
No Object Is an Island: Enhancing 3D Semantic Segmentation Generalization with Diffusion Models

Appendix and Supplementary Material

Fan Li¹ Xuan Wang¹ Xuanbin Wang¹ Zhaoxiang Zhang¹ Yuelei Xu^{1*}

¹Northwestern Polytechnical University
lifan.messages@gmail.com

A Appendix and Supplemental Material

The supplementary material is organized as follows: Section A.1 outlines the dataset split details and provides additional implementation information related to the stable diffusion model. Section A.2 presents an extended class-wise IoU comparison, highlighting the performance of our method against state-of-the-art baselines. Section A.3 provides further analysis on the ablation studies. In Section A.4, we compare our method with existing approaches in the source-free domain adaptation setting. Finally, qualitative visualization results of our method are shown in Section A.5.

A.1 Dataset Splits and Implementation Details

Dataset Splits. Following previous works [14, 2, 8], we partition the public datasets into training and testing subsets and perform class merging as detailed in Table 1. It is worth noting that we select 6 merged categories shared between VirtualKITTI and SemanticKITTI: Car, Truck, Road, Object, Building, and Vegetation. In the A2D2→SemanticKITTI setting, we expand the merged set to include ten categories: Car, Truck, Bike, Person, Road, Sidewalk, Building, Object, Parking, and Vegetation.

VirtualKITTI [5] is a synthetic dataset generated with the Unity engine by cloning five KITTI scenes (1, 2, 6, 18, 20) using real bounding box annotations. The dataset contains 2126 frames rendered under six weather and lighting conditions, all of which are used in training with random sampling.

SynLiDAR [16] is a large-scale synthetic LiDAR segmentation dataset generated with Unreal Engine [3], covering urban, suburban, residential, and harbor scenes. It contains 198396 scans with point-level annotations for 32 semantic classes. Following prior work [11, 16, 17, 19], we use 19840 point clouds for training and 1976 for validation.

SemanticKITTI [1] is a large-scale LiDAR dataset derived from the KITTI benchmark [6], providing front-view images and 64-beam LiDAR scans annotated with 19 semantic classes. Following the official split, we use sequences 00–07 and 09–10 for training, and sequence 08 for validation.

nuScenes [4] is a large-scale dataset collected in Boston and Singapore, containing 1,000 driving scenes and 40,000 annotated LiDAR frames with point-wise semantic labels from a 32-beam sensor. Each frame is paired with six RGB images. We follow the official train/val splits and adopt Day/Night and Boston/Singapore settings for domain generalization, using 6 merged semantic categories.

A2D2 [7] contains 28,637 frames from 20 driving sequences, collected using three front-facing 16-beam LiDARs. Point-wise labels for 38 classes are generated by projecting 2D semantic annotations onto the 3D point clouds. We follow prior work by using scene 20180807 145028 for testing and the rest for training.

*Corresponding author

Table 1: Details of dataset splits and merged classes for all cross-domain learning scenarios in this study.

Scenarios	Source		Target		Categories
	Train	Train	Val/Test	Val/Test	
nuScenes:Day/Night	24745	2779	606/602		Vehicle: [bicycle, bus, car, construction_vehicle, motorcycle, trailer, truck];
nuScenes:USA/Sing	15695	9665	2770/2929		Driveable Surface; Sidewalk; Terrain; Manmade; Vegetation
vKITTI/sKITTI	2126	18029	1101/4071		vKITTI: Car; Truck; Road; Object: [traffic sign, traffic light, pole, misc];
					Building; Vegetation: [terrain, tree, vegetation]
A2D2/sKITTI	27695	18029	1101/4071		sKITTI: Car; Truck; Road; Object: [fence, pole, traffic-sign, other-object];
					Building; Vegetation: [vegetation, trunk, terrain]
A2D2/sKITTI	27695	18029	1101/4071		A2D2: Car; Truck; Bike: [bicycle, small vehicle]; Person; Road; Parking;
					Sidewalk: [sidewalk, curbstone]; Object; Building; Vegetation
A2D2/sKITTI	27695	18029	1101/4071		sKITTI: Car; Truck; Bike: [bicycle, motorcycle, bicyclist, motorcyclist]; Person; Road;
					Parking; Sidewalk; Object; Building; Vegetation: [terrain, trunk, vegetation]

Table 2: Comparison of SOTA domain generalization methods on SemanticKITTI→SemanticSTF and SynLiDAR→SemanticSTF benchmarks. The best-performing results on each benchmark are highlighted in bold for clarity.

