MSA-UDA: Multi-scale Alignment Guided method in Cross-modal Unsupervised Domain Adaptation in 3D Semantic Segmentation

Jialong Zhang 36920240157322[1], Hao Wang 23020241154444[2], Jiawen Lin 23020241154419[2], Zihan Cheng 23020241154384[2], Yiping Wang 24520241154767[3]

[1]Class for Institute of AI
[2]Class for Institute of Information
[3]Class for Institute of Medical
{36920240157322,23020241154444,23020241154419,23020241154384,24520241154767}@stu.xmu.edu.cn

Abstract

Cross-modal Unsupervised Domain Adaptation (xMUDA) extract both 2D image semantic knowledge and 3D point cloud semantic information to eliminate the domain gap between source domain with target domain. It has achieved promising semantic segmentation in target point cloud by using model that train by source data. However, amount of existing methods either focus on improving the general performance of the model in both domain to indirectly enhances its performance in the target domain or use pseudo label to align output distribution which will introduce incorrect bias in feature representation. These does not directly solve the problem of domain differences. Therefore, this paper develop a Multi-scale Alignment Guided method (MSA-UDA) to focus on aligning multi-scale information in input-level to alleviate the domain gap between source and target domain. Specifically, MSA-UDA mainly contain two step: (a) Align the values in each RGB channel between source and target domain to reduce differences in color, lighting and other factors. (b) Align the intensity values to eliminate differences caused by external environmental factors such as lighting. Extensive experimental results demonstrate that our method achieve competitive results in several widely-recognized adaptation scenarios.

1 Introduction

3D segmentation plays a crucial role in many applications, such as autonomous driving, robotics, and medical imaging, where understanding spatial and structural information is essential. 3D semantic segmentation can densely assign specific semantic classes to each point. Like other computer-vision tasks, 3D semantic segmentation faces the domain shift issue, which results in performance degradation on a new unlabeled dataset (target-domain) with a different distribution from the labeled training dataset (source-domain). For instance, a 3D model learned on synthetic point clouds collected by the Unity game engine usually performs terribly on real point clouds collected by the LiDAR sensor. Annotating large-scale real datasets for every new scenario is a straightforward solution, but it leans on labor-intensive and time-consuming manual operations, especially for the tasks demanding point-wise annotations.

To overcome the issue above, recently, Unsupervised Domain Adaptation (UDA) technique has been proposed to alleviate the domain gap in 3D point cloud (Wu et al. 2019). UDA is mainly divided into uni-modal methods (Langer et al. 2020; Jiang and Saripalli 2021; Ding et al. 2022; Kong, Quader, and Liong 2023; Xiao et al. 2024; Yuan et al. 2024; Zhao et al. 2024) and cross-modal method (Jaritz et al. 2020; Peng et al. 2021; Li et al. 2022; Zhang et al. 2022; Wu et al. 2024). Cross-modal method consider the fusion of information from 2D images and 3D point clouds (Jaritz et al. 2020). It improves the confidence of the output distribution by aligning distribution between two modalities. To reduce the domain gap, some existing methods (Yi, Gong, and Funkhouser 2021; Wu et al. 2024) try to improve the segmentation capability of model to indirectly boost the performance in the target domain which do not actually bridge the domain gap.

Other methods (Yuan et al. 2022, 2023) focus on aligning the latent space distribution or output distribution between source and target domain. Although these methods can let the model output similar feature representation in both domain, they will also introduce extra bias. For example, due to the existence of domain differences, samples of different categories on two domains may have similar feature outputs, and further aligning this latent spatial distribution will result in the model learning incorrect parameter expressions.

Unlike the above methods, this paper propose MSA-UDA, which aims to align at the input level to approximate the attribute representation of the source data as closely as possible to the target data. Enable models trained on the source domain to be more suitable for target domain data. Specifically, We first align the pixel values of the RGB image channel by channel, so that the potential information such as color tone in the source domain image can be similar to that in the target domain image. A alignment result is show in Fig. 1. Secondly, We align the intensity of the 3D point cloud and project it onto the 2D image as an additional channel input into the network based on camera intrinsic parameters. By aligning the above two parameters (RGB and intensity), we can alleviate domain gap at the input level.

