Learning to Adapt SAM for Segmenting Cross-domain Point Clouds

Xidong Peng[1], Runnan Chen[2], Feng Qiao[3], Lingdong Kong[4], Youquan Liu[5], Yujing Sun[2], Tai Wang[6], Xinge Zhu[7⋆], and Yuexin Ma[1⋆]

[1]ShanghaiTech University, [2]The University of Hong Kong,
[3]RWTH Aachen University, [4]National University of Singapore,
[5]Hochschule Bremerhaven, [6]Shanghai AI Laboratory,
[7]The Chinese University of Hong Kong
{pengxd,mayuexin}@shanghaitech.edu.cn

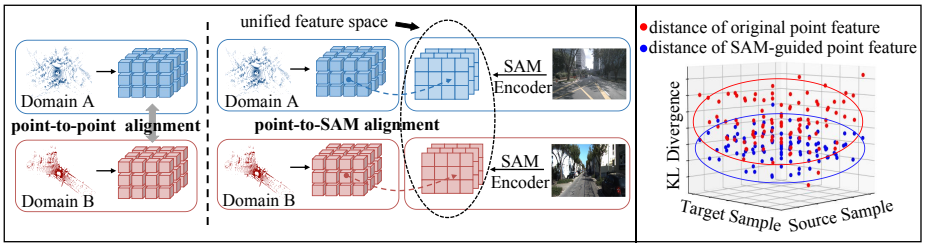
Abstract. Unsupervised domain adaptation (UDA) in 3D segmentation tasks presents a formidable challenge, primarily stemming from the sparse and unordered nature of point clouds. Especially for LiDAR point clouds, the domain discrepancy becomes obvious across varying capture scenes, fluctuating weather conditions, and the diverse array of LiDAR devices in use. Inspired by the remarkable generalization capabilities exhibited by the vision foundation model, SAM, in the realm of image segmentation, our approach leverages the wealth of general knowledge embedded within SAM to unify feature representations across diverse 3D domains and further solves the 3D domain adaptation problem. Specifically, we harness the corresponding images associated with point clouds to facilitate knowledge transfer and propose an innovative hybrid feature augmentation methodology, which enhances the alignment between the 3D feature space and SAM's feature space, operating at both the scene and instance levels. Our method is evaluated on many widely-recognized datasets and achieves state-of-the-art performance.

Keywords: Unsupervised Domain Adaptation · 3D Segmentation · Feature Alignment · Vision Foundation Model

1 Introduction

3D scene understanding is fundamental for many real-world applications, such as autonomous driving, robotics, smart cities, etc. Based on the point cloud, 3D segmentation is a critical task for scene understanding, which requires assigning semantic labels for each point. Current deep learning-based solutions [44,48] rely heavily on massive annotated data, which are high-cost and lack generalization capability for handling domain shifts. Unsupervised domain adaptation is

⋆ Corresponding author. This work was supported by NSFC (No.62206173), Natural Science Foundation of Shanghai (No.22dz1201900), Shanghai Sailing Program (No.22YF1428700), MoE Key Laboratory of Intelligent Perception and Human-Machine Collaboration (ShanghaiTech University), Shanghai Frontiers Science Center of Human-centered Artificial Intelligence (ShangHAI).



(a) comparison of feature alignment for 3D UDA (previous methods vs. ours) (b) cross-dataset feature distance

Fig. 1: (a) Comparison of 3D UDA paradigms. Different from aligning two point feature domains directly, our method makes both the source domain and target domain align with the SAM feature space. (b) Visualization of the feature distance across different datasets, where smaller values indicate a more similar distribution. It is obvious that after mapping to SAM feature space, point feature distributions from disparate domains become much more aligned.

significant for alleviating data dependency. However, unlike images with dense and regular representation, point clouds, especially LiDAR point clouds of large scenes, are unstructured and sparse, and have overt differences in patterns for various capture devices. Although some studies [38, 39, 46] have extended 2D techniques to solve the 3D UDA problem, the performance is still limited due to the essential defect of point cloud representation.

Considering that RGB cameras yield dense, color-rich, and structured data, and more importantly, they represent minor discrepancies across various devices, certain 3D UDA methods [4, 5, 19] utilize the synergy of LiDAR and camera capabilities to achieve more comprehensive and precise perception, and further enhance adaptation capabilities for 3D segmentation tasks. However, these methods usually train 2D and 3D networks simultaneously, demanding substantial online computing resources. Vision foundation models (VFMs), such as the Segment Anything Model (SAM) [23], have garnered significant attention due to their remarkable performance in addressing open-world vision tasks. Such models are trained on massive image data with tremendous parameters. Compared with a common model trained on limited data, VFMs have more general knowledge and much stronger generalization capability. Many works such as [7,8] have emerged recently to transfer the general 2D vision knowledge of VFMs to 3D and have achieved promising performance.