Method	car	bi.cle	mt.cle	truck	oth-v.	pers.	bi.clst	mt.clst	road	parki.	side.w.	oth-g.	build.	fence	veget.	trunk	terra.	pole	traf.
Oracle	89.4	42.1	0.0	59.9	61.2	69.6	39.0	0.0	82.2	21.5	58.2	45.6	86.1	63.6	80.2	52.0	77.6	50.1	61.7
SemanticKITTI→SemanticSTF																			
Source-only	55.9	0.0	0.2	1.9	10.9	10.3	6.0	0.0	61.2	10.9	32.0	0.0	67.9	41.6	49.8	27.9	40.8	29.6	17.5
Dropout [13]	62.1	0.0	15.5	3.0	11.5	5.4	2.0	0.0	58.4	12.8	26.7	1.1	72.1	43.6	52.9	34.2	43.5	28.4	15.5
Perturbation [17]	74.4	0.0	0.0	23.3	0.6	19.7	0.0	0.0	60.3	10.8	33.9	0.7	72.0	45.2	58.7	17.5	42.4	22.1	9.7
PolarMix [15]	57.8	1.8	3.8	16.7	3.7	26.5	0.0	2.0	65.7	2.9	32.5	0.3	71.0	48.7	53.8	20.5	45.4	25.9	15.8
MMD [9]	63.6	0.0	2.6	0.1	11.4	28.1	0.0	0.0	67.0	14.1	37.9	0.3	67.3	41.2	57.1	27.4	47.9	28.2	16.2
PCL [18]	65.9	0.0	0.0	17.7	0.4	8.4	0.0	0.0	59.6	12.0	35.0	1.6	74.0	47.5	60.7	15.8	48.9	26.1	27.5
PointDR [17]	67.3	0.0	4.5	19.6	9.0	18.8	2.7	0.0	62.6	12.9	38.1	0.6	73.3	43.8	56.4	32.2	45.7	28.7	27.4
UniMix [19]	82.7	6.6	8.6	4.5	15.1	35.5	15.5	37.7	55.8	10.2	36.2	1.3	72.8	40.1	49.1	33.4	34.9	23.5	33.5
XDiff3D (Ours)	83.3	7.9	9.5	24.1	16.7	35.9	14.9	37.9	67.8	14.9	39.6	2.5	73.1	49.3	50.7	35.8	45.6	30.7	36.1
SynLiDAR→SemanticSTF																			
Source-only	27.1	3.0	0.6	15.8	0.1	25.2	1.8	5.6	23.9	0.3	14.6	0.6	36.3	19.9	37.9	17.9	41.8	9.5	2.3
Dropout [13]	28.0	3.0	1.4	9.6	0.0	17.1	0.8	0.7	34.2	6.8	19.1	0.1	35.5	19.1	42.3	17.6	36.0	14.0	2.8
Perturbation [17]	27.1	2.3	2.3	16.0	0.1	23.7	1.2	4.0	27.0	3.6	16.2	0.8	29.2	16.7	35.3	22.7	38.3	17.9	5.1
PolarMix [15]	39.2	1.1	1.2	8.3	1.5	17.8	0.8	0.7	23.3	1.3	17.5	0.4	45.2	24.8	46.2	20.1	38.7	7.6	1.9
MMD [9]	25.5	2.3	2.1	13.2	0.7	22.1	1.4	7.5	30.8	0.4	17.6	0.2	30.9	19.7	37.6	19.3	43.5	9.9	2.6
PCL [18]	30.9	0.8	1.4	10.0	0.4	23.3	4.0	7.9	28.5	1.3	17.7	1.2	39.4	18.5	40.0	16.0	38.6	12.1	2.3
PointDR [17]	37.8	2.5	2.4	23.6	0.1	26.3	2.2	3.3	27.9	7.7	17.5	0.5	47.6	25.3	45.7	21.0	37.5	17.9	5.5
UniMix [19]	65.4	0.1	3.9	16.9	5.3	32.3	2.0	19.3	52.1	5.0	27.3	3.0	49.4	20.3	58.5	22.7	23.2	26.9	10.4
XDiff3D (Ours)	66.3	3.5	5.6	23.9	5.7	34.9	2.7	18.6	53.3	6.3	30.5	2.9	49.9	25.7	60.1	25.2	36.7	30.1	11.9

