

Supplementary Materials: CLIP2UDA: Making Frozen CLIP Reward Unsupervised Domain Adaptation in 3D Semantic Segmentation

Anonymous Authors

1 Overview

This material starts with more information on CLIP2UDA, including method detail, additional ablation study, error prediction under BEV, per-class ensembling results, and dataset split.

1.1 Architecture of Transformer Decoder

As shown in Fig. 1, it takes textual embedding e_t and visual embedding e_v as inputs and generates refined textual embedding \tilde{e}_t . Among \mathcal{G}_T , the output from a multi-head self-attention (MHSA) layer is used as the query for a multi-head cross-attention (MHCA) layer. The normalized operation uses LayerNorm [1]. This implementation can be written as:

$$\bar{e}_{t,q} = \text{MHSA}(e_{t,q}, e_{t,k}, e_{t,v}), \quad (1)$$

$$\bar{e}_v = \text{MHCA}(\bar{e}_{t,q}, e_{v,k}, e_{v,v}). \quad (2)$$

$$\tilde{e}_t = \bar{e}_v + \bar{e}_{t,q} + \text{FFN}(\bar{e}_v + \bar{e}_{t,q}), \quad (3)$$

where $\text{FFN}(\cdot)$ means feed-forward network.

1.2 Length of Image-specific Token

How many image-specific tokens $[\tilde{V}]_m$ should be used? And is it better to have more tokens? The results in Tab. 1 suggest having a shorter length of image-specific token benefits domain adaptation, probably due to less overfitting of source data as fewer parameters are learned.

Table 1: Length of image-specific token $[\tilde{V}]_m$.

Length	nuScenes: Day→Night		
	2D	3D	2D+3D
L=4	72.7	71.4	73.8
L=8	73.1	71.5	74.1
L=12	72.5	71.2	73.3

1.3 Per-class Ensembling Results

We present Tab. 2 to compare the per-class ensembling results of CLIP2UDA with recent MM-UDA methods. Our approach ranks first in all categories when compared to SSE [8] and BFtD [7]. Specifically, it is observed that the discrimination between “drivable surface” and “sidewalk” in nuScenes, recognition of small “object” in Virt.KITTI → Sem.KITTI significantly outperforms other methods.

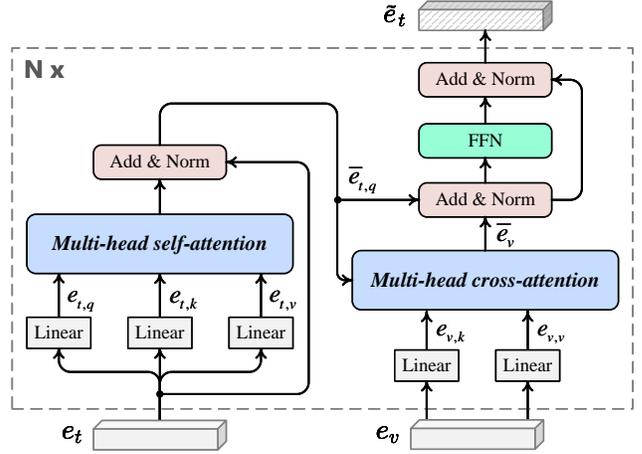


Figure 1: Architecture of Transformer decoder \mathcal{G}_T in VisPA.

Table 2: Per-class ensembling result “2D+3D”.

Method	Vehicle	Driv. Surf.	Sidewalk	Terrain	Manmade	Vegetation	mIoU(%)
nuScenes: Day → Night							
SSE [†] [8]	90.13	94.39	34.70	44.82	64.57	84.67	68.9
BFtD [†] [7]	88.86	93.53	36.62	39.81	65.85	85.35	68.3
CLIP2UDA	93.62	95.24	41.46	51.09	74.40	88.85	74.1
nuScenes: USA → Sing.							
SSE [†] [8]	82.80	96.09	43.06	51.82	63.06	78.61	69.2
BFtD [†] [7]	85.46	95.51	39.26	53.51	63.46	79.31	69.4
CLIP2UDA	88.34	96.58	48.62	58.73	69.08	82.52	74.0
Virt.KITTI → Sem.KITTI							
SSE [†] [8]	78.98	44.40	84.60	16.49	2.29	71.21	49.6
BFtD [†] [7]	75.81	52.47	84.28	17.66	0.00	78.46	51.5
CLIP2UDA	82.07	69.21	87.36	35.63	5.35	82.46	60.4

1.4 Error Prediction under BEV

In this section, we present some visualization results. Fig. 2 and Fig. 3 show the ensemble results on four adaptation scenarios. In the odd-numbered rows of the figures, we give the segmentation

Table 3: 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	Vehicle: [bicycle, bus, car, construction_vehicle, motorcycle, trailer, truck]; Driveable Surface; Sidewalk; Terrain; Manmade; Vegetation
nuScenes: USA→Singapore	15695	9665	2770/2929	
VirtualKITTI→SemanticKITTI	2126	18029	1101/4071	Car; Truck; Road; Object: [traffic sign, traffic light, pole, misc]; Building; Vegetation: [terrain, tree, vegetation]
A2D2→SemanticKITTI	27695	18029	1101/4071	Car; Truck; Bike; Person; Road; Sidewalk; Parking; Object; Building; Vegetation

result and ground truth. The even-numbered rows give points under Bird’s Eye View (BEV) that are misclassified (red points) so that we can see the position of the misclassification more clearly. It is observed that xMUDA and BfTD have a higher error recognizing small objects and region boundaries, while CLIP2UDA recognizes better thanks to the injection of multi-modal domain-invariant representation.