In a nutshell, our contributions can be summarized as follows:

(1). We align multiple attributes at the input level, im-

proving the similarity of feature representations between the source and target domains without reducing feature diversity, thereby reducing domain differences.

(2). Extensive experimental results demonstrate that our method achieve competitive results in several widely-recognized adaptation scenarios.

(a) source (b) target (c) target-like source

Figure 1: A example that align RGB image channel by channel from source to target

2 Related Work

2.1 3D semantice segmantation

The mainstream methods to process 3D semantic segmentation are divided into four types: Point-based method, Projection-based method, Voxel-based method, and Multi-representation method. In the point-based method, PointNet series (Qi et al. 2017), was a pioneering work that directly used multi-layer perceptrons (MLPs) to learn the features from the unordered sequences of point clouds. Later on, based on PointNet, convolution operations were implemented on the point-wise features output by MLPs (Wu, Qi, and Fuxin 2019) that performed well on the synthetic point cloud (Armeni et al. 2016). The projection-based method projected a point cloud to an image space in the bird's-eye-view (BEV) or range-view, achieving efficient 3D semantic segmentation with 2D CNNs, such as SqueezeSeg series (Wu et al. 2018) and others (Milioto et al. 2019). Recently, the voxel-based method (Graham, Engelcke, and Van Der Maaten 2018) (Peng et al. 2021) has been widely used for large-scale outdoor datasets, as it adopted sparse voxels to balance between efficiency and 183 effectiveness. Multi-representation method (Tang et al. 2020) used an ensemble of point-based or projection-based representation to potentially facilitate voxel representation. However, due to the sparsity of LiDAR sensors, these methods exhibit inferior performance in segmenting distant objects.

2.2 3D unsupervised domain adaptation

In recent years, 3D Unsupervised Domain Adaptation (3DUDA) for Lidar point clouds segmentation has gained traction, especially with methods focusing on point cloud segmentation. Existing work can be broadly divided into single-modality approaches (using Singleonly 3D point clouds) and multi-modality approaches (integrating 2D images and 3D point clouds).

Single-modality 3DUDA Early studies focused on transferring knowledge from one 3D domain (source) (e.g., synthetic to real-world LiDAR data) using techniques like adversarial training (Liu et al. 2021; Yuan et al. 2022, 2023; Li et al. 2023) or feature alignment (Langer et al. 2020) to another domain (target). For instance, certain methods (Liu et al. 2021) utilize domain discriminators to minimize the domain gap by making features from the source and target domains indistinguishable. Other approaches (Yuan et al. 2022, 2023) proposed the adversarial network based on category-level and prototype-level alignments to eliminate the side effect of global-level alignment and perform category-level alignment in a progressive manner. Furthermore, some method (Ding et al. 2022; Kong, Quader, and Liong 2023) constructed an intermediate domain to promote the cross-domain knowledge transfer.

Multi-modality 3DUDA Compared with Uni-modal methods, multi-modal methods exploit the exclusive information of 2d images, which can complement texture and other information with 3D point clouds. These approaches typically integrate 2D CNNs and 3D point cloud processing networks within the same framework. By aligning and fusing features across modalities, these methods enhance domain adaptation capabilities. xMuda (Jaritz et al. 2020) was the first to propose multi-modal segmentation methods in the field of 3DUDA. It aligns the output distribution between modalities in both the source and target domains, allowing the model to be used in both domains simultaneously, but it does not effectively eliminate the domain gap. Li et al. (Li et al. 2022) performed simple alignment at the input level based on xmuda and utilized distillation structures to extract target domain knowledge from the source domain, further reducing domain differences. Wu et al. (Wu et al. 2024) exploits prompt learning to transfer the generalization capability of the CLIP model to MM-UDA. It preserves the pre-existing target information from CLIP and learns vision-language-structure correlation. Peng et al. (Peng et al. 2025) leveraging the foundational model SAM to guide the alignment of features from diverse 3D data domains into a unified domain. Differently, we focus on align multi-scale input-level attribute to bridge the domain gap.

