UniDxMD: Towards Unified Representation for Cross-Modal Unsupervised Domain Adaptation in 3D Semantic Segmentation

Supplementary Material

In this document, we provide additional content to supplement the main manuscript. Section 1 offers a detailed introduction and division of the datasets. Section 2 provides additional quantitative and qualitative results. Section 3 discusses the potential failure cases.

1. Datasets

We evaluate our method on four typical cross-modal UDA scenarios: scene layout variations (**nuScenes: USA/Sing.**), lighting changes (**nuScenes: Day/Night**), synthetic-to-real (**v.KITTI/Sem.KITTI**), and sensor setup variations (**A2D2/Sem.KITTI**). These scenarios utilize data from nuScenes-Lidarseg [2], VirtualKITTI [3], SemanticKITTI [1], and A2D2 [4]. The nuScenes-Lidarseg dataset supports both *nuScenes: USA/Sing.* and *nuScenes: Day/Night* adaptation tasks, each defined over six adaptation classes: vehicle, drivable surface, sidewalk, terrain, manmade, and vegetation. *v.KITTI/Sem.KITTI* leverages VirtualKITTI for simulated data and SemanticKITTI for real-world samples. To reconcile differences in class definitions, a mapping from xMUDA [5] is applied, standardizing the adaptation to six classes: car, trunk, road, vegetation, building, and object. *A2D2/Sem.KITTI* is constructed from A2D2 and SemanticKITTI and is intended to evaluate robustness to variations in sensor configuration and data characteristics, such as resolution and field of view. More information about the data splits and class mappings is available in Table S1. All datasets are synchronized and calibrated for both LiDAR and camera data, and only the front camera and its corresponding LiDAR points are considered for consistency.

2. Quantitative and Qualitative Results

In this section, we offer additional quantitative and qualitative results. Section 2.1 shows the performance of different VQ-based approaches in cross-modal UDA for 3D semantic segmentation. Section 2.2 presents the more qualitative results discussed in the main manuscript.

2.1. Comparison with VQ-based approaches

Here, we extend the latest VQ-based methods [6, 7] to our task by replacing our CSQM (Cluster-Based Soft Quantization Mechanism) and LSR (Latent Space Regularization) with their proposed quantizers. The results are shown in Table S2. CVQ-VAE [7] attempts to improve codebook utilization by replacing inactive code vectors with encoded features and applying a decay factor for smoothing. However, this inevitably restricts the codebook's learning capacity, resulting in only a marginal 0.8% improvement in the xM score on the nuScenes: USA/Sing. scenario. Yang et al. [6] enhances the correlation between pre- and post-quantization features by incorporating self-attention mechanisms. While effective, it does not explicitly address codebook collapse and lacks structured learning of the latent space, making it unsuitable for our task. In contrast, we propose UniDxMD, introducing CSQM to solve the problems of insufficient representation and codebook collapse caused by single-code assignment strategies, and implementing structured learning of the latent space through LSR. Our method learns equivalent semantics from different modalities and domains, and derives a unified discrete representation that addresses both modality bias and domain shift. This significantly advances the current research benchmark, providing a novel solution for cross-modal UDA.

2.2. Qualitative Results

We provide more qualitative results to illustrate the effectiveness of our UniDxMD, as shown in Fig. S1, S2, S3 and S4.

3. Potential Failure Cases

Although our method quantizes heterogeneous data into a unified latent code to extract semantically equivalent features, the features of large objects may dominate during the quantization process due to their greater number of instances available for codebook updates. As a result, the features of small objects can be suppressed, which may limit the adaptation and segmentation performance of our method for them. In future work, we plan to explore leveraging the priors of VFMs to guide the learning of the entire latent space and to further enhance the discriminative capability for small objects.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019.

[2] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022.

