PanDA: Unsupervised Domain Adaptation for Multimodal 3D Panoptic Segmentation in Autonomous Driving

Supplementary Material

Table of Contents

A. Cross-Domain Benchmark Setting Details

A.1. Intra-Dataset Domain Shifts

For intra-dataset experiments, we follow the preprocessing pipeline of closed-domain 3D panoptic segmentation [6, 8, 11]. Different from semantic UDA methods [1, 4, 7] that use only the front camera images and corresponding point clouds, we use all available images and points. Using the official training/validation split and the nuScenes metadata, we construct three types of domain shifts: Day vs. Night, Sunny vs. Rainy, and Boston vs. Singapore. For example, in the Sunny/Rainy split, frames annotated as rainy serve as the rainy domain, while all remaining frames are assigned to the sunny domain. We retain the official class mapping with 10 thing classes and 6 stuff classes. For target domains with reduced class coverage, we normalize all metrics over the classes present in the target set. Dataset statistics are summarized in Table A, and representative LiDAR/image samples along with corruption examples are shown in Fig. A. The three intra-dataset domain shifts result in distinct modality-specific degradations. **USA vs. SG.** Both LiDAR and camera modalities are influenced by geographical and environmental discrepancies such as architectural styles, vegetation density, road layout, and weather patterns, all of which introduce substantial distribution shifts. **Sunny vs. Rainy.** Rain introduces multiple LiDAR corruptions, including spurious points, local point dropouts, unstable reflectance, and overall sparser and more irregular point patterns—particularly harmful for geometry-based segmentation and detection. Fig A column 2, row 3 and 4 provide examples of paired LiDAR and images. Images also deteriorate due to raindrops, and reduced visibility, significantly lowering contrast and impairing small or distant objects. **Day vs. Night.** Image degradation is the main degradation. Nighttime illumination is insufficient, yielding low brightness, intensified noise, and reduced texture/edge/appearance cues. See Fig. A row 3, column 3 for example.

A.2. Cross-Dataset Domain Shift

For cross-dataset experiments, the class taxonomies differ considerably between datasets. Following xMUDA [4, 5], we merge categories into 1 "thing" and 5 "stuff" classes; the exact mapping is provided in Table C. The SemanticKITTI→nuScenes transfer setting is particularly challenging, as it simultaneously introduces differences in LiDAR sensor characteristics, geolocation, and camera coverage (nuScenes provides six views, while SemanticKITTI contains only two front cameras).

A.3. Synthetic Corruption Setting

To further investigate the robustness of our proposed PanDA model under diverse modality degradations, we introduce a set of synthetic corruptions covering a wide range of practical conditions in autonomous driving. We adopt two corruption groups from nuScenes-C [2], selecting those applicable to segmentation tasks:

- **Weather Corruptions.** This group includes Snow, Rain, Fog, and Strong Sunlight, which simultaneously degrade **both** LiDAR and camera signals. For instance, fog reduces image visibility and scatters LiDAR beams. As neither modality remains fully reliable, this category represents the most challenging corruption type.
- **Sensor Corruptions.** This category includes sensor failure like LiDAR density reduction and sensor noise, which primarily affect a **single** modality while leaving the other largely intact. These corruptions test the model's ability to leverage the complementary sensor.

During the experiment, we use nuScenes-Sunny-Day, a clean subset collected during clear daytime conditions, as the source domain. For the target domain, we apply one

Table A. Overview of dataset splits and semantic categories across all cross-domain settings.

Scenarios	Source Train	Target Train	Target Val	Source Categories	Target-exclusive Categories
nuSc.:USA/SG	15,695	12,435	2,929	**Thing**: barrier, bicycle, bus, car, construction_vehicle, motorcycle, pedestrian, traffic_cone, trailer, truck; **Stuff**: driveable surface, other_flat, sidewalk, terrain, manmade, vegetation.	**Thing**: trailer
nuSc.:Sunny/Rainy	22,548	5,582	1,088		–
nuSc.:Day/Night	24,745	3,385	602		**Thing**: bus, construction_vehicle, trailer
Sem.KITTI/nuSc.	18,029	28,130	6,019	**Thing**: vehicle; **Stuff**: driveable surface, sidewalkl, terrain, manmade, vegetation.	–