SemanticSTF [17] is an adverse-weather point cloud dataset with 2,076 scans and point-wise labels for 21 classes, including fog, rain, and snow conditions. For DG3SS, following the official split, 1,326 scans are used for training and 250 for validation.

Implementation Details. The stable diffusion model’s pre-trained weights can be obtained from the official repository at <https://huggingface.co/stabilityai/stable-diffusion-2>.

A.2 Extended Comparison of Class-wise IoU

Table 2 provides a comprehensive evaluation of state-of-the-art domain generalization methods on the SemanticKITTI→SemanticSTF and SynLiDAR→SemanticSTF benchmarks, including detailed class-wise IoU comparisons to assess generalization across diverse semantic categories. Notably, XDiff3D achieves top performance across most categories and demonstrates significant improvements over the baseline model UniMix in several particularly challenging classes. Under the SemanticKITTI→SemanticSTF setting, it boosts performance in truck (4.5% → 24.1%), fence (40.1% → 49.3%), and pole (23.5% → 30.7%). Similar trends are observed on the SynLiDAR→SemanticSTF benchmark, with notable gains in truck, sidewalk, and pole, highlighting the robustness of our method to adverse weather conditions and cross-domain visual variations. These consistent improvements—particularly in categories where previous methods struggle—highlight the strength of XDiff3D in modeling complex semantic dependencies among objects within a scene and effectively transferring them across domains. This underscores the effectiveness of leveraging diffusion-guided object agent queries to enable robust and generalizable 3D semantic segmentation in challenging real-world scenarios.

Table 3: Performance comparison of different loss functions \mathcal{L}_{sup} under the DG3SS setting on the A2D2 \rightarrow SemanticKITTI benchmark.

Function	Cross entropy	Huber	MSE
mIoU	48.6	49.1	47.7

Table 4: Ablation study of the loss weight γ for \mathcal{L}_{sup} under the DG3SS setting on the A2D2 \rightarrow SemanticKITTI benchmark.

γ	0.5	0.8	1	1.5	2
mIoU	48.6	48.9	49.1	48.7	48.1

Table 5: Performance comparison of source-free domain adaptation methods for 3D semantic segmentation on the nuScenes:Day \rightarrow Night benchmark. \dagger indicates results obtained using the official code. Top three results are highlighted as **best**, **second** and **third**, respectively. xM denotes the result which is obtained by taking the mean of the predicted 2D and 3D probabilities after softmax.

Task	Method	3D	xM
SFDA	xMUDA \dagger [10]	66.1	65.7
	SUMMIT [12]	68.9	68.2
	UniDseg [14]	70.7	68.7
	XDiff3D (Ours)	72.1	70.9

A.3 More Ablation Experiments

Loss Function \mathcal{L}_{sup} . We explore several loss formulations for \mathcal{L}_{sup} , including cross-entropy, Huber, and MSE, as summarized in Table 3. The results demonstrate that the type of loss function considerably affects performance, with Huber achieving superior results. This can be attributed to its robustness to outliers and its balanced sensitivity to prediction errors, making it particularly well-suited to our setting. In our framework, the intermediate supervision signal is derived from 2D object agent queries, which inherently introduce uncertainty. The Huber loss effectively addresses this uncertainty by providing a smoother optimization landscape, thus enabling more stable training and efficient convergence.