UDA Scenarios	Source	Target		Classes
	Train	Train	Val/Test	
nuScenes:USA/Sing.	15,695	9,665	2,770/2,929	**Vehicle** [bicycle, bus, car, construction_vehicle, motorcycle, trailer, truck]; **Driveable Surface**; **Sidewalk**; **Terrain**; **Manmade**; **Vegetation**
nuScenes:Day/Night	24,745	2,779	606/602	**Vehicle** [bicycle, bus, car, construction_vehicle, motorcycle, trailer, truck]; **Driveable Surface**; **Sidewalk**; **Terrain**; **Manmade**; **Vegetation**
v.KITTI/Sem.KITTI	2,126	18,029	1,101/4,071	**v.KITTI**: **Car**; **Truck**; **Road**; **Object** [traffic sign, traffic light, pole, misc]; **Building**; **Vegetation** [terrain, tree, vegetation] **Sem.KITTI**: **Car**; **Truck**; **Road**; **Object** [fence, pole, traffic-sign, other-object]; **Building**; **Vegetation** [vegetation, trunk, terrain]
A2D2/Sem.KITTI	27,695	18,029	1,101/4,071	**A2D2**: **Car**; **Truck**; **Bike** [bicycle, small vehicle]; **Person**; **Road**; **Parking**; **Sidewalk** [sidewalk, curbstone]; **Object**; **Building**; **Vegetation** **Sem.KITTI**: **Car**; **Truck**; **Bike** [bicycle, motorcycle, bicyclist, motorcyclist]; **Person**; **Road**; **Parking**; **Sidewalk**; **Object**; **Building**; **Vegetation** [terrain, trunk, vegetation]

Table S1. Data splits for different UDA scenarios. The nuScenes:USA/Sing., nuScenes:Day/Night, and v.KITTI/Sem.KITTI scenarios each contain six adaptation classes (in **bold**). The A2D2/Sem.KITTI scenario contains ten adaptation classes (in **bold**).

Exp	nuScenes: USA/Sing.			nuScenes: Day/Night		
	2D	3D	xM	2D	3D	xM
xMUDA [5]	64.4	63.2	69.4	55.5	69.2	67.4
CVQ-VAE[7]	65.0	64.7	70.2	59.6	69.3	68.5
Yang et al. [6]	65.6	64.5	70.6	58.7	69.1	68.3
CSQM	70.2	67.5	72.3	72.2	70.2	73.1
CSQM + LSR	73.2	68.5	74.3	73.5	71.8	74.6

Table S2. Comparison with VQ-based methods. xMUDA serves as the baseline. (mIoU↑, %).

[3] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

[4] 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] Maximilian Jaritz, Tuan-Hung Vu, Raoul De Charette, Émilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3d semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1533–1544, 2022.

[6] Zhou Yang, Weisheng Dong, Xin Li, Mengluan Huang, Yulin Sun, and Guangming Shi. Vector quantization with self-attention for quality-independent representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24438–24448, 2023.

[7] Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22798–22807, 2023.