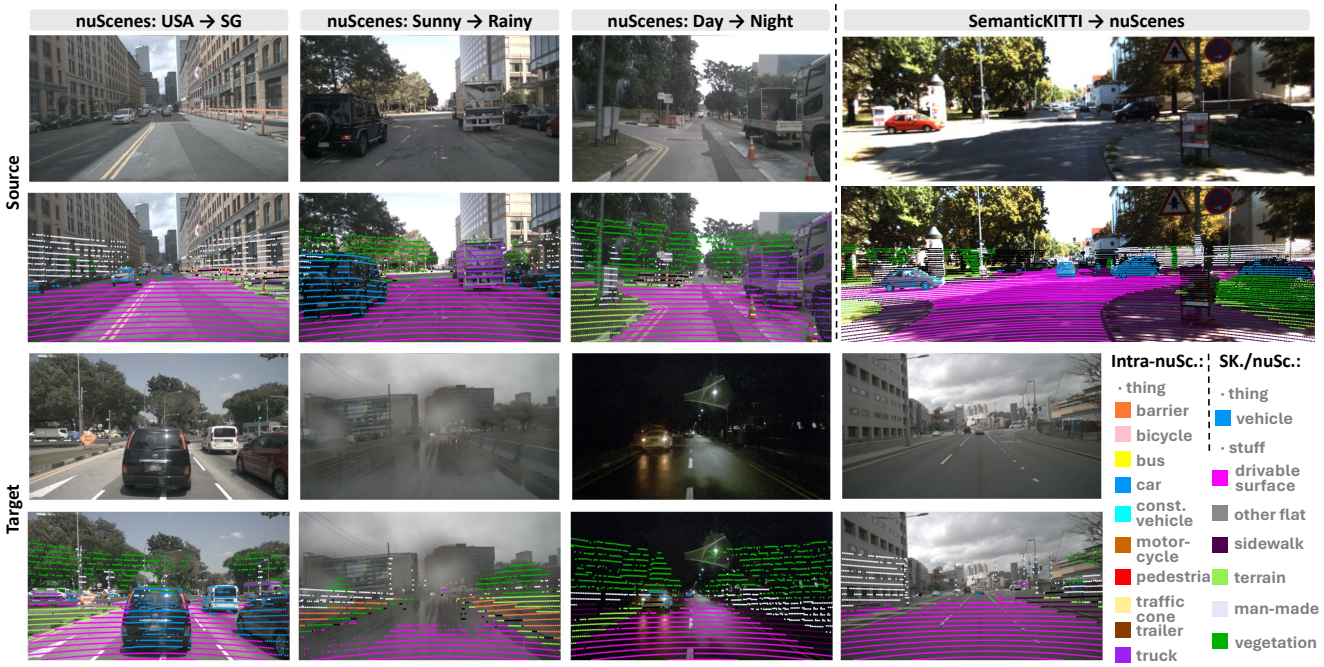

Figure A. Illustration of cross-domain benchmarks. To highlight domain discrepancies, LiDAR points are projected onto front-view images and color-coded by semantic classes. Best viewed in color.

type of corruption from the above groups to each sample by randomly selecting the corruption type and its severity level (from 1 to 5). This setup creates a controlled yet diverse corruption space for evaluating robustness. All panoptic settings and model hyperparameters remain consistent with our previous experiments to ensure a fair comparison.

Table B. Comparative study for cross-domain 3D multimodal panoptic segmentation under synthetic multi-modal corruptions. The highest UDA results are marked in **bold**.