Based on SAM, focusing on image segmentation, we propose a novel paradigm for 3D UDA segmentation. As shown in Fig. 1(a), different from previous UDA approaches that strive to align the target domain to the source domain so that the model trained on labeled source data can also work on target data without annotation, our method makes both the source domain and target domain align with the SAM feature space. SAM feature space contains more general knowledge, which provides a friendly space to unify the feature representation from different domains. We utilize RGB images to assist point clouds in our framework. However, unlike the methods mentioned above only using images to provide auxiliary information, we take images as a bridge to align diverse 3D feature spaces to the SAM feature space, so we do not need to train extra 2D

networks and we can process the image offline for less computing resources. Moreover, considering that the 3D feature space created by the source-domain data and the target-domain data is still much smaller than the SAM feature space, we propose a hybrid feature augmentation method at both scene and instance levels to generate more 3D data with diverse feature patterns in a broader data domain, which can further benefit the 3D-to-SAM feature alignment. In particular, we make full use of the masks generated by SAM to mix instance-level point clouds with the other domains. This technique can maintain the geometric completeness of instances, which is beneficial for semantic recognition.

To verify that our idea of SAM-guided UDA is reliable, we randomly choose data from the source and target datasets and calculate the differences of feature distributions from source and target domain by KL divergence. The qualitative results are shown in Fig. 1(b), where the feature distribution differences after mapping to SAM feature space truly become much smaller. Moreover, to verify the effectiveness of our method, we compare it with current SOTA works on extensive 3D UDA segmentation settings and our method outperforms others by a large margin, improving about 14% mIoU for VirtualKITTI-to-SematicKITTI, about 15% mIoU for Waymo-to-nuScenes, and about 20% mIoU for nuScenes-to-SemanticKITTI domain adaptation. Surprisingly, our unsupervised method achieves comparable performance with the supervised method for city-changing and light-changing settings on the nuScenes dataset. We also test our method on more challenging tasks, such as panoptic segmentation and domain generalization, showing that our method is robust and has good generalization capability.

In summary, our contributions are as follows:

- We propose a novel unsupervised domain adaptation approach for 3D segmentation, leveraging the foundational model SAM to guide the alignment of features from diverse 3D data domains into a unified domain.
- We introduce a hybrid feature augmentation strategy at both scene and instance levels, generating more distinct feature patterns across a broader data domain for better feature alignment.
- We conduct extensive experiments on large-scale datasets and achieve SOTA performance.

2 Related Work

2.1 Point Cloud Semantic Segmentation

Point cloud semantic segmentation [17,48] is a rapidly evolving field, and numerous research works have contributed to advancements in this area. The pioneering approach PointNet [34] directly processes point clouds without voxelization and revolutionizes 3D segmentation by providing a novel perspective on point cloud analysis. Further, PointNet++ [35] extends PointNet with hierarchical feature learning through partitioning point clouds into local regions. To handle sparse point cloud data efficiently within large-scale scenes, a framework called SparseConvNet [16] has been specifically crafted. It excels in processing sparse

3D data and has been effectively utilized in various applications, including 3D semantic segmentation. MinkUNet [9] represents a significant advancement in point cloud semantic segmentation. Employing multi-scale interaction networks, MinkUNet enhances the segmentation of point clouds, effectively addressing the challenges posed by 3D spatial data. Our 3D segmentation networks are the popular SparseConvNet and MinkUNet. Due to the sparse characteristics of point cloud data, many current methods [18, 25, 45] add corresponding dense image information to facilitate point cloud segmentation tasks. Our method also uses image features to assist point cloud segmentation, and additionally, we utilize the 2D segmentation foundation model to achieve effective knowledge transfer.

2.2 Domain Adaptation for 3D Segmentation

Unsupervised Domain Adaptation (UDA) aims at transferring knowledge learned from a source annotated domain to a target unlabelled domain, and there are already several UDA methods proposed for 2D segmentation [6, 22, 47, 50]. In recent years, domain adaptation techniques have gained increasing traction in the context of 3D segmentation tasks. [46] leverage a "Complete and Label" strategy to enhance semantic segmentation of LiDAR point clouds by recovering underlying surfaces and facilitating the transfer of semantic labels across varying LiDAR sensor domains. CosMix [38] introduces a sample mixing approach for UDA in 3D segmentation, which stands as the pioneering UDA approach utilizing sample mixing to alleviate domain shift. It generates two new intermediate domains of composite point clouds through a novel mixing strategy applied at the input level, mitigating domain discrepancies. However, due to the sparsity and irregularity of the point cloud, the disparity across different point cloud data domains is larger compared to that across 2D image domains, which makes it difficult to mitigate the variation across domains.