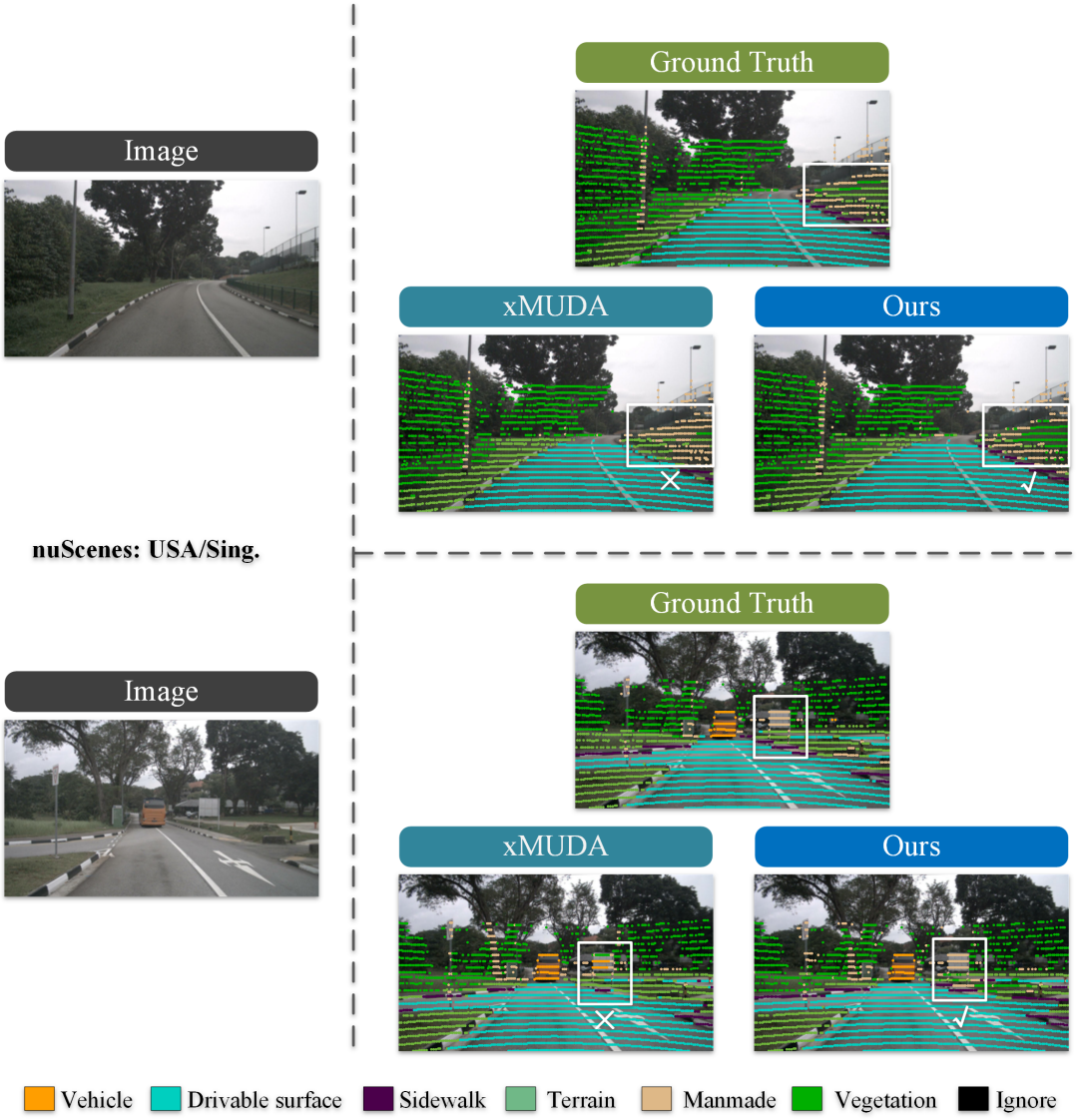

Figure S1. Additional qualitative results on *nuScenes: USA/Sing.* scenario.