Method	Weather				Sensor			
	PQ	PQ^{th}	PQ^{st}	mIoU	PQ	PQ^{th}	PQ^{st}	mIoU
Baseline	48.3	49.8	45.7	48.5	58.9	60.5	56.1	51.1
Pano-xMUDA	52.0	50.7	54.3	48.7	51.0	48.4	55.2	46.1
PanDA (Ours)	**64.7**	**62.3**	**68.6**	**62.2**	**60.2**	**59.6**	**61.2**	**52.4**
Oracle-Target	69.5	68.7	70.8	65.9	64.7	65.9	62.6	62.2

A.4. Results on Synthetic Corruptions

As shown in Table B, PanDA demonstrates strong robustness across all corruption types, with particularly significant gains under challenging weather corruptions, where it outperforms Pano-xMUDA by +12.7% PQ and the source-only Baseline by +16.4% PQ. This substantial improvement confirms our framework's effectiveness when both modalities are degraded simultaneously. For sensor corruptions that asymmetrically affect individual modalities, PanDA maintains a +9.2% PQ advantage over Pano-xMUDA while substantially closing the gap with Oracle-Target.

B. Visualization of Core Modules

Asymmetric Multi-modal Drop (AMD). The AMD module selectively removes informative regions across modalities to enhance cross-modal robustness and generalization. As shown in Fig. B, AMD first extracts boundary regions from 2D images using Canny edges and identifies instance-

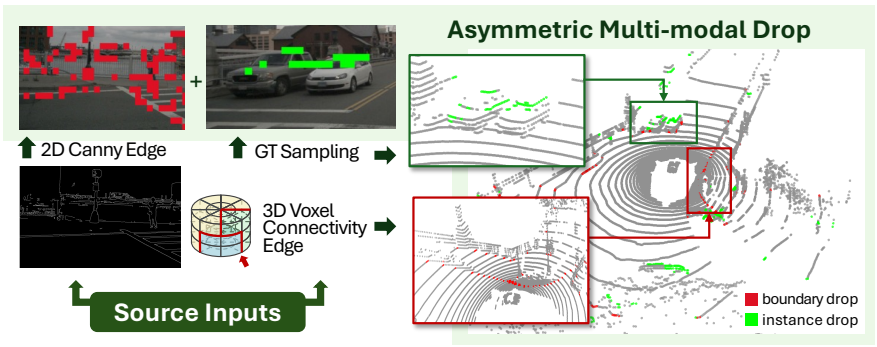

Figure B. Illustration of the AMD module, which identifies and selectively drops **boundary** and **instance-interior** regions in an asymmetric multi-modal manner.

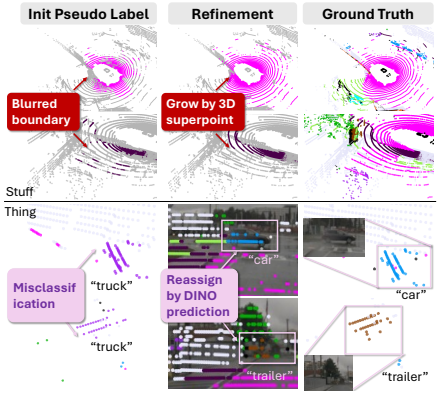

Figure C. Result visualization in DualRefine.

interior regions by sampling points based on ground-truth instance masks. In parallel, a 3D voxel connectivity–based edge detector identifies geometric boundaries in LiDAR space.

Leveraging these modality-specific cues, AMD applies asymmetric region dropout. **Boundary** drop focuses on sharp contours, which are important to both thing and stuff classes. While **instance-interior** drop targets on the instance contextual understanding. This asymmetric masking enables the model to learn structure-aware, cross-modal representations that remain stable under diverse domain shifts.

Dual-Expert Pseudo-Label Refinement (DualRefine). The DualRefine module addresses two major limitations of conventional pseudo-labeling: blurred boundaries and instance misclassification, which typically arise from hard thresholding. The visualization in Fig. C shows our dual-strategy tailored for **thing** and **stuff** classes: for stuff classes, it employs a grow mechanism that expands instance regions using matched 3D geometric superpoints, restoring complete and clear boundaries; for thing classes, it applies class reassignment based on DINO-predicted instance labels to correct and refine the predicted semantic labels. By leveraging 3D geometric consistency and instance-aware cues, DualRefine significantly improves the accuracy of pseudo-labels, leading to more precise segmentation results.

Table C. Class mapping between SemanticKITTI and nuScenes.