3 Method

3.1 Problem definition

Giving source domain $D_s = \{(X_i^{2D,S}, X_i^{3D,S}, Y_i^{3D,S})\}_{i=1}^{n_s}$ with n_s unlabeled 2D images and labels 3D point clouds, and a target domain $D_t = \{X_i^{2D,T}, X_i^{3D,T}\}_{i=1}^{n_t}$ with n_t unlabeled 2d images and 3D point clouds that the data distribution are inconsistent. Both domain share the same label space which contain C classes. The task is to learn a model that could predict the target labels of each points in point clouds $X_i^{3D,T}$.

3.2 Overview

The overall framework of MSA-UDA is describe in Fig. 2. Firstly, each of channel in RGB and intensity of source images are aligned to target images, improving the similarity between the source domain images and the target domain images. Then, we use the ResNet34 (He et al. 2016) as 2D backbond and SparseConvNet (Graham, Engelcke, and Van Der Maaten 2018) as 3D backbond to extract feature respectively. We call classifier head the last linear layer in the net-

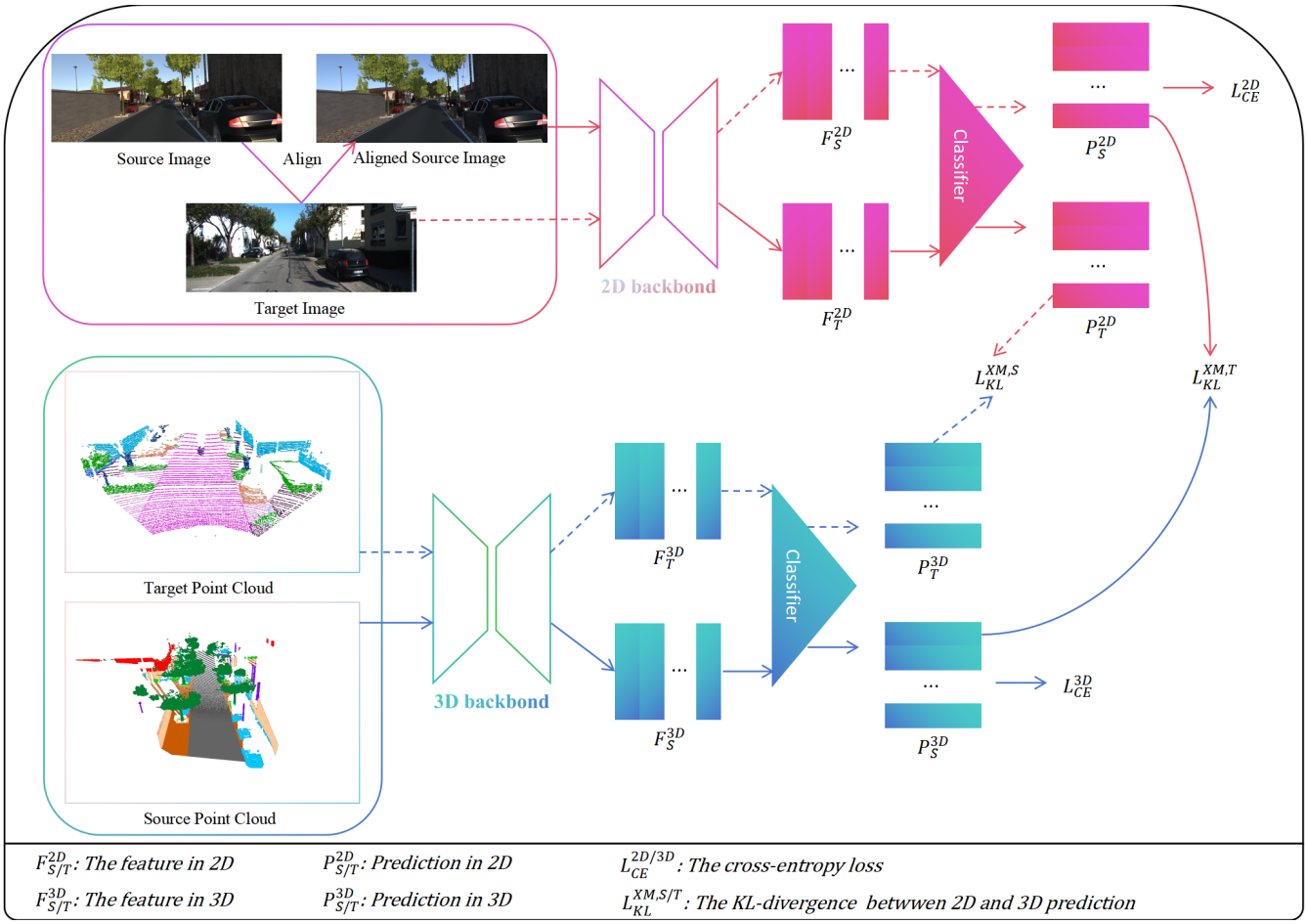

Figure 2: Overview framework of this paper. Note that in practice, a dual head classifier is used to align the output distribution (KL). For ease of understanding, only a single head classifier is shown here.

work that transforms the output features into segmentation logits followed by a softmax function to predict the class probabilities. Like xMuda (Jaritz et al. 2020), we use the dual head technique to construct a mimicry loss which use the KL divergence between logits from the first head of 2D and the second head of 3D, vice versa. Note that, before this step, the 3D feature will be project into 2D image using the camera intrinsic parameter and then sample the 2D feature that carry 3D feature. Finally, we use the ground truth namely semantic labels to supervised the semantic segmentation task, and cross-entropy loss is selected as the loss.

