# *Supplementary Material:* xMUDA: Cross-Modal Unsupervised Domain Adaptation for 3D Semantic Segmentation

Maximilian Jaritz[1,2,3], Tuan-Hung Vu[3], Raoul de Charette[1], Émilie Wirbel[2,3], Patrick Pérez[3]
[1]Inria, [2]Valeo DAR, [3]Valeo.ai

In this document we provide more details of the dataset splits used in our experiments and additional qualitative results.

## 1. Dataset Splits

### 1.1. nuScenes

The nuScenes dataset [2] consists of 1000 driving scenes, each of 20 seconds, which corresponds to 40k annotated keyframes taken at 2Hz. The scenes are split into train (28,130 keyframes), validation (6,019 keyframes) and hidden test set. The point-wise 3D semantic labels are obtained from 3D boxes like in [5]. We propose the following splits destined for domain adaptation with the respective source/target domains: Day/Night and Boston/Singapore. Therefore, we use the official validation split as test set and divide the training set into train/val for the target set (see Tab. 1 for the number of frames in each split). As the number of object instances in the target split can be very small (e.g. for night), we merge the objects into 5 categories: **vehicle** (car, truck, bus, trailer, construction vehicle), **pedestrian**, **bike** (motorcycle, bicycle), **traffic boundary** (traffic cone, barrier) and **background**.

### 1.2. A2D2 and SemanticKITTI

The A2D2 dataset [4] features 20 drives, which corresponds to 28,637 frames. The point cloud comes from three 16-layer front LiDARs (left, center, right) where the left and right front LiDARS are inclined. The semantic labeling was carried out in the 2D image for 38 classes and we compute the 3D labels by projection of the point cloud into the labeled image. We keep scene 20180807_145028 as test set and use the rest for training.

The SemanticKITTI dataset [1] provides 3D point cloud labels for the Odometry dataset of Kitti [3] which features large angle front camera and a 64-layer LiDAR. The annotation of the 28 classes has been carried out directly in 3D. We use the scenes $\{0, 1, 2, 3, 4, 5, 6, 9, 10\}$ as train set, 7 as validation and 8 as test set.

We select 10 shared classes between the 2 datasets by merging or ignoring them (see Tab. 2). The 10 final classes are car, truck, bike, person, road, parking, sidewalk, building, nature, other-objects.

| | source | | target | | |
|---|---|---|---|---|---|
| Split | train | test | train | val | test |
| Day - Night | 24,745 | 5,417 | 2,779 | 606 | 602 |
| Boston - Singapore | 15,695 | 3,090 | 9,665 | 2,770 | 2,929 |
| A2D2 - SemanticKITTI | 27,695 | 942 | 18,029 | 1,101 | 4,071 |

Table 1: Number of frames for the 3 splits.

## 2. Qualitative Results

We provide qualitative results in Fig. 1 where we show the output of the 2D and 3D stream individually to illustrate their respective strengths and weaknesses, e.g. that 3D works much better at night. We also provide a video in this supplementary that shows a driving scene in the test set of SemanticKITTI.

## References

[1] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall. SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences. In *ICCV*, 2019.

[2] 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. *arXiv preprint arXiv:1903.11027*, 2019.

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

[4] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, and Peter Schuberth. A2D2: AEV autonomous driving dataset. `http://www.a2d2.audi`, 2019.

| A2D2 class | mapped class | SemanticKITTI class | mapped class |
|---|---|---|---|
| Car 1 | car | unlabeled | ignore |
| Car 2 | car | outlier | ignore |
| Car 3 | car | car | car |
| Car 4 | car | bicycle | bike |
| Bicycle 1 | bike | bus | ignore |
| Bicycle 2 | bike | motorcycle | bike |
| Bicycle 3 | bike | on-rails | ignore |
| Bicycle 4 | bike | truck | truck |
| Pedestrian 1 | person | other-vehicle | ignore |
| Pedestrian 2 | person | person | person |
| Pedestrian 3 | person | bicyclist | bike |
| Truck 1 | truck | motorcyclist | bike |
| Truck 2 | truck | road | road |
| Truck 3 | truck | parking | parking |
| Small vehicles 1 | bike | sidewalk | sidewalk |
| Small vehicles 2 | bike | other-ground | ignore |
| Small vehicles 3 | bike | building | building |
| Traffic signal 1 | other-objects | fence | other-objects |
| Traffic signal 2 | other-objects | other-structure | ignore |
| Traffic signal 3 | other-objects | lane-marking | road |
| Traffic sign 1 | other-objects | vegetation | nature |
| Traffic sign 2 | other-objects | trunk | nature |
| Traffic sign 3 | other-objects | terrain | nature |
| Utility vehicle 1 | ignore | pole | other-objects |
| Utility vehicle 2 | ignore | traffic-sign | other-objects |
| Sidebars | other-objects | other-object | other-objects |
| Speed bumper | other-objects | moving-car | car |
| Curbstone | sidewalk | moving-bicyclist | bike |
| Solid line | road | moving-person | person |
| Irrelevant signs | other-objects | moving-motorcyclist | bike |
| Road blocks | other-objects | moving-on-rails | ignore |
| Tractor | ignore | moving-bus | ignore |
| Non-drivable street | ignore | moving-truck | truck |
| Zebra crossing | road | moving-other-vehicle | ignore |
| Obstacles / trash | other-objects | | |
| Poles | other-objects | | |
| RD restricted area | road | | |
| Animals | other-objects | | |
| Grid structure | other-objects | | |
| Signal corpus | other-objects | | |
| Drivable cobbleston | road | | |
| Electronic traffic | other-objects | | |
| Slow drive area | road | | |
| Nature object | nature | | |
| Parking area | parking | | |
| Sidewalk | sidewalk | | |
| Ego car | car | | |
| Painted driv. instr. | road | | |
| Traffic guide obj. | other-objects | | |
| Dashed line | road | | |
| RD normal street | road | | |
| Sky | ignore | | |
| Buildings | building | | |
| Blurred area | ignore | | |
| Rain dirt | ignore | | |

Table 2: Class mapping for A2D2 - SemanticKITTI UDA scenario.

