	41.8
	(4)	✓	✓	Consensus		✓	40.8	37.5	43.8
	(5)	✓	✓	Merge		✓	43.1	37.4	45.3
	(6)	✓		Merge		✓	39.3	36.7	41.6
	(7)	✓	✓		Entropy	✓	40.2	39.6	43.4
	MM-TTA	✓	✓		Consistency (Hard)	✓	43.3	42.4	47.0
	MM-TTA	✓	✓		Consistency (Soft)	✓	43.7	42.5	47.1

Table 2. **Ablation study on effects of Intra-PG and Inter-PR in the A2D2 → SemanticKITTI benchmark.** We provide two variants with different fusion: 1) Consensus: using pseudo-labels that are consistent between 2D and 3D, and 2) Merge: taking the mean of two output probabilities. For the selection process, "Entropy" calculates and compares the entropy of 2D and 3D predictions.

based methods (e.g., TENT, xMUDA$_{PL}$) perform better than the consistency loss (e.g., xMUDA), due to the difficulty of capturing the correct consistency across modalities. In addition, although some TTA baselines (*e.g.*, TENT$_{Ens}$, xMUDA$_{PL}$) improve the performance of individual 2D and 3D predictions, the ensemble results are all worse than the "Source-only" model. This is because these methods do not have a well-designed module to jointly consider multi-modal outputs, where we use our Inter-PR to adaptively generate cross-modal pseudo-labels.

For nuScenes Day-to-Night, different from the other settings, the domain gap is larger for RGB than for LiDAR, and thereby the challenge mainly lies in how to improve the 2D branch and obtain effective ensemble results. For all the baselines and our methods, IoU for the 3D branch is competitive, while our results in the 2D branch and the ensemble are significantly improved, which shows the benefits of our designed Intra-PG and Inter-PR modules. Surprisingly, the ensemble results of our MM-TTA methods are better than the ones in the xMUDA approaches that use the UDA setting. This shows the effectiveness of our proposed MM-TTA framework for fast test-time adaptation. Fig. 3 show example results of 3D semantic segmentation on SemanticKITTI. Our MM-TTA method gradually improves the initial prediction throughout adaptation, and produces more complete and accurate outputs compared to TENT.

4.4. Ablation Study

4.4.1 Analysis on MM-TTA

Inter-PR for pseudo-label refinement. We show different pseudo-label refinement methods for Inter-PR, and compare them with our *Hard Select* and *Soft Select* schemes. First, in Method (4) and (5) of Table 2 respectively, we use two simple fusion techniques: 1) only using points that are consistent between pseudo-labels of 2D and 3D (Consensus), and 2) taking the mean of two output probabilities for pseudo-labeling (Merge). Second, for selecting pseudo-labels from either the 2D or 3D branch, one alternative is to calculate and compare the entropy of 2D and 3D predictions (Entropy) as in Method (7). Overall, our MM-TTA methods perform better than these model variants. In addition, we show that using the threshold on pseduo-labels is a good choice, e.g., comparing Method (3) with (4).

Method	Threshold	2D	3D	Softmax avg
Hard	0.1	40.3	41.3	45.2
	0.3	43.3	42.4	47.0
	0.5	43.2	42.5	46.7
	0.7	43	42.3	46.2
Soft	0.1	41.2	41.5	45.7
	0.3	43.7	42.5	47.1
	0.5	43.9	42.6	46.9
	0.7	43.7	42.3	46.3

(a) Pseudo-label threshold ratio $\theta^{(k)}$

Method	Momentum	2D	3D	Softmax avg
Hard	1.00	42.8	42.0	46.3
	0.99	43.3	42.4	47.0
	0.95	42.0	42.2	46.1
Soft	1.00	43.2	42.1	46.5
	0.99	43.7	42.5	47.1
	0.95	42.6	42.4	46.3

(b) Momentum factor λ

Table 3. Sensitivity analysis in A2D2 → SemanticKITTI.