3.3 Multi-scale Alignment

Maintaining diversity in feature expression while aligning the distribution of source and target domains is an important prerequisite for eliminating domain differences without compromising model performance. Aligning distribution in input-level directly is an effective means. We have experimentally (§) demonstrated that aligning inputs at multiple scales can effectively reduce domain differences. And this operation does not force changes in feature expression, which can effectively maintain diversity in feature expression.

Specifically, we align the pixel values in each channel of RGB and intensity, making the input of source domain distribution similar with that of target domain. Assuming f_D^{ij} represent the j_{th} $(j \in (1, 2, ..., n))$ sample in i_{th} $(i \in (1, 2, 3, 4))$ channel from Domain $D \in (S, T)$, where i indicate the R,G,B,intensity channel respectively and they have the same sample size. We align each channel by using the follow formula:

$$S^i = \frac{\frac{1}{n}\sum_{j=1}^{n} f_T^{ij}}{\frac{1}{n}\sum_{j=1}^{n} f_S^{ij}} \tag{1}$$

$$f_{S2T}^i = f_S^i * S^i \tag{2}$$

where the S^i indicate the scale value in channel i, f_{S2T}^i represent the scaled pixel values in channel i.

3.4 Adding Depth and Intensity Channels

To further improve domain adaptation, we incorporate depth and intensity information from 3D point clouds into the 2D