SemanticKITTI	mapped classes	nuScenes	mapped classes
unlabeled	ignore	ignore	ignore
outlier	ignore	barrier	ignore
car	vehicle	bicycle	vehicle
bicycle	vehicle	bus	vehicle
bus	ignore	car	vehicle
motorcycle	vehicle	construction_vehicle	vehicle
on-rails	ignore	motorcycle	vehicle
truck	vehicle	pedestrian	ignore
other-vehicle	ignore	traffic_cone	ignore
person	ignore	trailer	vehicle
bicyclist	vehicle	truck	vehicle
motorcyclist	vehicle	driveable_surface	driveable_surface
road	driveable_surface	other_flat	ignore
parking	driveable_surface	sidewalk	sidewalk
sidewalk	sidewalk	terrain	terrain
other-ground	ignore	manmade	manmade
building	manmade	vegetation	vegetation
fence	manmade		
other-structure	ignore		
lane-marking	driveable_surface		
vegetation	vegetation		
trunk	vegetation		
terrain	terrain		
pole	manmade		
traffic-sign	manmade		
other-object	manmade		
moving-car	vehicle		
moving-bicyclist	vehicle		
moving-person	ignore		
moving-motorcyclist	vehicle		
moving-on-rails	ignore		
moving-bus	ignore		
moving-truck	vehicle		
moving-other-vehicle	ignore		

C. Extended Comparison and Ablation Studies

C.1. Domain Analysis of AMD

Table D evaluates the effectiveness of our Asymmetric Multimodal Drop (AMD) module under different application strategies. We compare applying AMD only to the source domain, only to the target domain (as a strong perturbation for the student model, similar to MIC [3]), and to both domains (following preliminary UDA methods MIC-Drop [9]). Across all domain shifts, our source-only AMD (row 1) significantly surpasses the target-only (row 2) by +1.5%, +0.6% and +3.7% on PQ. Our strategy also consistently outperforms dual-domain variants (row 3), especially on Day/Night shift. This superiority stems from the **intrinsic difference** between domains: clean, fully annotated source data provides reliable and stable supervision, enabling the model to learn domain-invariant representations. In contrast, augmenting noisy and unlabeled target data introduces additional uncertainty and may further am-

Table D. Ablation study of asymmetric training module. "Src."/"Tgt." denote applying AMD on source/target domain.

Domain		USA/SG		Sunny/Rainy		Day/Night	
Src.	Tgt.	PQ	mIoU	PQ	mIoU	PQ	mIoU
✓		**77.3**	72.3	**72.4**	**67.9**	**73.1**	**59.9**
	✓	75.8	72.1	71.8	65.9	69.4	58.6
✓	✓	77.0	**72.6**	72.3	67.1	71.7	58.3

Table E. Comparison of masking strategies ("Mask") and multimodal interaction modes ("Sync") across domain shifts.

Mask	Sync	USA/SG			Sunny/Rainy			Day/Night		
		PQ	PQ^{th}	PQ^{st}	PQ	PQ^{th}	PQ^{st}	PQ	PQ^{th}	PQ^{st}
Rand	Sym	76.8	78.7	74.0	72.0	71.6	**72.8**	70.0	74.0	65.2
	Comp	76.8	78.6	74.1	72.0	71.9	72.1	70.4	74.7	65.3
	Asym	77.0	78.9	74.1	72.3	72.0	72.6	70.8	76.7	63.9
AMD	Asym	**77.3**	**79.3**	**74.2**	**72.4**	**72.3**	72.4	**73.1**	**79.6**	**65.5**

plify the domain discrepancy, ultimately undermining adaptation performance.

C.2. Augmentation Strategy Comparison

We compare our AMD module with masking-based UDA approaches that typically rely on random patch selection, such as MIC, MICDrop, and Mx2M [10]. For a fair comparison, we first extend random masking to the multimodal setting by additionally removing the LiDAR points projected onto the masked image regions. We further evaluate three multimodal interaction schemes: (1) synchronous masking, which drops the same spatial regions in both modalities; (2) complementary masking, where the two modalities compensate for each other following MICDrop; and (3) asymmetric masking, as adopted in our AMD design. All experiments are conducted on the source domain with the same hyperparameters.