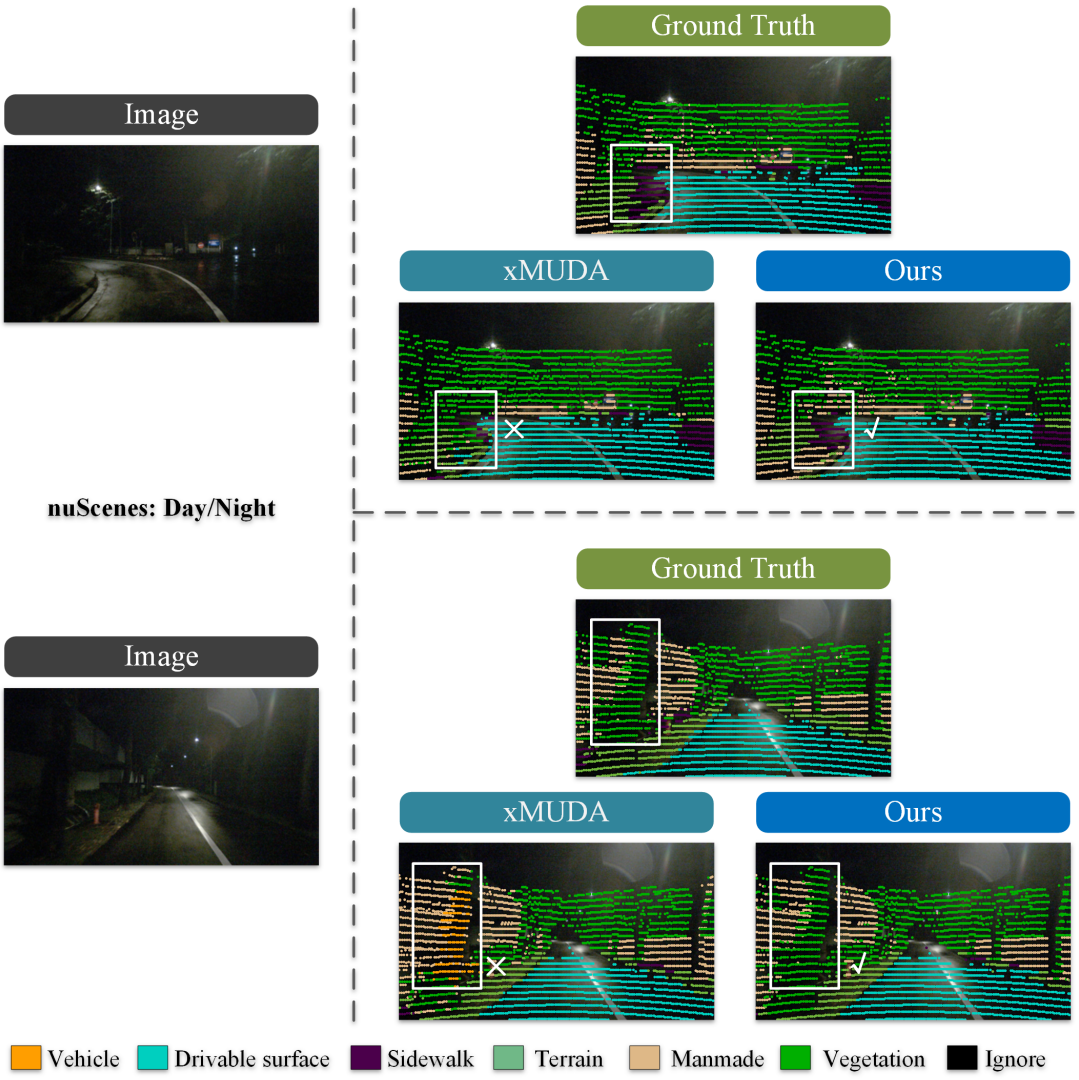

Figure S2. Additional qualitative results on *nuScenes: Day/Night* scenario.

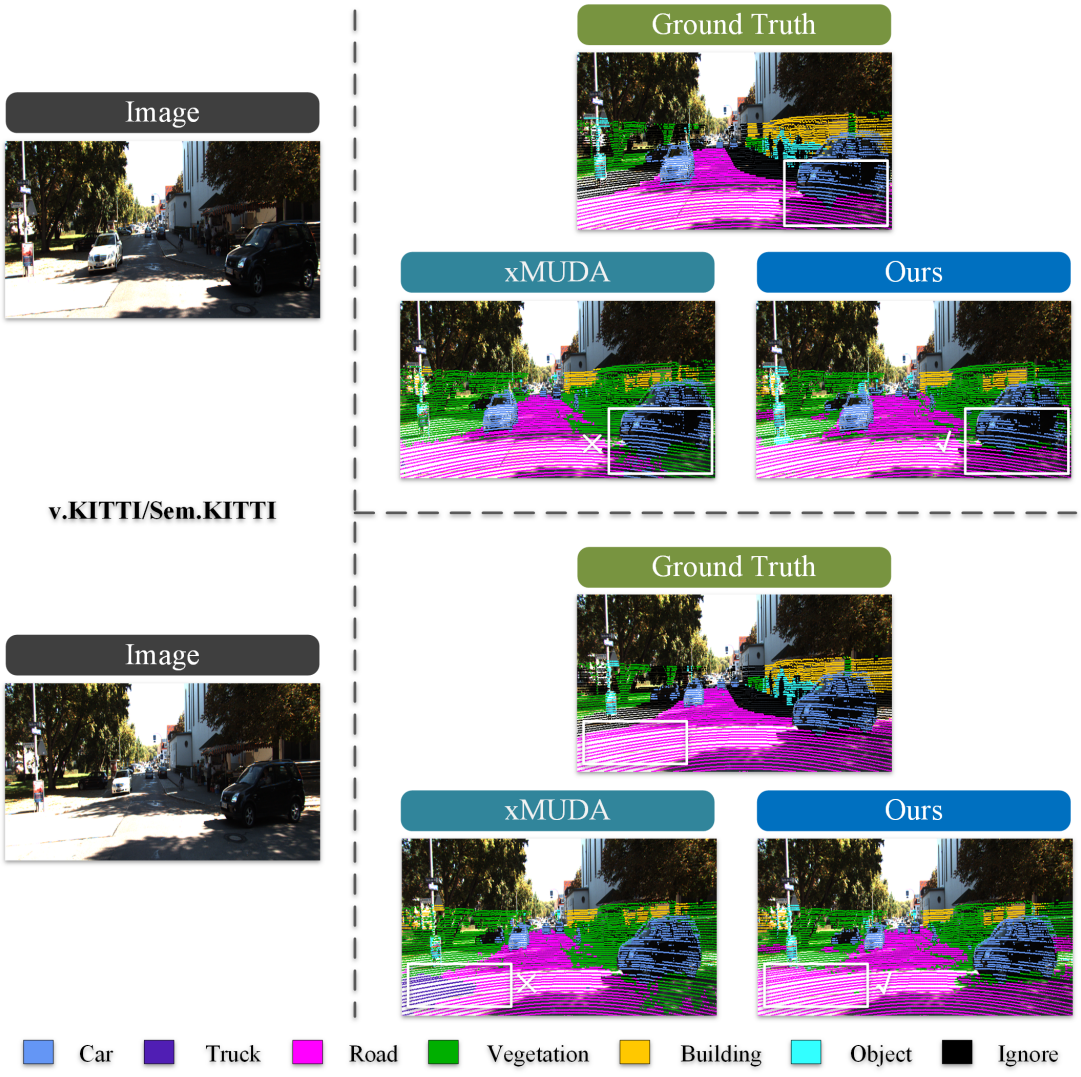

Figure S3. Additional qualitative results on *v.KITTI/Sem.KITTI* scenario.