Intra-PG with slow-fast modeling. We design model variants to validate the effectiveness of Intra-PG. In Method (2)/(5) of Table 2, using the slowly-updated model improves Method (1)/(6), respectively. This shows that Intra-PG is useful with different pseudo-labeling schemes, e.g., without fusion in Method (2) or "Merge" in Method (5). Note that our Inter-PR module requires slow-fast modeling and thus these two modules are coupled together as our final model, which shows performance gains compared to other variants.

4.4.2 Sensitivity Analysis

Threshold $\theta^{(k)}$. This threshold is critical for pseudo-labeling, where low values filter more points in a class-wise manner, and vice versa. Table 3a shows the robustness of our method to $\theta^{(k)}$, and a value of 0.3 performs best.

Momentum factor λ. We use a slow-fast modeling strategy to slowly update the source pre-trained batch norm statistics with a momentum λ from the fast-updated model. Table 3b shows the effect of changing λ. Setting it as 1.0 would simply keep the source statistics and is not optimal.

Stability during TTA. Since TTA only sees the test data once during adaptation, the stability can be largely affected by hyperparameters like the learning rate. In Fig. 4, we run different methods with various learning rates, and find that our MM-TTA methods perform robustly and show a good stability during adaptation with the result of higher mean (44.2/44.3) and lower standard deviation (2.45/2.55).

4.4.3 Analysis on Pseudo-labeling Accuracy

We measure the pseudo-label accuracy at different iterations during adaptation for our proposed modules. We test 6 phases from the iterations of 100 to 600. In each phase, we collect pseudo-labels for valid points and calculate the average accuracy over all categories. In Fig. 5, we first observe that using slow-fast modeling in Intra-PG improves the accuracy from the baseline (only using the fast model) by 2%. Then, combining our proposed two modules consistently shows improvement in all iterations, with a 5% gain.

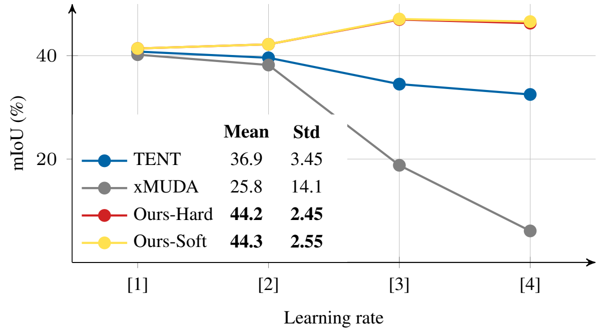

Figure 4. **Stability on using different learning rates in A2D2 → SemanticKITTI.** For the 2D/3D branch, we use four sets of learning rates: [1] 1.0x10^{-5}/2.4x10^{-5}, [2] 1.0x10^{-5}/2.4x10^{-4}, [3] 1.0x10^{-4}/2.4x10^{-4}, [4] 1.0x10^{-4}/2.4x10^{-3}.

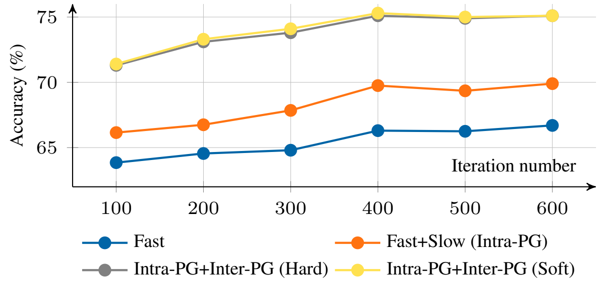

Figure 5. Pseudo-label accuracy during adaptation in A2D2 → SemanticKITTI.

5. Conclusions

In this paper, we present a new problem setting, Multi-Modal Test-Time Adaptation (MM-TTA) for 3D semantic segmentation. We first identify several baselines and their limitations, and then propose a simple yet effective self-training framework consisting of two modules, Intra-PG and Inter-PR, to produce reliable cross-modal pseudo-labels. In experiments, we demonstrate our MM-TTA framework in several benchmark settings. In addition, we provide extensive ablation studies and analysis to show the benefits of our proposed modules.