Table E shows that our method (in row 4) consistently outperforms random masking under all interaction schemes and across all three domain shifts, with notable improvements on thing classes (e.g., gaining up to +5.6% on PQ^{th} for Day/Night). This confirms that selectively dropping boundary- and instance-sensitive regions yields **more discriminative panoptic representations**. Compared with the other interaction schemes, synchronous augmentation distorts the spatial integrity of multimodal content, while the complementary strategy introduces excessive information loss. Our AMD, with its selective and asymmetric design, maintains learning stability while achieving more effective domain adaptation.

C.3. Impact of EMA Teacher

To examine the impact of the EMA teacher, we introduce an EMA teacher into Pano-xMUDA and Pano-UniDSeg, respectively. As shown in Table F, compared with the original models, EMA slightly improves performance under the geolocation shift (USA/Singapore). However, both methods suffer significant performance degradation under the Day/Night shift, where the domain gap is substantially larger. We attribute this behavior to error accumulation caused by the huge domain discrepancy. In contrast, PanDA effectively mitigates this issue via enhancing cross-modal complementarity and refining the pseudo-label strategy, consistently outperforming all baselines.

Table F. Performance of existing methods with an EMA teacher under different domain shifts. Numbers in parentheses denote the change relative to the original models without EMA.

Method	USA/SG	Sunny/Rainy	Day/Night
Pano-xMUDA + EMA	73.4(↑6.2)	64.8 (↑2.6)	61.0 (↓ 8.7)
Pano-UniDSeg + EMA	75.0(↑2.1)	65.4 (↓ 0.1)	65.8 (↓ 4.7)

D. Additional Qualitative Results

Fig. D presents extended panoptic error map visualizations, complementing Fig.3 in the main paper. Denser colored regions indicate more prediction errors. We highlight representative boundary and instance segmentation results with red and gold boxes, respectively.

Our method demonstrates clear advantages over xMUDA and UniDSeg, showing not only improved recognition quality (fewer **false positives** and **false negatives**) but also enhanced segmentation quality (reduced **mismatched** regions). Specifically, our predictions exhibit clearer boundaries and more complete instance structures. Under Sunny/Rainy shifts where both LiDAR and image modalities degrade, our approach maintains robustness against sparse point clouds (red box, first row) and blurred objects (gold box, second row). For Day/Night adaptation, where nighttime images provide limited information—particularly along object boundaries—our method effectively leverages LiDAR data to maintain reliable perception. In USA/SG shifts, even though no significant modality degradation was showcased, our framework shows consistent generalization when recognizing objects with different appearances and distributions.

E. Computational Efficiency

We evaluate model complexity and module latency during training against the base model IAL [6]. As shown in Table G, our method (combining AMD and DualRefine modules) maintains a model size and training latency comparable to Pano-xMUDA, while being significantly smaller and faster than Pano-UniDSeg. All latencies were measured on an NVIDIA H100 GPU with a batch size of 1.