1.5 Dataset Split

To compose our domain adaptation scenarios, following [6], we exploit public datasets, including nuScenes [3], VirtualKITTI [4], SemanticKITTI [2], and A2D2 [5]. The split details are tabulated in Tab. 3.

1.5.1 nuScenes. It contains 1,000 scenes, each of 20 seconds, corresponding 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 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 6 categories: **Vehicle, Driveable Surface, Sidewalk, Terrain, Manmade, and Vegetation.**

1.5.2 VirtualKITTI. It consists of 5 driving scenes which are created with the Unity game engine by real-to-virtual cloning of the scenes 1, 2, 6, 18, and 20 of the real KITTI dataset. Different from real KITTI, VirtualKITTI does not simulate LiDAR, but rather provides a dense depth map, alongside semantic, instance, and flow ground truth. Each of the 5 scenes contains between 233 and 837 frames, *i.e.*, in total 2126 for the 5 scenes. Each frame is rendered with 6 different weather/lighting variants (clone, morning, sunset, overcast, fog, rain) which we use all.

1.5.3 SemanticKITTI. It 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.

1.5.4 A2D2. It 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.

Note that, we select 6 merged categories between the VirtualKITTI and SemanticKITTI, including **Car, Truck, Road, Object, Building, and Vegetation.** Between A2D2 and SemanticKITTI, 4 shared categories are considered, which are **Bike, Person, Sidewalk, and Parking.**

References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [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] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. 2016. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*. 4340–4349.
- [5] Jakob Geyer, Johannes 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).
- [6] 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.
- [7] 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 *ACM MM*. 490–498.
- [8] 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 *ACM MM*. 3338–3346.