With the development of multi-modal perception [1, 10] in autonomous driving, prevalent 3D datasets [2, 13, 15, 29] include both 3D point clouds and corresponding 2D images, making leveraging multi-modality for addressing domain shift challenges in point clouds convenient. xMUDA [19, 20] shows the power of combining 2D and 3D networks within a single framework, which achieves outstanding performance by aggregating the scores from these two branches. This achievement is attributed to the complementary nature resulting from the diverse modalities processed by each branch. [32] introduce Dynamic Sparse-to-Dense Cross-Modal Learning (DsCML) to enhance the interaction of multi-modality information, ultimately boosting domain adaptation sufficiency, while [5] elucidate this complementarity of image and point cloud through an intuitive explanation centered on the effective receptive field, and proposes to feed both modalities to both branches. However, in practice, training two networks with distinct architectures is difficult to converge and demands substantial computing resources due to increased memory. Our method uses the pre-trained foundation model to process the image data, guaranteeing the quality of the image features and enabling the training process to focus on the 3D model.

2.3 Vision Foundation Models

The rise of foundation models [12, 21, 40] has garnered significant attention which are trained on extensive datasets, consequently demonstrating exceptional performance. Foundation models [42, 49] have seen significant advancements in the realm of 2D vision, and several research studies extend these foundation models to comprehend 3D information. Representative works CLIP [36] leverage contrastive learning techniques to train both text and image encoders. CLIP2Scene [8] extends the capabilities of CLIP by incorporating a 2D-3D calibration matrix to facilitate a deeper comprehension of 3D scenes and OpenScene [33] focuses on zero-shot learning for 3D scenes through aligning point features in CLIP feature space to enable open vocabulary queries for 3D points. The Meta Research team recently launched the 'Segment Anything Model' [23], trained on an extensive dataset of over 1 billion masks from 11 million images. Utilizing efficient prompting, SAM can generate high-quality masks for image instance segmentation. The integration of flexible prompting and ambiguity awareness enables SAM with robust generalization capabilities for various downstream segmentation challenges. Many methods [7, 8, 28] take it as an off-the-shelf tool and distillate the knowledge to solve 3D problems by 2D-3D feature alignment. In our work of tackling the UDA of 3D segmentation, we utilize SAM to provide 2D prior knowledge for 3D feature alignment in a wider data domain.

3 Method

3.1 Problem Statement

We explore UDA for 3D segmentation, in which we have the source domain, denoted as $D_S = \{P_S, I_S, Y_S\}$ with paired input, namely point cloud P_S and image I_S, as well as annotated labels Y_S for each point, and the target domain denoted as $D_T = \{P_T, I_T\}$ without any annotation. Using these data, we train a 3D segmentation model that can generalize well to the target domain. 3D data from various domains have obvious differences in distribution and patterns, leading to over-fitting problems when models trained in one domain try to analyze data from another. The main solution is to align different features despite domain differences to achieve the generalization capability of the model. Different from previous works, our novel paradigm is to map data from distinct domains into a unified feature space, ensuring the model performs consistently across domains.

3.2 Framework Overview

The vision foundation model, SAM, is trained by massive image data, which contains general vision knowledge and provides a friendly feature space to unify diverse feature representations. Taking 2D images as the bridge, the 3D feature space of different domains can be indirectly unified by bringing them closer to the SAM feature space based on 2D-to-3D knowledge distillation. Based on this, we design a novel SAM-guided UDA method for 3D segmentation, as Fig. 2 shows.

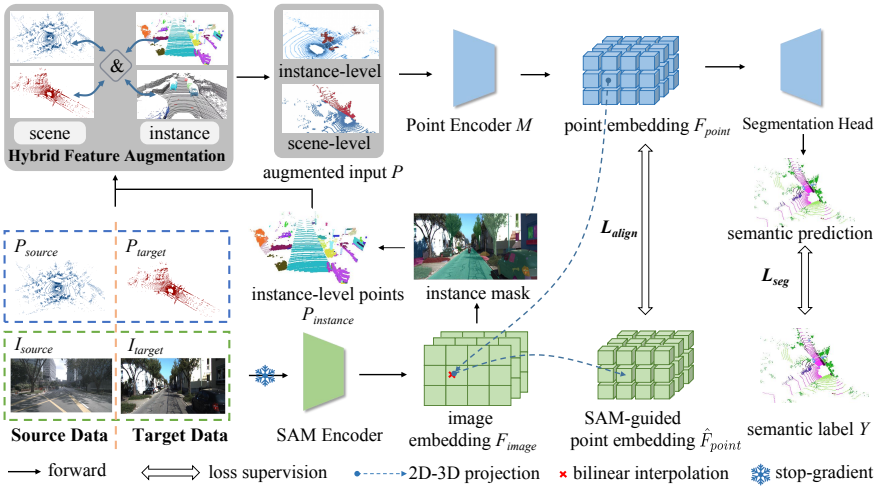


Fig. 2: Pipeline of our method. The point cloud is fed into the point encoder for point embeddings at the top, and the corresponding images are passed through the SAM encoder for image embeddings at the bottom, from which we obtain SAM-guided point embedding with the 2D-3D projection. Alignment loss L_{align} is calculated based on the SAM-guided features and original features. Furthermore, augmented inputs provide diverse feature patterns boosting the 3D-to-SAM feature alignment.