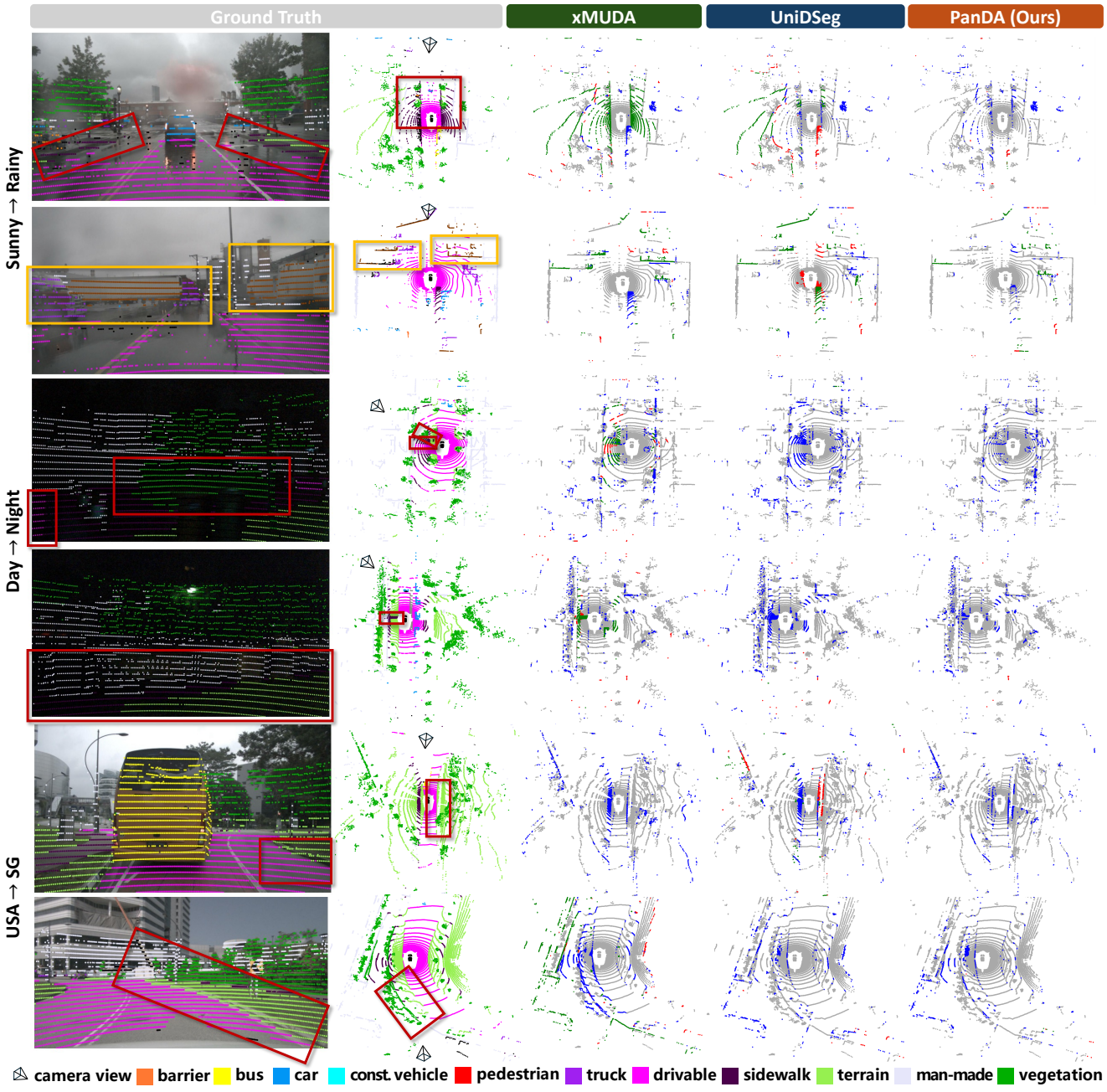

Figure D. Additional visualization of panoptic segmentation error maps across various domain shifts, comparing our method with xMUDA and UniDSeg. Point colors indicate error types: **false positives** and **false negatives** reflect recognition errors, while **well-matched** and **mismatched** points (within true positives) highlight segmentation inaccuracies. Raw point clouds, camera images, and semantic-colored ground truth (GT) are shown for reference. Best viewed in color. The boundary (with **scarlet** boxes) and instance segmentation (with **gold** boxes) results are highlighted.

F. Broader Impact and Limitations

F.1. Broader Impact

Our work presents the first framework for unsupervised domain adaptation in multimodal 3D panoptic segmentation. By addressing the challenges of cross-domain generalization and label scarcity, PanDA advances the robustness and adaptability of autonomous perception systems. This has the potential to reduce reliance on expensive manual annotations, enabling safer deployment of perception models

Table G. Additional training time and module size compared to the base model. "*" denotes the trainable parameters in UniDSeg. "#" denotes the size of auxiliary semantic branch. All measurements are conducted on the same device.

	Model Size (MB)	Lantency (ms)
Pano-xMUDA	+0.01	+105
Pano-UniDSeg	+1.82*	+760
AMD	+0.01#	+110
DualRefine	-	+108

in diverse real-world environments, such as under-explored cities, adverse weather conditions, and nighttime scenarios.