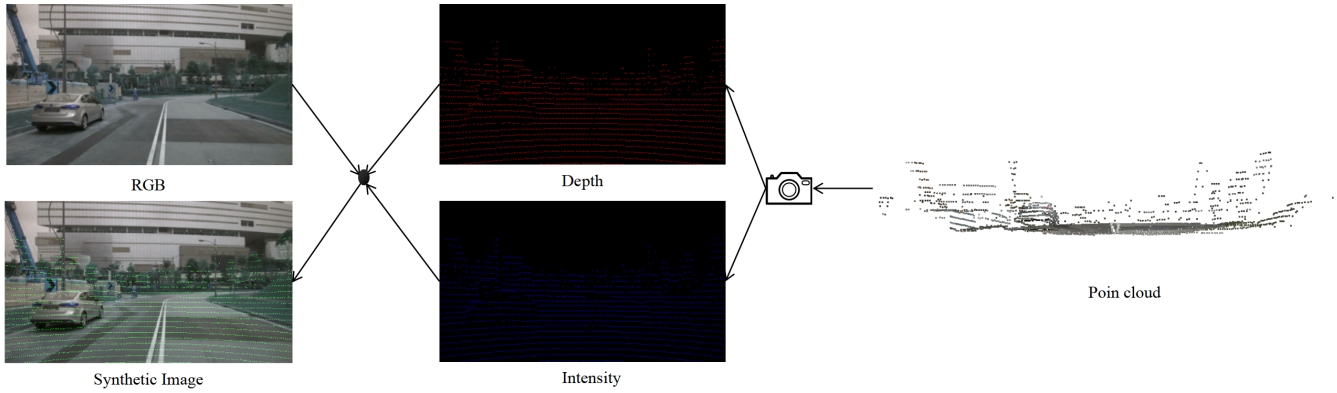

Figure 3: Depth and Intensity Channel Projection.

network. This is done by projecting depth and intensity values onto the 2D image plane using the camera's intrinsic parameters.

Depth Information The 3D point cloud's depth is computed as the distance from each point to the sensor. Using the camera's intrinsic matrix, we project this depth information onto the 2D image plane, generating a depth map.

Intensity Information Similarly, we project the intensity of each point in the 3D point cloud onto the 2D image, creating an intensity map.

By adding these depth and intensity maps as additional channels to the RGB input, the model gains richer spatial information, enhancing its ability to understand object structure. This helps the model better handle domain shifts by improving the alignment between source and target domain data.

Fig. 3 shows a schematic of this process, illustrating how depth and intensity information from the 3D point cloud are projected onto the 2D image. The depth and intensity channels are then concatenated with the RGB channels to form an enhanced multi-channel input.

3.5 Training Objective

Overall, the model is optimized with two objectives, *i.e.*, cross-entropy loss ($L_{CE}^{2D/3D}$) and KL-divergence loss ($L_{KL}^{XM,S/T}$):

$$\min_{\theta} L = L_{CE}^{2D} + L_{CE}^{3D} + L_{KL}^{XM,2D} + L_{KL}^{XM,3D} \quad (3)$$

where θ denotes the parameters of the model. For each modality, the source domain and target domain share the same model parameters.

4 Experiment

4.1 Datasets

In this paper, we use four public autonomous driving datasets, including three real scenarios: *nuScenes* (Caesar et al. 2020), *SemanticKITTI* (Behley et al. 2019), A2D2 (Geyer et al. 2020) and one synthetic scenario: VirtualKITTI (Gaidon et al. 2016). For all real datasets, LiDAR and RGB cameras are synchronized and calibrated, allowing 2D-to-3D projection, and for the synthetic dataset, VirtualKITTI provides depth maps so we simulate LiDAR scanning via uniform point sampling. Furthermore, following xMUDA (Jaritz et al. 2020), we only use the front camera image and the corresponding LiDAR points.

Our experimental scenarios cover typical real-to-real domain adaptation challenges like lighting changes (nuScenes: Day $\rightarrow$ Night), scene layout of country (nuScenes: USA $\rightarrow$ Singapore), and sensor setups (A2D2 $\rightarrow$ SemanticKITTI). For the first two scenarios, we choose 6 merged classes while for the last scenario, we select 10 shared classes between two datasets. In addition, the syntheticto-real domain adaptation challenge also be considered (VirtualKITTI $\rightarrow$SemanticKITTI, simulated depth, and RGB to real LiDAR and camera, with 6 merged classes). Details are provided in supplementary materials.

