—Supplementary Material—
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

Appendix

In this supplementary material, we provide,

A) Dataset construction

B) Algorithm for our MM-TTA

C) Oracle test

D) Analysis on class-wise adaptation

E) Qualitative results for pseudo labels

F) More qualitative results

G) Limitations

A. Dataset Construction

A.1. A2D2, SemanticKITTI and nuScenes

We strictly follow the dataset setting of A2D2/SemanticKITTI/nuScenes from xMUDA [1]. For A2D2 and SemanticKITTI, 10 classes are shared and used to train and test in each dataset. The 10 classes are car, truck, bike, person, road, parking, sidewalk, building nature, other-objects.

A.2. Synthia

We re-organize the synthia dataset [2] to simulate synthetic-to-real setting. It initially contains 9,000 RGB images with corresponding labels of segmentation and depth. Since every pixel in RGB images can be formatted into point cloud with depth ground truth, we randomly sample about 15k number of pixels to make up the point cloud of that scene (see Fig. 1). We use all of the data as the training set and merge 23 classes into 10 categories to be shared with the SemanticKITTI dataset.

B. Algorithm for MM-TTA

Here, we provide an algorithm for MM-TTA consisting of the proposed two modules: Intra-PG and Inter-PR in Alg. 1.

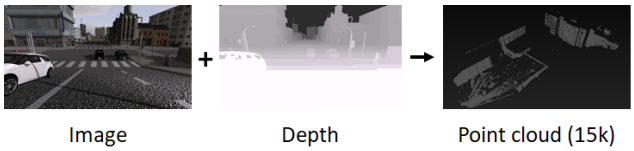

Figure 1. Construction of the Synthia dataset to generate point clouds (15k points). In that sense, we can simulate the multi-modal dataset.

C. Oracle Test

We provide two kinds of oracle tests: 1) Oracle TTA: only BN parameters of 2D/3D models are finetuned using the real target label during 1 epoch. 2) Oracle Full: all layers are updated with real target label from scratch during 30 epochs. In both cases, our MM-TTA is able to generate reliable pseudo labels, where the performance is achieved closer to that of oracle. Specifically, on nuScenes Day $\rightarrow$ Night, our MM-TTA using *Soft Select* obtains comparable results to "Oracle TTA" that uses real labels.

D. Analysis on Class-wise Adaptation

We analyze how the test-time adaptation has class-wise effect while proceeding with the iterations. We visualize class-wise adaptation with t-SNE [3] and conduct an analysis comparing between our MM-TTA and the TENT baseline (see Fig. 2). We map the final logit of all of test data's points to the 2-D space via t-SNE. We observe that our MM-TTA performs better category-level feature alignment during test-time adaptation at across different iterations.

E. Qualitative Results for Pseudo Labels

In Fig. 3, we visualize the pseudo labels generated from MM-TTA and compare with other baselines. We can find that our two modules achieve more refined and accurate pseudo label that is more similar to GT.

Method	Adapt	A2D2 → SemanticKITTI			Synthia → SemanticKITTI			nuScenes Day → Night		
		2D	3D	Softmax avg	2D	3D	Softmax avg	2D	3D	Softmax avg
Source-only	-	37.4	35.3	41.5	21.1	25.9	29.3	42.2	41.2	47.8
TENT [4]	TTA	39.2	36.6	40.8	25.3	23.8	27.8	39.0	43.6	43.0
MM-TTA (*Hard Select*)	TTA	43.3	42.4	47.0	31.4	29.9	35.2	42.6	43.6	51.1
MM-TTA (*Soft Select*)	TTA	43.7	42.5	47.1	31.5	30.0	35.1	44.2	43.7	51.8
Oracle TTA	TTA	48.5	45.8	52.4	38.8	31.1	41.4	45.6	43.6	51.5
Oracle Full	-	57.9	66.6	69.5	57.9	66.6	69.5	48.6	47.1	55.2

Table 1. Quantitative results with using real target labels as oracles. Depending on whether we only finetune the batchnorm parameters or update all layers, the oracles are "Oracle TTA" and "Oracle Full".