Specifically, given a point cloud input P, the point encoder M generates a point embedding $F_{point} \in \mathbb{R}^{n \times d}$ in the d-dimensional latent feature space. Concurrently, the corresponding image input I is passed through the SAM encoder for a c-channels image embedding $F_{image} \in \mathbb{R}^{h \times w \times c}$. Utilizing the correspondence between the point cloud and image, we acquire SAM-guided point embedding $\hat{F}_{point} \in \mathbb{R}^{n \times d}$ to compute the alignment loss L_{align} with the original point embedding F_{point}, serving the purpose of using SAM as a bridge to integrate the features of diverse data domains into a unified feature space. Notably, during training, the input for feature alignment consists of data from both source and target domains. We named this process as **SAM-guided Feature Alignment**. At the same time, as for labeled data Y, segmentation loss L_{seg} is also calculated as semantic supervision. During model training, only the point cloud branch of the whole pipeline is trained, and the gradient is not calculated in the image branch, which makes our method more lightweight. Furthermore, a **Scene-Instance Hybrid Feature Augmentation** is designed, which consists of scene-level and instance-level mix-up strategy. These mix-up strategies boost the variance of training data and generalize the network capability under the convex combination of the source domain and target domain data. Notably, the instance-level feature augmentation could maintain the local geometric relationship between two domains and make the subsequent alignment efficient.

3.3 SAM-guided 3D Feature Alignment

Previous UDA methods usually align the feature space of the target domain to that of the source domain so that the model trained on the source domain with

labeled data can also recognize the data from the novel domain. However, the distributions and patterns of 3D point clouds in various datasets have substantial differences, making the alignment very difficult. SAM [23], a 2D foundation model, is trained with a huge dataset of 11M images, granting it robust generalization capabilities to address downstream segmentation challenges effectively. If we can align features extracted from various data domains into the unified feature space represented by SAM, the model trained on the source domain can effectively handle the target data with the assistance of the universal vision knowledge existing in the SAM feature space.

We focus on training a point-based 3D segmentation model, while SAM is a foundation model trained on 2D images, which presents a fundamental challenge: how to bridge the semantic information captured in 2D images with the features extracted from 3D points. Most outdoor large-scale datasets with point clouds and images provide calibration information to project the 3D points into the corresponding images. With this information, we can easily translate the coordination of points P from the 3D LiDAR coordinate system P_{lidar} to the 2D image coordinate system P_{image}. This transformation can be formally expressed as Eq. 1, where rotation R_{ext} and translation T_{ext} represent the extrinsic parameters of the camera, and matrix K represents the intrinsic parameters of the camera.

$$zP_{image} = K(R_{ext}P_{lidar} + T_{ext}) \tag{1}$$

Once we calculate the projected 2D positions of points in the image coordinate system, we can determine their corresponding positions in the SAM-guided image embedding F_{image}, which is generated from the image by the SAM feature extractor. As the positions of points in the image embedding typically are not integer values, we perform bilinear interpolation based on the surrounding semantic features in the image embedding corresponding to the point, which to some extent alleviates the effect of calibration errors and allows us to derive the SAM-guided feature of each point, denoted as Eq. 2.

$$\hat{F}_{point} = \mathbf{Bilinear}(F_{image},\ P_{image}) \tag{2}$$

Then, the original point embedding F_{point} from both source and target domains are all required to align with their corresponding SAM-guided features $\hat{F}_{point}$. Specifically, we utilize the cosine function to measure the similarity of F_{point} and $\hat{F}_{point}$, employing it as the alignment loss L_{align} during training. With the supervision of L_{align}, features obtained by the point encoder M will gradually converge towards the feature space represented by SAM, achieving the purpose of extracting features within a unified feature space from the input of different domains. The formulation of the loss function for feature alignment is shown as Eq. 3.