4.2 Implementation Details

For the 2D network, we use a modified version of U-Net with ResNet34 (He et al. 2016) encoder and a decoder with transposed convolutions and skip connections. For the 3D network, we use the offical SparseConvNet (Peng et al. 2021) implementation and a U-Net architecture with 6 times downsampling.

We employ standard 2D/3D data augmentation and log-smoothed class weights on point-wise supervised segmentation loss to address the class imbalance. The batch size is set to 8. Our model is trained on real-to-real adaptation for 100k iterations. We utilize an iteration-based learning schedule where the initial learning rate is 0.001 and then it is divided by 10 at 80k and 90k iterations. For synthetic-to-real, the training is performed for 30k iterations, and the learning rate is divided by 10 at the 25k and 28k iterations. All experiments are conducted on a single NVIDIA RTX 3090 GPU.

4.3 Quantitative Comparison

We present a comprehensive evaluation of our method, MSA-UDA, by comparing it with several state-of-the-

Table 1: Quantitative results (mIoU, %) on four settings

Method	Usa→Singapore			Day→night			vKITTI→sKITTI			A2D2→sKITTI		
	2D	3D	2D+3D	2D	3D	2D+3D	2D	3D	2D+3D	2D	3D	2D+3D
xMUDA(Jaritz et al. 2020)	64.4	63.2	69.4	55.5	69.2	67.4	42.1	46.7	48.2	38.3	46.0	44.0
AUDA(Liu et al. 2021)	64.0	64.0	69.2	55.6	69.8	64.8	35.8	37.8	41.3	43.0	43.6	46.8
DSCML(Peng et al. 2021)	65.6	56.2	66.1	50.9	49.3	53.2	38.4	38.4	45.5	39.6	45.1	44.5
DUAL-Cross(Li et al. 2022)	64.7	58.1	66.5	58.5	69.7	68.0	40.7	35.1	44.2	44.3	46.1	48.6
SSE-xMUDA(Zhang et al. 2022)	64.9	63.9	69.2	62.8	69.0	68.9	45.9	40.0	49.6	44.5	46.8	48.4
BFtD-xMUDA(Wu et al. 2023)	63.7	62.2	69.4	57.1	70.4	68.3	41.5	45.5	51.5	40.5	44.4	48.7
CLIP2UDA(Wu et al. 2024)	71.6	68.3	74.0	73.1	71.5	74.1	57.8	53.0	60.4	45.4	45.5	50.0
MSA-UDA(ours)	71.7	69.2	**74.5**	69.9	71.2	73.4	51.5	47.3	53.0	39.1	34.9	40.4

art multi-modal Unsupervised Domain Adaptation (UDA) methods, including xMUDA (Jaritz et al. 2020), AUDA (Liu et al. 2021), DSCML (Peng et al. 2021), DUAL-Cross (Li et al. 2022), SSE-xMUDA (Zhang et al. 2022), BFtD-xMUDA (Wu et al. 2023), and CLIP2UDA (Wu et al. 2024). The results of our experiments are presented in Table 1, where we report the mean Intersection over Union (mIoU, %) across four distinct domain adaptation scenarios.

From Table 1, we find that our method consistently outperforms the baseline methods across various settings. In the USA to Singapore adaptation scenario, MSA-UDA achieves a mIoU of 74.5% with the combination of both 2D and 3D modalities, outperforming xMUDA (69.4%) and CLIP2UDA (74.0%)—the latter being one of the closest competitors. Similarly, in the Day to Night scenario, MSA-UDA demonstrates a significant improvement in both 2D and 3D modality alignments, with a mIoU of 73.4%, surpassing xMUDA (67.4%) and other methods like SSE-xMUDA (68.9%).

The vKITTI to sKITTI adaptation scenario also shows the effectiveness of MSA-UDA. While our method delivers a mIoU of 53.0% in the 2D+3D setup, other approaches like DUAL-Cross and BFtD-xMUDA yield lower results (60.4% and 51.5%, respectively). This highlights the ability of MSA-UDA to handle domain shifts more effectively by aligning input-level features, including both RGB images and 3D point clouds, without losing feature diversity.