Loss Weight γ . As demonstrated in Table 4, our approach shows robustness across various choices of the weight coefficient γ . It consistently outperforms the baseline (UniDseg), regardless of the exact value of γ . In practice, setting $\gamma = 1$ provides a reliable initial choice when applying our method to new environments. This finding highlights the stability of our method across various weighting schemes and reinforces its reliability and effectiveness across diverse settings.

A.4 Source-Free Domain Adaptive 3D Semantic Segmentation

Source-Free Domain Adaptation (SFDA) aims to adapt a pre-trained source model to an unlabeled target domain without access to the original source data. This setting is especially relevant in scenarios where source data cannot be shared due to privacy, security, or storage constraints. The core objective is to bridge the domain gap using only the target-domain inputs and the pre-trained model, enabling effective target-domain performance while maintaining source-domain confidentiality. As shown in Table 5, our method significantly outperforms previous approaches on the nuScenes:Day \rightarrow Night benchmark. This result underscores the effectiveness of our framework in handling challenging domain shifts within the SFDA setting, highlighting its capability to robustly adapt to new domains even in the absence of source-domain data.

Limitation. While our method achieves outstanding performance in the SFDA setting, the lack of source-domain data makes it reliant on paired images and point clouds in the target domain. This requirement may constrain its applicability in real-world scenarios where obtaining complete multi-modal target data is challenging. Future work will explore utilizing depth maps derived from point clouds as surrogate image inputs, enabling diffusion models to provide useful semantic priors without explicit RGB images. This would improve the flexibility and applicability of our framework in more practical and unconstrained settings.

A.5 Qualitative Examples

In this section, we provide qualitative comparisons between our method and the previous state-of-the-art approach, UniDSEG, under the DG3SS setting on the VirtualKITTI→SemanticKITTI and A2D2→SemanticKITTI benchmarks. Our method demonstrates superior performance across both benchmarks, highlighting its capability to capture comprehensive spatial structures and effectively model contextual relationships among objects.

Specifically, our approach produces more complete and coherent spatial predictions for road (e.g., rows 1 and 3 in Figure 1), sidewalk (e.g., rows 7, 8, and 10 in Figure 2), and vegetation (e.g., rows 4, 7, and 8 in Figure 2). Additionally, it achieves more accurate category predictions for building (e.g., rows 1, 6, and 7 in Figure 1) and object (e.g., rows 1, 2, and 3 in Figure 2), effectively resolving ambiguities frequently encountered by UniDSEG. Moreover, our method demonstrates a robust understanding of scene semantics, significantly reducing unreasonable or trivial predictions commonly observed with UniDSEG (e.g., rows 3, 4, and 12 in Figure 1, and rows 4, 9, 10, and 11 in Figure 2). These qualitative results further validate the effectiveness of our diffusion-guided framework in capturing instance semantic dependencies and delivering reliable segmentation performance across diverse, unseen domains.

To better illustrate the effect of the feature refinement based on 3D agent queries, we conducted a clustering analysis on intermediate point cloud features before and after refinement. As shown in Figure 3, the refined features exhibit more comprehensive spatial coverage of scene elements, contextually consistent category predictions, and effective suppression of fragmented or noisy outputs. These results indicate that the model effectively captures semantic dependencies and learns the spatial and contextual arrangements within the scene, leading to more structured and coherent feature representations.