$$L_{align} = 1 - \mathbf{cos}(\hat{F}_{point},\ F_{point}) \tag{3}$$

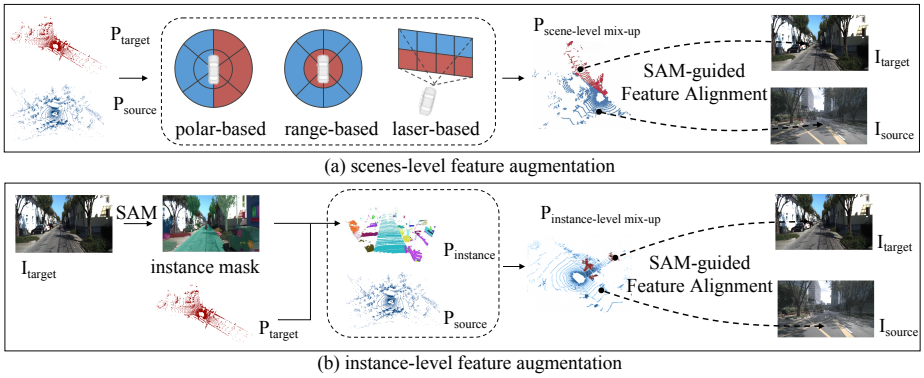


Fig. 3: Hybrid feature augmentation by data mixing for better 3D-to-SAM feature alignment. Part(a) illustrate all the scene-level approaches including polar-based, range-based, and laser-based point mix-up, where different color represents points from distinct domain. Part(b) shows the data flow of mixing the point data with instance-level data from another domain with an instance mask, where we take source data as an example for instance-level point generation and vice versa.

3.4 Scene-Instance Hybrid Feature Augmentation

3D point features of the source-domain data and the target-domain data only cover subsets of the 3D feature space, which are limited to align with the whole SAM feature space with more universal knowledge. Therefore, more 3D data with diverse feature patterns in a broader data domain is needed to achieve more effective 3D-to-SAM feature alignment.

Previous works [24, 43] usually focus on synthesizing data by combining data in the source domain and target domain at the scene level, including polar-based, range-based, and laser-based, as shown in Fig.3(a). Polar-based point mix-up selects semi-circular point cloud data from two different domains based on the polar coordinates of the point cloud. Range-based point mix-up divides the point cloud by its distance from the center, synthesizing circular point data close to the center and ring point data farther away from the center. Laser-based point mix-up determines the part of point clouds based on the number of laser beams, combining points with positive and negative laser pitch angles from different domains for synthesis. These ways of scene-level feature augmentation can maintain the general pattern of LiDAR point clouds as much as possible and improve the data diversity. Moreover, they are simple to process without any requirement for additional annotations such as real or pseudo-semantic labels. We adopt these three kinds of scene-level data augmentation in our method.

However, scene-level data augmentation will, to some extent, destroy the completeness of the point cloud of instances in the stitching areas and affect the exploitation of local geometric characteristics of point clouds. To further increase the data diversity and meanwhile keep the instance feature patterns of LiDAR point clouds for better semantic recognition, we propose an instance-level augmentation method. Benefiting from the instance mask output from SAM, we can thoroughly exploit the instance-level geometric features. Compared with

pre-trained 3D segmentation models, SAM provides more accurate and robust instance masks and enables us to avoid extra warm-up for a pre-train model, simplifying the whole training process. Therefore, we perform instance-level data synthesis as Fig.3(b) shows. Specifically, we begin by employing SAM to generate instance masks for input images from either the source or target domain (We take target data as the example in the figure). Next, we use the calibration matrix to project the corresponding point cloud into the image. The instance information of each point is determined according to whether the projection position of the point cloud falls within a specific instance mask, and then we randomly select some points with $20 \sim 30$ specific instances, mixed with the point cloud from the other domain by direct concatenation to achieve point augmentation at the instance level.

In practice, we combine all the ways of feature augmentation at both scene level and instance level with a random-selection strategy for a more comprehensive feature augmentation, which generates a more diverse set of point cloud data with varied feature patterns. Then, the augmented points are fed into the point encoder M to obtain the point embedding F_{point} with distinct feature patterns in a broader data domain beyond the source domain and target domain for more effective SAM-guided feature alignment. Notably, to maintain the consistency of the point cloud and the image, we extract SAM-guided point embedding based on the corresponding original image embedding.

4 Experiment

We first introduce datasets and implementation details. After that, we explore several domain shift scenarios and conduct comparisons for 3D segmentation. Then, we conduct extensive ablation studies to assess submodules of our method. Finally, we extend our method to more challenging tasks to show its generalization capability.

4.1 Dataset Setup

We first follow the benchmark introduced in xMUDA [20] to evaluate our method, comprehending four domain shift scenarios, including (1) USA-to-Singapore, (2) Day-to-Night, (3) VirtualKITTI-to-SemanticKITTI and (4) A2D2-to-SemanticKITTI. The first two leverage nuScenes [3] as their dataset, consisting of 1000 driving scenes in total with 40k annotated point-wise frames. Specifically, the former differs in the layout and infrastructure while the latter exhibits severe illumination changes between the source and the target domain. The third is more challenging since it is the adaptation from synthetic to real data, implemented by adapting from VirtualKITTI [14] to SemanticKITTI [2] while the fourth involves A2D2 [15] and SemanticKITTI as different data domains, where the domain discrepancy lies in the distinct density and arrangement of 3D point clouds captured by different devices since the A2D2 is captured by 16-beam LiDAR and the SemanticKITTI uses 64-beam LiDAR. For the above settings,

noted that only 3D points visible from the camera are used for training and testing, specifically, only one image and corresponding points for each sample are used for training.