In the A2D2 to sKITTI scenario, which involves more complex sensor setup differences, our method achieves a mIoU of 40.4%. Although this is lower than in the previous scenarios, it still provides a clear advantage over methods such as CLIP2UDA (50.0%) and xMUDA (44.0%), demonstrating the robustness of MSA-UDA in dealing with challenging cross-domain tasks involving significant sensor variations.

4.4 Ablation Study

In accordance with the methodology introduced in Sec. 3.4, we conduct an ablation study to evaluate the impact of progressively adding depth, intensity, and RGB alignment information on the model's performance. As shown in Table 2, the baseline model achieves an mIoU of 69.4%. Adding depth information improves the performance to 72.23%, demonstrating the importance of depth for better spatial awareness and object relationship understanding. When both depth and intensity information are added, the mIoU increases further to 74.22%, highlighting the value of intensity data in providing additional surface details that complement depth information. Finally, incorporating RGB alignment results in a slight but noticeable improvement, bringing the mIoU to 74.46%. This final addition aligns pixel-level features across domains, reducing the domain gap and improving model performance.

Table 2: Ablation Study Results (mIoU, %)

Method	**mIoU (%)**
Baseline	69.4
+ Depth Information	72.2
+ Intensity Information	74.2
+ RGB Alignment	**74.5**

While the addition of depth and intensity provides substantial improvements, the marginal gain from RGB alignment suggests that it is particularly useful for addressing residual domain discrepancies. Overall, our ablation study confirms that the combination of depth, intensity, and RGB alignment provides the best performance, supporting the effectiveness of our multi-source input strategy in enhancing model robustness.

5 Conclusion

We propose MSA-UDA, Multi-scale Alignment Guided method in Cross-modal Unsupervised Domain Adatation, where four channel values in source domain align with that of target domain, eliminate the domain gap from input-level while keep the variation of feature representation. And the modalities learn from each other to improve performance on the target domain.

We reckon that our input-level alignment can bring inspiration in Unsupervised Domain Adaptation. And provide support for other related research and tasks.

References

Armeni, I.; Sener, O.; Zamir, A. R.; Jiang, H.; Brilakis, I.; Fischer, M.; and Savarese, S. 2016. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1534–1543.

Behley, J.; Garbade, M.; Milioto, A.; Quenzel, J.; Behnke, S.; Stachniss, C.; and Gall, J. 2019. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9297–9307.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

Ding, R.; Yang, J.; Jiang, L.; and Qi, X. 2022. Doda: Data-oriented sim-to-real domain adaptation for 3d semantic segmentation. In *European Conference on Computer Vision*, 284–303. Springer.

Gaidon, A.; Wang, Q.; Cabon, Y.; and Vig, E. 2016. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4340–4349.

Geyer, J.; Kassahun, Y.; Mahmudi, M.; Ricou, X.; Durgesh, R.; Chung, A. S.; Hauswald, L.; Pham, V. H.; Mühlegg, M.; Dorn, S.; et al. 2020. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*.

Graham, B.; Engelcke, M.; and Van Der Maaten, L. 2018. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9224–9232.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Jaritz, M.; Vu, T.-H.; Charette, R. d.; Wirbel, E.; and Pérez, P. 2020. xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12605–12614.

Jiang, P.; and Saripalli, S. 2021. Lidarnet: A boundary-aware domain adaptation model for point cloud semantic segmentation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2457–2464. IEEE.

Kong, L.; Quader, N.; and Liong, V. E. 2023. Conda: Unsupervised domain adaptation for lidar segmentation via regularized domain concatenation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 9338–9345. IEEE.

Langer, F.; Milioto, A.; Haag, A.; Behley, J.; and Stachniss, C. 2020. Domain transfer for semantic segmentation of LiDAR data using deep neural networks. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8263–8270. IEEE.

Li, G.; Kang, G.; Wang, X.; Wei, Y.; and Yang, Y. 2023. Adversarially masking synthetic to mimic real: Adaptive noise injection for point cloud segmentation adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20464–20474.