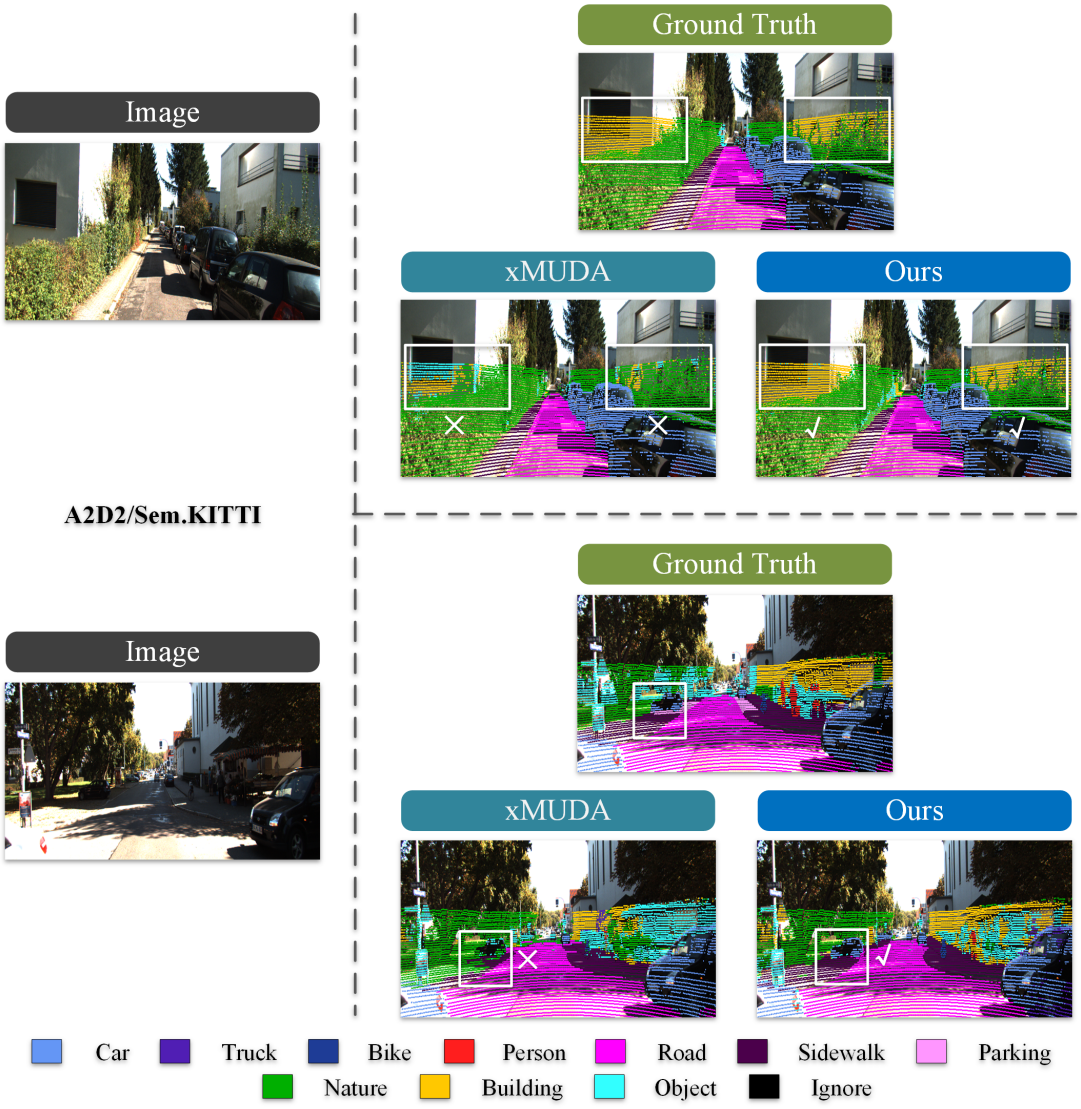

Figure S4. Additional qualitative results on *A2D2/Sem.KITTI* scenario.