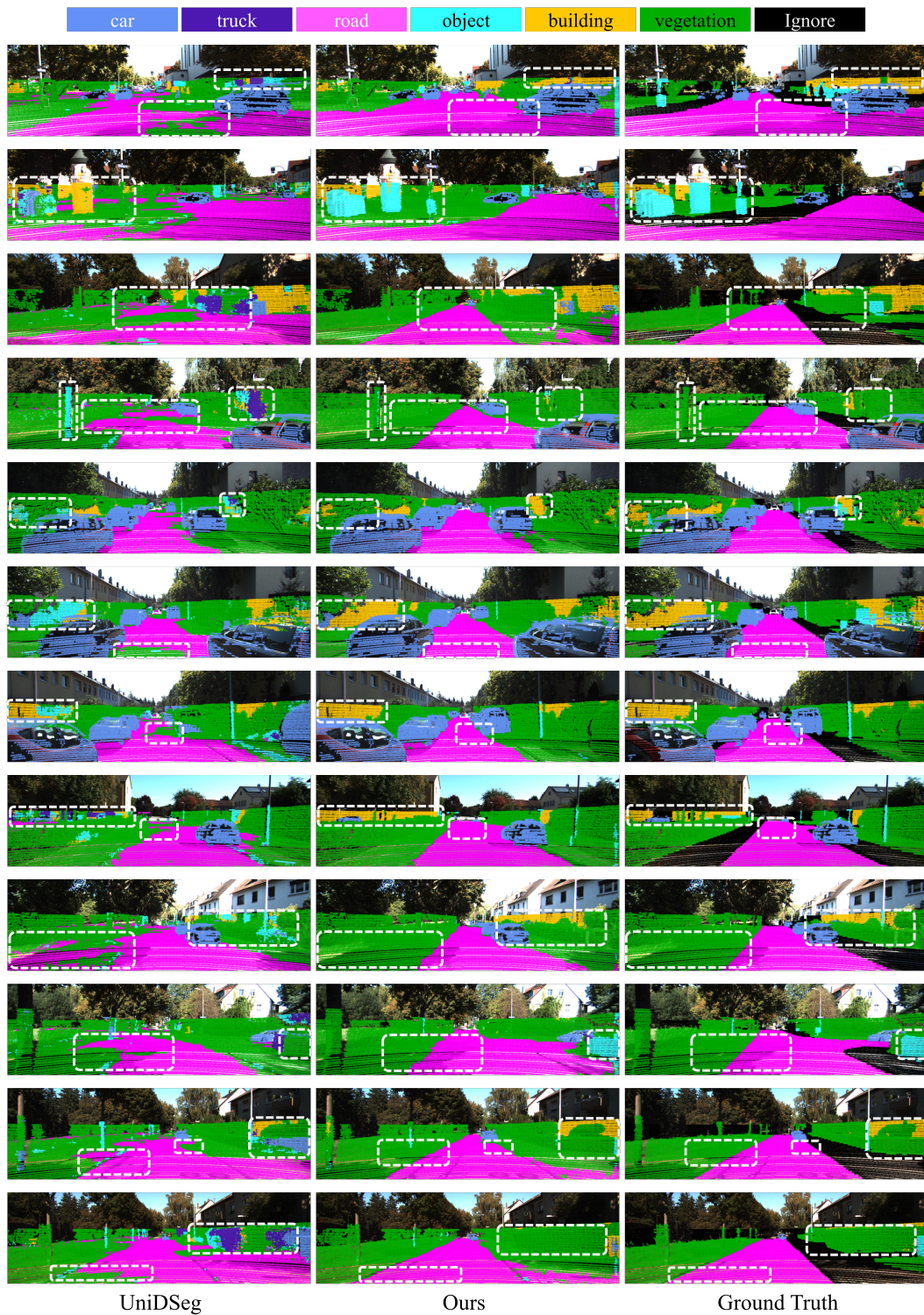




Figure 2: Qualitative results on A2D2→SemanticKITTI. From left to right: the visual results predicted by UniDSeg, Ours, and Ground Truth. We deploy the white dash boxes to highlight different prediction parts.