[5] 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. In *ICRA*, 2018.

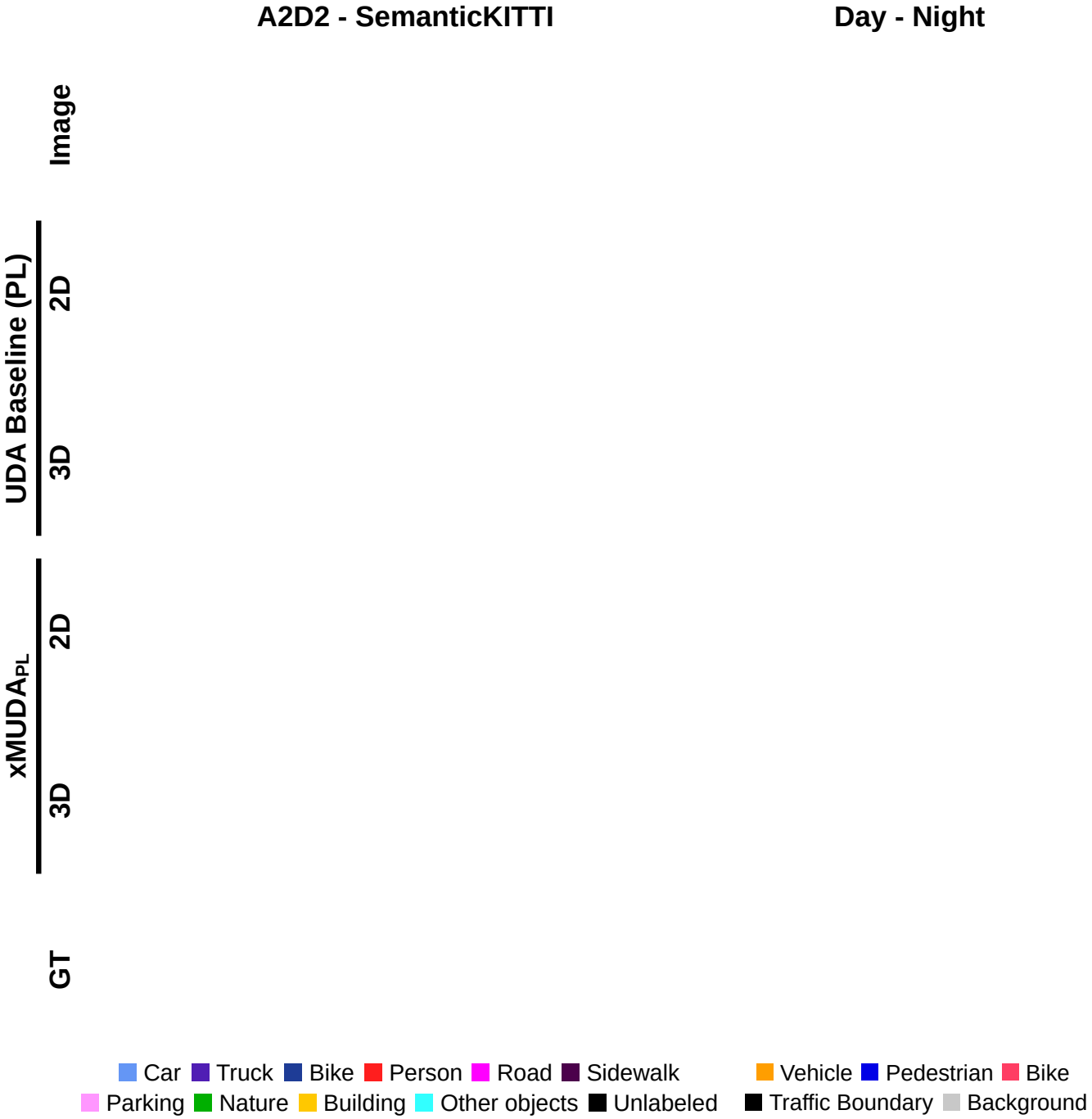


Figure 1: **Qualitative results on two UDA scenarios**. For UDA Baseline (PL) and xMUDA$_{PL}$, we separately show the predictions of the 2D and 3D network stream.
**A2D2/SemanticKITTI**: For the uni-modal UDA baseline (PL), the 2D prediction lacks consistency on the road and 3D is unable to recognize the bike and the building on the left correctly. In xMUDA$_{PL}$, both modalities can stabilize each other and obtain better performance on the bike, the road, the sidewalk and the building.
**Day/Night**: For the UDA Baseline, 2D can only partly recognize one car out of three while the 3D prediction is almost correct, with one false positive car on the left. With xMUDA$_{PL}$, the 2D and 3D predictions are both correct.