Algorithm 1: Algorithm for MM-TTA

Input: Target data $x_t = (x_t^{2D}, x_t^{3D})$, Source pre-trained model $F^M = (F^{2D}, F^{3D})$
Output: The model with adapted weights on the target dataset $F^M = (F^{2D}, F^{3D})$

begin
  Define the slow model and copy weights
  $F \leftarrow S$
  F.train(), S.eval()
  for *1 epoch* **do**
    # 1.Intra-PG
    Fuse the outputs of slow-fast model (Eq.(7)).
    $p(x_t^M) = Fuse(S^M(x_t^M), F^M(x_t^M))$.
    Obtain aggregated pseudo labels (Eq.(8)).
    $\hat{y}_t^M = \arg\max_{k \in K} p(x_t^M)^{(k)}$
    # 2.Inter-PR
    Calculate a consistency measure between slow and fast models (Eq.(9), (10)).
    $\zeta_M = Sim(S^M(x_t^M), F^M(x_t^M))$
    if *Hard Select* **then**
      Select from one of the modalities (Eq.(11).
      $\hat{y}_t^{Ens} = \begin{cases} \hat{y}_t^{2D}, & \text{if } \zeta_{2D} \geq \zeta_{3D}, \\ \hat{y}_t^{3D}, & \text{otherwise.} \end{cases}$
    else if *Soft Select* **then**
      Weighted sum from the two modalities (Eq.(12), (13)).
      $\hat{y}_t^{Ens} = \arg\max_{k \in K} p_t^{W(k)}$
      $p_t^{W(k)} = Weight(p(x_t^{2D})^{(k)}, p(x_t^{3D})^{(k)})$
    # 3.Update the model
    Update the Ω^F with Eq.(14).
    Momentum update for Ω^S with Eq.(6)
    $\Omega_{t_i}^S = (1-\lambda)\Omega_{t_i}^F + \lambda\Omega_{t_{i-1}}^S$

F. More Qualitative Results

Given several multi-modal data with image and point cloud, we visualize the qualitative 3D segmentation results of our MM-TTA and compare with other baselines (TENT and xMUDA). In all adaptation scenarios, we can observe that our MM-TTA (both *Hard Select* and *Soft Select*) shows more similar results to ground truth (GT).

G. Limitations and Discussion

Since our method focuses on selecting or giving adaptive weights between two modalities for general pseudo-label generation, one limitation is that its effectiveness may

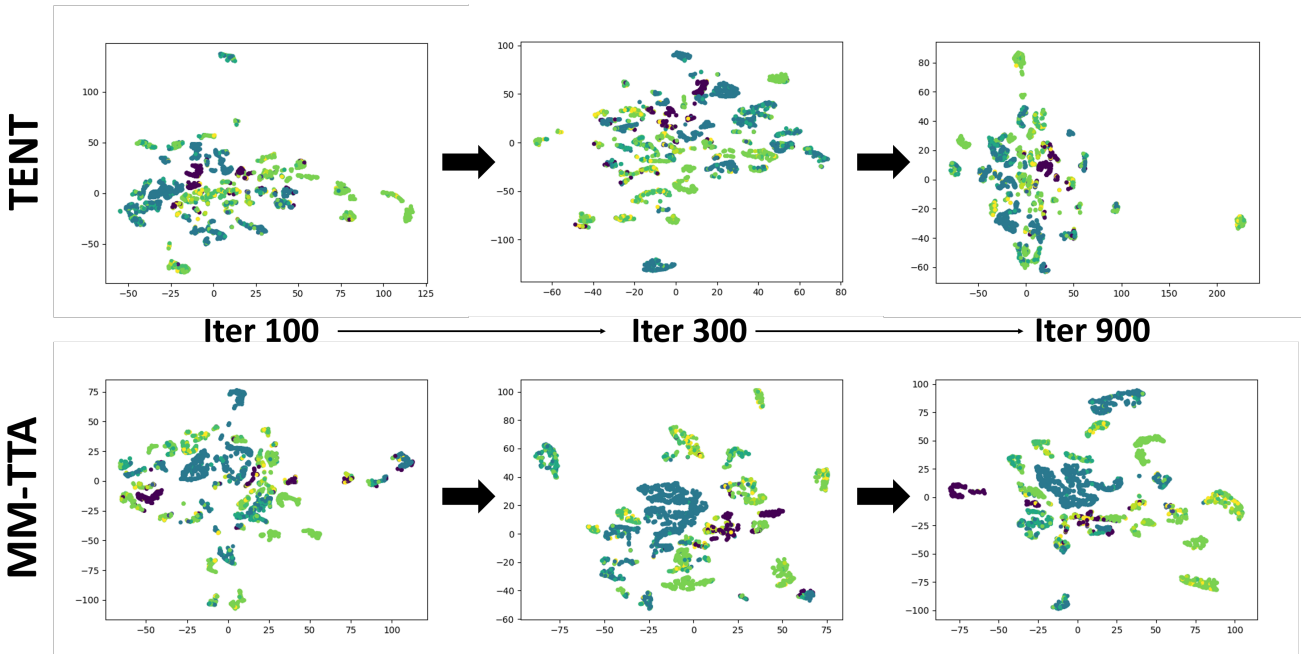

Figure 2. Qualitative results of t-SNE on TENT and MM-TTA. Each color represents one category, where our MM-TTA produces more compact clusters for each category