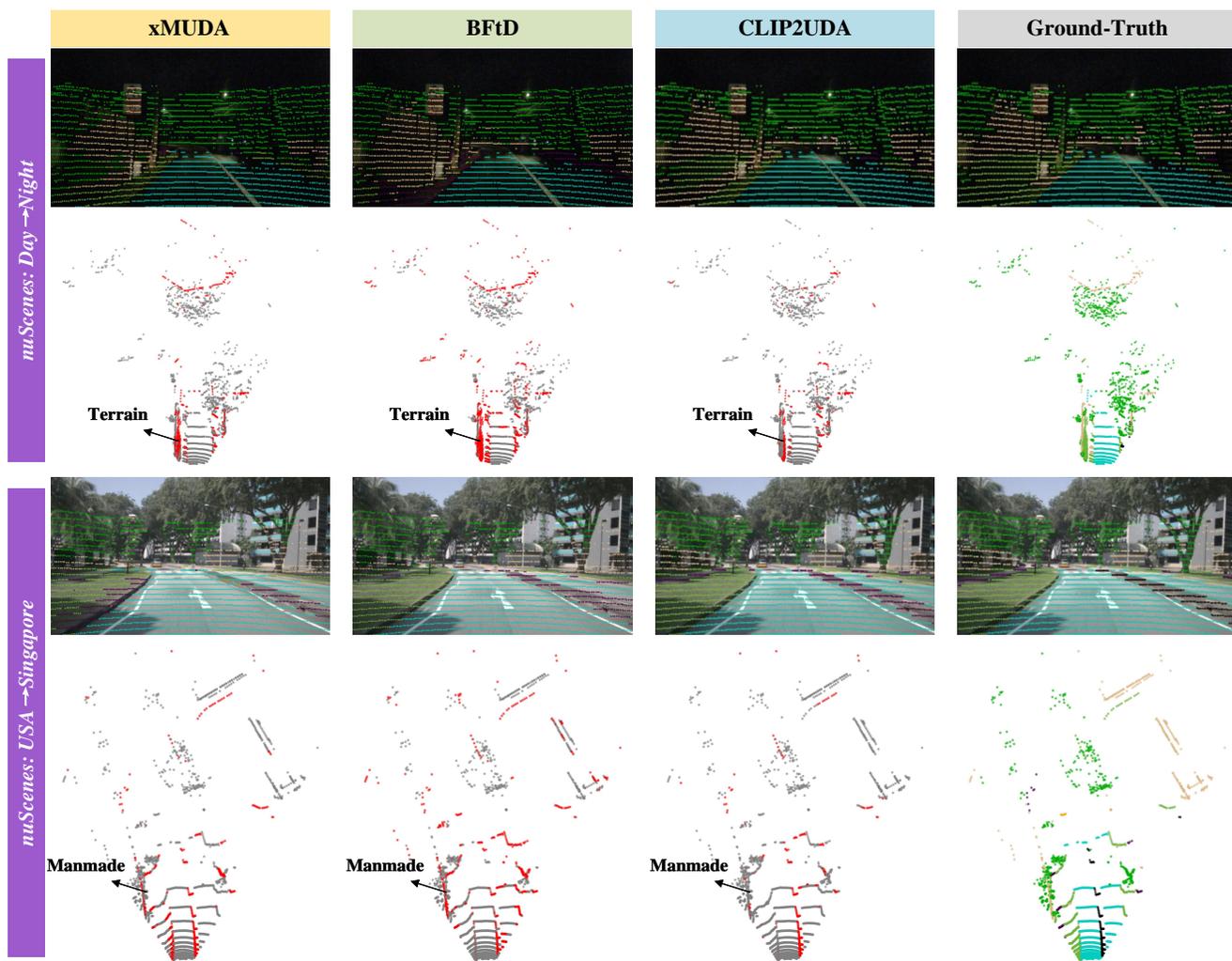


Figure 2: Qualitative results of nuScene adaptation test sets under the perspective projection (odd-numbered rows) and BEV (even-numbered rows).

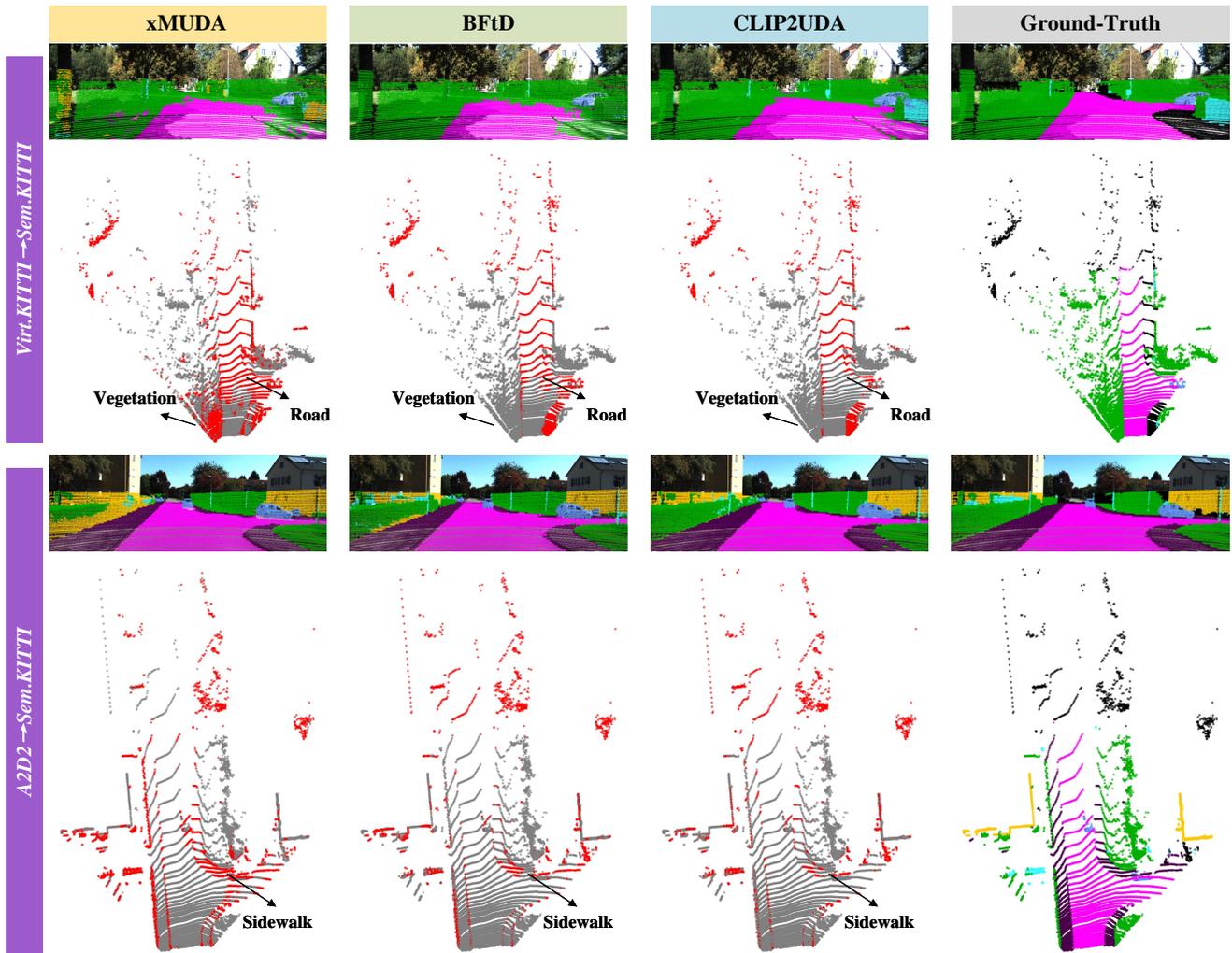


Figure 3: Qualitative results of SemanticKITTI adaptation test sets under the perspective projection (odd-numbered rows) and BEV (even-numbered rows).