Moreover, our framework highlights the synergy between visual foundation models and geometric priors for reliable label estimation, promoting future research in cross-modal self-supervised learning. However, as with any perception system, responsible deployment requires consideration of failure cases and appropriate safeguards in safety-critical applications like autonomous driving.

F.2. Potential Limitations

While PanDA demonstrates strong cross-domain performance, several limitations remain. First, the method assumes temporal synchronization between LiDAR and camera inputs, which may not always hold in real-world systems. Second, though AMD improves robustness to missing or degraded modalities, performance may decline under extreme corruption (e.g., both LiDAR and cameras have significant corruptions). Third, our pseudo-label refinement relies on the quality of superpoints and VFM outputs, which may propagate errors under highly ambiguous scenes.

Future work will explore test-time adaptation strategies to further close the domain gap online, extend the framework to sequential or multi-frame settings, and investigate more efficient super-alternatives for large-scale deployment.

- Pytorch [1] . . . . . Pytorch License
- Grounding DINO [2] . . . . . Apache License 2.0
- Segment Anything Model (SAM) [3] . Apache License 2.0
- nuScenes Dataset [4] . . . . . CC BY-NC-SA 4.0 License[5]
- SemanticKITTI Dataset [6] . . . CC BY-NC-SA 4.0 License
- CLIP [7] . . . . . MIT License
- MMDetection3D [8] . . . . . Apache License 2.0
- IAL [9] . . . . . Apache License 2.0

[1] https://github.com/pytorch/pytorch
[2] https://github.com/IDEA-Research/GroundingDINO
[3] https://github.com/facebookresearch/segment-anything
[4] https://www.nuscenes.org/
[5] https://www.nuscenes.org/terms-of-use
[6] http://semantic-kitti.org/
[7] https://github.com/openai/CLIP
[8] https://github.com/open-mmlab/mmdetection3d
[9] https://github.com/IMPL-Lab/IAL

References

[1] Adriano Cardace, Pierluigi Zama Ramirez, Samuele Salti, and Luigi Di Stefano. Exploiting the complementarity of 2d and 3d networks to address domain-shift in 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 98–109, 2023. 1

[2] Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3d object detection to common corruptions in autonomous driving. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1022–1032. IEEE. 1

[3] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. 3

[4] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1

[5] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3D semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[6] Yining Pan, Qiongjie Cui, Xulei Yang, and Na Zhao. How do images align and complement lidar? towards a harmonized multi-modal 3d panoptic segmentation. In *Proceedings of the 42st International Conference on Machine Learning*, 2025. 1, 4

[7] Yao Wu, Mingwei Xing, Yachao Zhang, Xiaotong Luo, Yuan Xie, and Yanyun Qu. Unidseg: Unified cross-domain 3d semantic segmentation via visual foundation models prior. In *Proceedings of the 38th International Conference on Neural In- formation Processing Systems*, 2024. 1

[8] Zeqi Xiao, Wenwei Zhang, Tai Wang, Chen Change Loy, Dahua Lin, and Jiangmiao Pang. Position-Guided Point Cloud Panoptic Segmentation Transformer. *International Journal of Computer Vision*, 133(1):275–290, 2025. 1

[9] Linyan Yang, Lukas Hoyer, Mark Weber, Tobias Fischer, Dengxin Dai, Laura Leal-Taixé, Marc Pollefeys, Daniel Cremers, and Luc Van Gool. Micdrop: Masking image and depth features via complementary dropout for domain-adaptive semantic segmentation. In *European Conference on Computer Vision*, pages 329–346. 2025. 3

[10] Boxiang Zhang, Zunran Wang, Yonggen Ling, Yuanyuan Guan, Shenghao Zhang, and Wenhui Li. Mx2m: Masked cross-modality modeling in domain adaptation for 3d semantic segmentation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*, pages 3401–3409, 2023. 4

[11] Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-PolarNet: Proposal-free LiDAR Point Cloud Panoptic Seg-

mentation . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13189–13198, 2021. 1