vary across categories. Therefore, one future direction is to develop category-aware test-time adaptation methods, so that the model can further boost the performance for certain classes that do not perform well.

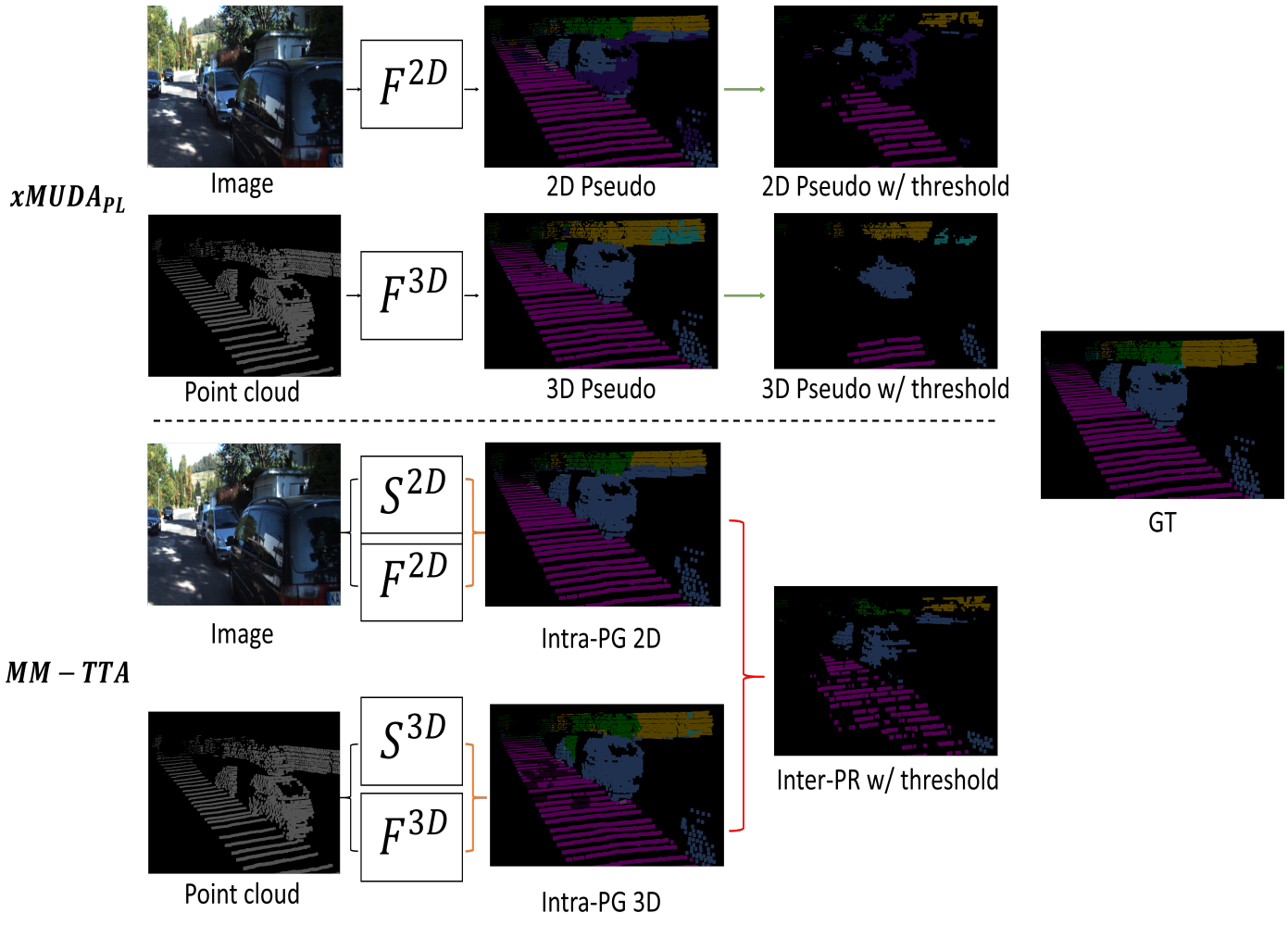

Figure 3. Qualitative results of pseudo labels on xMUDA$_{PL}$ and MM-TTA.