Since we only use the image combined with SAM as offline assistance for the training of a 3D segmentation network instead of training a new 2D segmentation network, we focus on comparing the performance of the 3D segmentation network and enable model training with the whole point cloud sample because of less computational cost, even if some part of it is not visible in the images. For the part of the point cloud that cannot be covered by the image, alignment loss is not calculated, only segmentation loss is calculated. Thus, we also compare our method with others trained with the whole 360° view of the point cloud, in which three datasets are involved including nuScenes, SemanticKITTI, and Waymo [29]. In these settings, we use 6 images in nuScenes covering 360° view, 1 image in SemanticKITTI covering 120° view, and 5 images in Waymo covering 252° view. More information is introduced in the Appendix. For metric, We compute the Intersection over the Union per class and report the mean Intersection over the Union (mIoU).

4.2 Implementation Details

We make source and target labels compatible across these experiments. For all benchmarks in prior multi-modal UDA methods, we strictly follow class mapping like xMUDA for a fair comparison, while we map the labels of the dataset in other experiments into 10 segmentation classes in common. Our method is implemented by using the public PyTorch [31] repository MMDetection3D [11] and all the models are trained on a single 24GB GeForce RTX 3090 GPU. To compare fairly, we use SparseConvNet [16] with U-Net architecture as the 3D backbone network when following the benchmark introduced in xMUDA to evaluate our method and use MinkUNet32 [9] as the 3D backbone network when following the setup of taking the whole 360° point cloud as input, which is also the backbone of the state-of-the-art uni-modal method CosMix. For the image branch, the ViT-h variant SAM model is utilized to generate image embedding for SAM-guided feature alignment and instance masks for hybrid feature augmentation in an offline manner. We keep the proportion of mixed data and normal data from the source and target domain the same during model training. Before the data is fed into the 3D network, data augmentation such as vertical axis flipping, random scaling, and random 3D rotations are widely used like all the compared methods. For the model training strategies, we choose a batch size of 8 for both source data and target data, then mix the data batch for training at each iteration. Besides, we adopt AdamW as the model optimizer and One Cycle Policy as the learning-rate scheduler.

4.3 Experimental Results and Comparison

Tab. 1 and Tab. 2 show the experimental results and performance comparison with previous UDA methods for 3D segmentation under the setup intro-

Table 1: Results under four domain shift scenarios introduced by xMUDA. We report all the 3D network performance of compared multi-modal UDA methods in terms of mIoU. Note that the 3D backbone in these experiments is SparseConvNet [16].

Method	USA → Singapore		Day → Night		v.KITTI → Sem.KITTI		A2D2 → Sem.KITTI	
Source only	62.8	+0.0	68.8	+0.0	42.0	+0.0	35.9	+0.0
xMUDA [20]	63.2	+0.4	69.2	+0.4	46.7	+4.7	46.0	+10.1
DsCML [32]	52.3	−10.5	61.4	−7.4	32.8	−9.2	32.6	−3.3
MM2D3D [5]	66.8	+4.0	70.2	+1.4	50.3	+8.3	46.1	+10.2
Ours	**73.6**	**+10.8**	**70.5**	**+1.7**	**64.9**	**+22.9**	**52.1**	**+16.2**
Oracle	76.0	–	69.2	–	78.4	–	71.9	–

Table 2: Results under four domain shift scenarios with 360° point cloud, where not all the points are visible in the images. We report the 3D network performance in terms of mIoU. Note that the 3D backbone in these experiments is MinkUNet32 [9].

Method	nuScenes → Sem.KITTI		Sem.KITTI → nuScenes		nuScenes → Waymo		Waymo → nuScenes	
Source only	27.7	+0.0	28.1	+0.0	29.4	+0.0	21.8	+0.0
PL [30]	30.0	+2.3	29.0	+0.9	31.9	+2.5	22.3	+0.5
CosMix [37]	30.6	+2.9	29.7	+1.6	31.5	+2.1	30.0	+8.2
MM2D3D [5]	30.4	+2.7	31.9	+3.8	31.3	+1.9	33.5	+11.7
MM2D3D*	32.9	+5.2	33.7	+5.6	34.1	+4.7	37.5	+15.7
Ours	**48.5**	**+20.8**	**42.9**	**+14.8**	**44.9**	**+15.5**	**48.2**	**+26.4**
Oracle	70.3	–	78.3	–	79.9	–	78.3	–