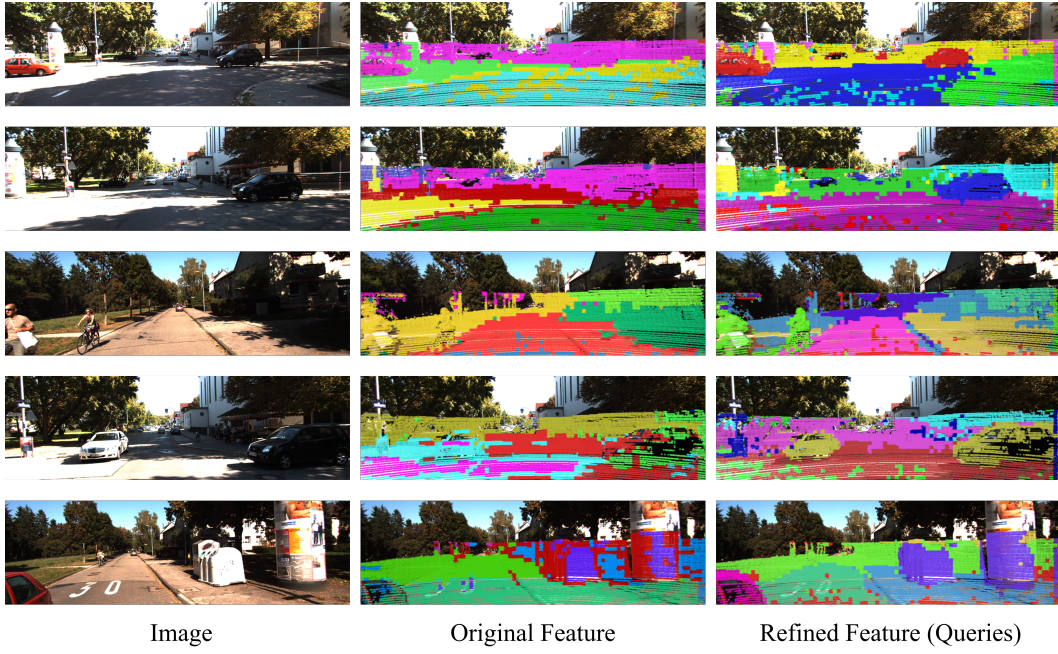


Figure 3: Clustering analysis of point cloud features on A2D2→SemanticKITTI. Visualization of feature distributions before and after query-based refinement. The “Original Feature” column shows the raw feature clustering results, while the “Refined Feature (Queries)” column presents the corresponding refined feature clusters obtained through the feature refinement process based on 3D agent queries.

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] 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.
- [3] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.
- [4] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nusenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022.
- [5] 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.
- [6] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.
- [7] 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.
- [8] 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.
- [9] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018.
- [10] Wei Liu, Zhiming Luo, Yuanzheng Cai, Ying Yu, Yang Ke, José Marcato Junior, Wesley Nunes Gonçalves, and Jonathan Li. Adversarial unsupervised domain adaptation for 3d semantic segmentation with multi-modal learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176:211–221, 2021.
- [11] Cristiano Saltori, Fabio Galasso, Giuseppe Fiameni, Nicu Sebe, Elisa Ricci, and Fabio Poiesi. Cosmix: Compositional semantic mix for domain adaptation in 3d lidar segmentation. In *European Conference on Computer Vision*, pages 586–602. Springer, 2022.
- [12] Cody Simons, Dripta S Raychaudhuri, Sk Miraj Ahmed, Suyu You, Konstantinos Karydis, and Amit K Roy-Chowdhury. Summit: Source-free adaptation of uni-modal models to multi-modal targets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1239–1249, 2023.
- [13] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.
- [14] 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. *Advances in Neural Information Processing Systems*, 37:101223–101249, 2024.
- [15] Aoran Xiao, Jiaxing Huang, Dayan Guan, Kaiwen Cui, Shijian Lu, and Ling Shao. Polarmix: A general data augmentation technique for lidar point clouds. *Advances in Neural Information Processing Systems*, 35:11035–11048, 2022.

- [16] Aoran Xiao, Jiaying Huang, Dayan Guan, Fangneng Zhan, and Shijian Lu. Transfer learning from synthetic to real lidar point cloud for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2795–2803, 2022.
- [17] Aoran Xiao, Jiaying Huang, Weihao Xuan, Ruijie Ren, Kangcheng Liu, Dayan Guan, Abdulmotaleb El Saddik, Shijian Lu, and Eric P Xing. 3d semantic segmentation in the wild: Learning generalized models for adverse-condition point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9382–9392, 2023.
- [18] Xufeng Yao, Yang Bai, Xinyun Zhang, Yuechen Zhang, Qi Sun, Ran Chen, Ruiyu Li, and Bei Yu. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7097–7107, 2022.
- [19] Haimei Zhao, Jing Zhang, Zhuo Chen, Shanshan Zhao, and Dacheng Tao. Unimix: Towards domain adaptive and generalizable lidar semantic segmentation in adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14781–14791, 2024.