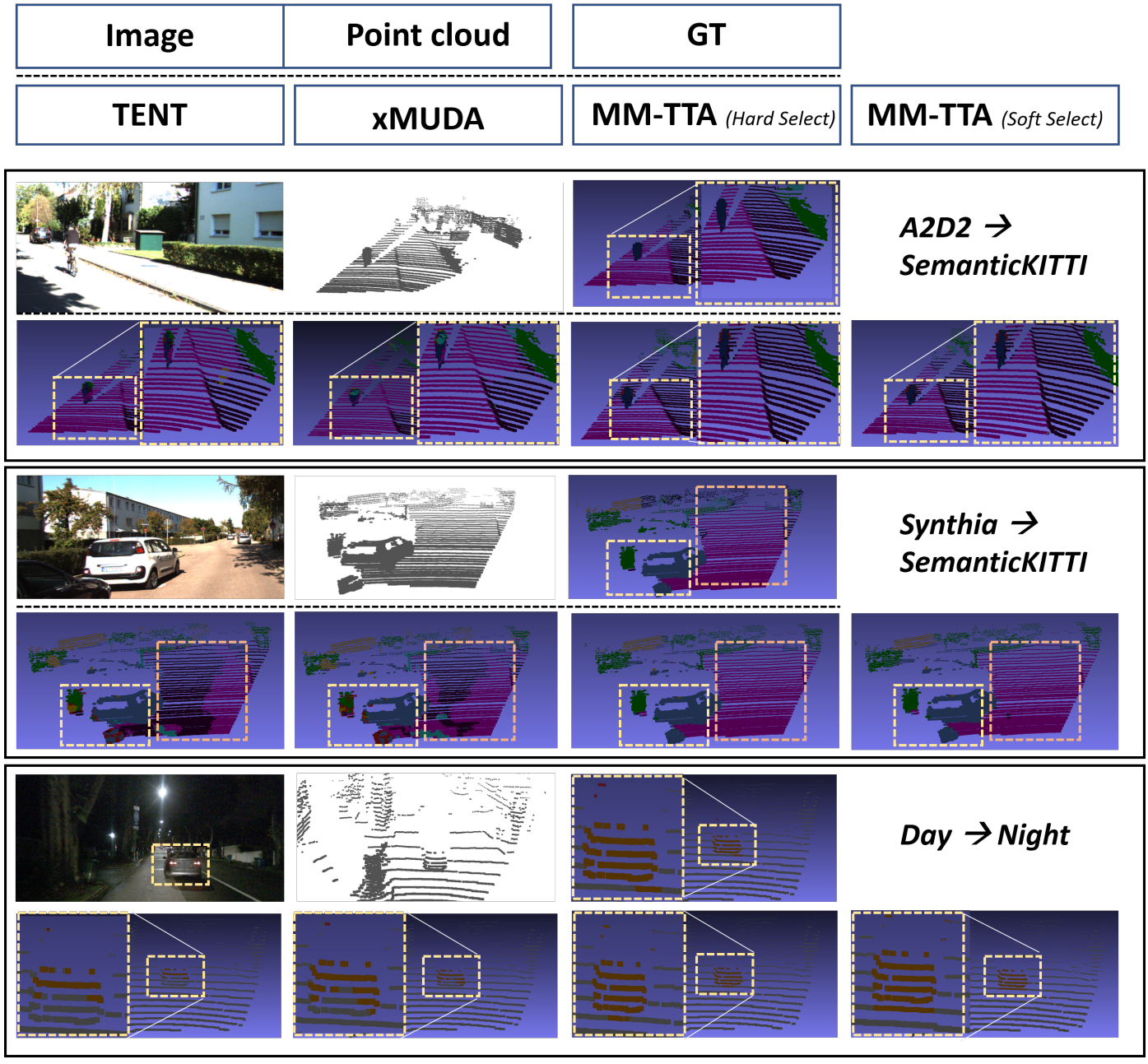

Figure 4. Qualitative 3D segmentation result of TENT, xMUDA, MM-TTA (*Hard Select*), MM-TTA (*Soft Select*) on three adaptation benchmarks.

References

[1] 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] 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. 1

[3] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 11 2008. 1

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