duced in Sec. 4.1. Each experiment contains two reference methods in common, a baseline model named **Source only** trained only on the source domain and an upper-bound model named **Oracle** trained only on the target data with annotations. Tab. 1 focuses on four domain shift scenarios introduced by xMUDA [20] and comparison with these multi-modal methods based on xMUDA such as DsCML [32] and MM2D3D [5]. Among them, MM2D3D fully exploits the complementarity of image and point cloud and proposes to feed two modalities to both branches, achieving better performance. Our method outperforms it by +6.8% (USA → Singapore), +0.3% (Day → Night), +14.6% (v.KITTI → Sem.KITTI), +6.0% (A2D2 → Sem.KITTI) respectively, because our method aligns all the features into a unified feature space with the guidance of SAM instead of simply aligning features from image and point cloud in 2D and 3D network. Tab. 2 focuses on the scenarios where not all the point clouds are visible in the images and we re-implement three methods by their official codes. PL [30] uses the prediction from the pre-trained model as pseudo labels for unlabelled data to retrain this model, which is widely used in UDA methods. CosMix [37] trains a 3D network with only the utilization of a point cloud, which generates new intermediate domains through a mixing scene-level strategy to mitigate domain discrepancies. MM2D3d is the SOTA multi-modal method as described above, but it needs all the points visible in the image for the best performance. Our method surpasses them by at least +17.9% (nuScenes → Sem.KITTI), +11.0% (Sem.KITTI → nuScenes), +13.0% (nuScenes → Waymo), +14.7% (Waymo → nuScenes) respectively by a large margin, since hybrid feature augmentation can provide more intermediate domains and SAM-guided

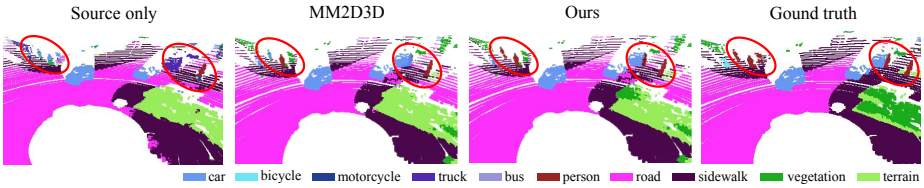


Fig. 4: Visualization of the domain adaptation from nuScenes to SemanticKITTI.

feature alignment can help map the whole point cloud into the unified feature space.

According to the results above, our method outperforms others by a large margin, attributed to the full utilization of the general knowledge provided by SAM. To further prove that the achievement is due to not only SAM but also our novel paradigm, we also use SAM for current SOTA multi-modal UDA work and get results in the row of "MM2D3D*". Because it trains both 2D and 3D networks simultaneously, under its original framework, we can only refine the supervision signals of 2D network using the instance mask output of SAM. As seen from the results, SAM can improve its performance but very limited. Our method not only uses instance masks but also makes full use of the general features extracted by SAM, ensuring the superiority of our method. Qualitative results are shown in Fig. 4, where predictions in the ellipses demonstrate that source-only and MM2D3D models often infer wrong and mingling results, especially for the person category, while our method can provide correct and more fine-grained segmentation. More qualitative results are in the Appendix.

Table 3: Ablation study. Baseline means the result of the source-only model indicating the lower-bound and Pseudo Label means re-training the model with pseudo labels.

Setting	Baseline	SAM-guided Feature Alignment	Hybrid Feature Augmentation		Pseudo Label	mIoU
			Scene-level	Instance-level		
(1)	✓					27.7
(2)	✓	✓				34.0
(3)	✓		✓	✓		28.6
(4)	✓	✓	✓			40.1
(5)	✓	✓		✓		39.0
(6)	✓	✓	✓	✓		44.0
(7)	✓	✓	✓	✓	✓	**48.5**

4.4 Ablation Study

To show the effectiveness of each module of our method, we conduct ablation studies on nuScenes-to-SemanticKITTI UDA. We also analyze and show the effect of other vision foundation models on our method.

Effectiveness of Model Components We first analyze the effects of all the submodules in our method in Tab. 3, containing SAM-guided Feature Alignment, Hybrid Feature Augmentation, and Pseudo Label. SAM-guided Feature Alignment aligns all the point features with the corresponding feature embeddings output by SAM, guiding the 3D network map point cloud into the unified feature space represented by SAM while Hybrid Feature Augmentation generates additional point cloud data of the intermediate domain for feature extraction to maximize the effect of feature alignment. Setting (1), (2), (3), and (6) in the table shows that combining the two submodules improves performance by a large margin. Besides, re-training the model with pseudo labels is a strategy widely used in UDA tasks and it also improves the performance.