Li, M.; Zhang, Y.; Xie, Y.; Gao, Z.; Li, C.; Zhang, Z.; and Qu, Y. 2022. Cross-domain and cross-modal knowledge distillation in domain adaptation for 3d semantic segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3829–3837.

Liu, W.; Luo, Z.; Cai, Y.; Yu, Y.; Ke, Y.; Junior, J. M.; Gonçalves, W. N.; and Li, J. 2021. Adversarial unsupervised domain adaptation for 3D semantic segmentation with multi-modal learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 176: 211–221.

Milioto, A.; Vizzo, I.; Behley, J.; and Stachniss, C. 2019. Rangenet++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, 4213–4220. IEEE.

Peng, D.; Lei, Y.; Li, W.; Zhang, P.; and Guo, Y. 2021. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7108–7117.

Peng, X.; Chen, R.; Qiao, F.; Kong, L.; Liu, Y.; Sun, Y.; Wang, T.; Zhu, X.; and Ma, Y. 2025. Learning to adapt sam for segmenting cross-domain point clouds. In *European Conference on Computer Vision*, 54–71. Springer.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

Tang, H.; Liu, Z.; Zhao, S.; Lin, Y.; Lin, J.; Wang, H.; and Han, S. 2020. Searching efficient 3d architectures with sparse point-voxel convolution. In *European conference on computer vision*, 685–702. Springer.

Wu, B.; Wan, A.; Yue, X.; and Keutzer, K. 2018. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *2018 IEEE international conference on robotics and automation (ICRA)*, 1887–1893. IEEE.

Wu, B.; Zhou, X.; Zhao, S.; Yue, X.; and Keutzer, K. 2019. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *2019 international conference on robotics and automation (ICRA)*, 4376–4382. IEEE.

Wu, W.; Qi, Z.; and Fuxin, L. 2019. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 9621–9630.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; Fan, J.; Shi, Z.; and Qu, Y. 2023. Cross-modal Unsupervised Domain Adaptation for 3D Semantic Segmentation via Bidirectional Fusion-then-Distillation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 490–498.

Wu, Y.; Xing, M.; Zhang, Y.; Xie, Y.; and Qu, Y. 2024. Clip2uda: Making frozen clip reward unsupervised domain

adaptation in 3d semantic segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8662–8671.

Xiao, A.; Huang, J.; Liu, K.; Guan, D.; Zhang, X.; and Lu, S. 2024. Domain Adaptive LiDAR Point Cloud Segmentation via Density-Aware Self-Training. *IEEE Transactions on Intelligent Transportation Systems*.

Yi, L.; Gong, B.; and Funkhouser, T. 2021. Complete & label: A domain adaptation approach to semantic segmentation of lidar point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15363–15373.

Yuan, Z.; Cheng, M.; Zeng, W.; Su, Y.; Liu, W.; Yu, S.; and Wang, C. 2023. Prototype-guided multitask adversarial network for cross-domain LiDAR point clouds semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Yuan, Z.; Wen, C.; Cheng, M.; Su, Y.; Liu, W.; Yu, S.; and Wang, C. 2022. Category-level adversaries for outdoor lidar point clouds cross-domain semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 24(2): 1982–1993.

Yuan, Z.; Zeng, W.; Su, Y.; Liu, W.; Cheng, M.; Guo, Y.; and Wang, C. 2024. Density-guided Translator Boosts Synthetic-to-Real Unsupervised Domain Adaptive Segmentation of 3D Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23303–23312.

Zhang, Y.; Li, M.; Xie, Y.; Li, C.; Wang, C.; Zhang, Z.; and Qu, Y. 2022. Self-supervised exclusive learning for 3d segmentation with cross-modal unsupervised domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*, 3338–3346.

Zhao, H.; Zhang, J.; Chen, Z.; Zhao, S.; and Tao, D. 2024. 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*, 14781–14791.