Acknowledgement This work was part of Inkyu Shin's internship at NEC Laboratories America and was also partially supported by Samsung Electronics Co., Ltd (G01200447), and under the framework of international cooperation program managed by the National Research Foundation of Korea (NRF-2020M3H8A1115028, FY2021).

References

[1] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1 – 32, 2021. 1

[2] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In *ICCV*, 2019. 5

[3] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 5

[4] Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jingshan Tang. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *CVPR*, 2021. 2

[5] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, 2019. 1, 3, 6

[6] F. Duerr, H. Weigel, M. Maehlisch, and J. Beyerer. Iterative deep fusion for 3d semantic segmentation. In *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, pages 391–397, Los Alamitos, CA, USA, nov 2020. IEEE Computer Society. 1, 3

[7] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 5

[8] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018. 6

[9] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 3, 6

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[11] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and¡ 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016. 3

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 3

[13] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Émilie Wirbel, and Patrick Pérez. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *CVPR*, 2020. 1, 2, 3, 4, 6

[14] Donghyun Kim, Yi-Hsuan Tsai, Bingbing Zhuang, Xiang Yu, Stan Sclaroff, Kate Saenko, and Manmohan Chandraker. Learning cross-modal contrastive features for video domain adaptation. In *ICCV*, 2021. 3

[15] Yizhuo Li, Miao Hao, Zonglin Di, Nitesh Bharadwaj Gundavarapu, and Xiaolong Wang. Test-time personalization with a transformer for human pose estimation. *NeurIPS*, 2021. 2

[16] Khaled El Madawy, Hazem Rashed, Ahmad El Sallab, Omar Nasr, Hanan Kamel, and Senthil Yogamani. Rgb and lidar fusion based 3d semantic segmentation for autonomous driving. In *IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019. 1

[17] Gregory P. Meyer, Jake Charland, Darshan Hegde, Ankit Laddha, and Carlos Vallespi-Gonzalez. Sensor fusion for joint 3d object detection and semantic segmentation. In *CVPR Workshop on Autonomous Driving*, 2019. 1, 3

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

[19] Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. S4t: Source-free domain adaptation for semantic segmentation via self-supervised selective self-training. *arXiv preprint arXiv:2107.10140*, 2021. 1, 2

[20] Christoph B. Rist, Markus Enzweiler, and Dariu M. Gavrila. Cross-sensor deep domain adaptation for lidar detection and segmentation. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1535–1542, 2019. 3

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 6

[22] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*, 2016. 6

[23] Khaled Saleh, Ahmed Abobakr, Mohammed Attia, Julie Iskander, Darius Nahavandi, and Mohammed Hossny. Domain adaptation for vehicle detection from bird's eye view lidar point cloud data. In *IEEE SMC*, 2019. 3

[24] Inkyu Shin, Sanghyun Woo, Fei Pan, and In So Kweon. Two-phase pseudo label densification for self-training based domain adaptation. In *ECCV*, 2020. 3

[25] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *ICML*, 2020. 2

[26] Haotian* Tang, Zhijian* Liu, Shengyu Zhao, Yujun Lin, Ji Lin, Hanrui Wang, and Song Han. Searching efficient 3d architectures with sparse point-voxel convolution. In *ECCV*, 2020. 6

[27] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *ICCV*, 2019. 1, 3

[28] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, 2018. 1, 3

[29] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 3

[30] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021. 1, 2, 3, 6, 7

[31] Yuan Wang, Tianyue Shi, Peng Yun, Lei Tai, and Ming Liu. Pointseg: Real-time semantic segmentation based on 3d lidar point cloud. *arXiv preprint arXiv:1807.06288*, 2018. 3

[32] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. *ICRA*, 2018. 1, 3

[33] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *ICRA*, 2019. 3

[34] Jianyun Xu, Ruixiang Zhang, Jian Dou, Yushi Zhu, Jie Sun, and Shiliang Pu. Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation. In *ICCV*, 2021. 1, 3

[35] Li Yi, Boqing Gong, and Thomas Funkhouser. Complete & label: A domain adaptation approach to semantic segmentation of lidar point clouds. In *CVPR*, 2021. 1, 3

[36] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018. 3

[37] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training. In *ICCV*, 2019. 3