Effectiveness of Hybrid Feature Augmentation For the detailed ablation of feature alignment, we adopt hybrid strategies for diverse data with distinct feature patterns, which not only mix up points at the scene level in polar-based, range-based, and laser-based ways but also at the instance level with the help of instance mask output by SAM. Random selection in all these point mix-up ways forms this feature augmentation. Setting (4), (5), (6) in Tab. 3 shows that both mix-up methods can help feature alignment with more distinct features but the hybrid strategy raises the best performance. More ablations are in the Appendix. Moreover, since masks generated by SAM do not contain semantic labels, we conduct an additional experiment to prove the validity of instance-level augmentation by replacing the generated SAM instance masks with the ground truth semantic masks under the setting (6). Compared with the original result (mIoU=44.0), the performance of using ground truth semantic masks is mIoU=42.7, demonstrating that although masks generated by SAM cannot be identical to the ground truth, the contained semantic information is consistent, and the randomness of our augmentation further improve the performance.

Effect of Different Point-to-Pixel Coverage for Alignment. For point clouds not covered by images, we do not calculate feature alignment loss and solely calculate segmentation loss with ground truth or pseudo label. When conducting the experiments, we used all available images to ensure that as many point cloud data as possible could find corresponding features on the pictures for SAM-guided feature alignment. We also conduct additional experiments to demonstrate the impact of the number of available images for feature alignment, as shown in Tab.4. The results indicate that a greater number of available images correlates with improved experimental outcomes. Importantly, as long as SAM-guided feature alignment is achievable, the performance does not significantly degrade even with limited coverage.

Table 4: The effect of image-point coverage. Different numbers represent the number of used pictures of nuScenes.

Covered Images for Alignment	baseline	2	4	6
nuScenes → Sem.KITTI	27.7	43.9	45.6	**48.5**
Sem.KITTI → nuScenes	28.1	38.4	40.6	**42.9**

Effect of Vision Foundation Model We also extend our method to other vision foundation models, such as InternImage [41], serving for image-based

tasks. Specifically, we replace the SAM-based image encoder with InternImage to guide the feature alignment in a similar manner. Compared with the baseline (mIoU=27.7), the performance of using InternImage is mIoU=36.9. A consistent performance gain can be obtained, which also verifies and validates our insight, *i.e.*, the generic feature space of the VFM can ease the feature alignment.

4.5 More Challenging Tasks

Since we achieve the purpose of mapping data from different domains into a unified feature space, the extracted feature can be used for some more challenging tasks. We show some extension results of our method in Tab. 5. The left subtable shows the results of UDA for panoptic segmentation, a more challenging task requiring instance-level predictions. With more accurate and fine-grained semantic prediction, our method achieves promising results. The right subtable shows the results of domain generalization, in which target data only can be used for testing. In this subtable, models are trained with nuScenes and SemanticKITTI and then evaluated with A2D2 dataset. With the ability of stronger data-to-feature mapping, our method outperforms the current SOTA method [26]. In the future, we seek to explore the potential of our method on more tasks, such as 3D detection.

Table 5: Extension on more challenging tasks, such as UDA for Panoptic Segmentation(left) and Domain Generalization(right), where N, S, A represent nuScenes, SemanticKITTI and A2D2 dataset.

Task	Method	PQ	PQ†	RQ	SQ	mIoU
nuScenes → Sem.KITTI	Source only	14.0	21.6	19.9	55.8	27.7
	PL	15.9	22.7	22.2	58.1	29.7
	Ours	**34.3**	**38.4**	**42.6**	**55.9**	**48.5**
	Oracle	50.5	52.2	57.8	77.2	70.3
Sem.KITTI → nuScenes	Source only	15.6	22.1	20.7	52.7	28.2
	PL	16.8	23.0	21.7	48.3	29.0
	Ours	**24.6**	**30.7**	**30.8**	**60.0**	**42.9**
	Oracle	40.7	44.9	47.2	83.8	78.3

Method	N,S → A
Baseline	45.0
xMUDA [20]	44.9
Dual-Cross [27]	41.3
BEV-DG [26]	55.1
Ours	**57.2**

5 Conclusion

In this paper, we acknowledge the limitations of existing UDA methods in handling the domain discrepancy present in 3D point cloud data and propose a novel paradigm to unify feature representations across diverse 3D domains by leveraging the powerful generalization capabilities of the vision foundation model, significantly enhancing the adaptability of 3D segmentation models. Hybrid feature augmentation strategy is also proposed for better 3D-SAM feature alignment. Extensive experiments show that our method surpasses all compared SOTA methods by a large margin.

References

1. Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., Tai, C.L.: Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1090–1099 (2022)
2. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9297–9307 (2019)
3. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11621–11631 (2020)
4. Cao, H., Xu, Y., Yang, J., Yin, P., Yuan, S., Xie, L.: Mopa: Multi-modal prior aided domain adaptation for 3d semantic segmentation. arXiv preprint arXiv:2309.11839 (2023)
5. Cardace, A., Ramirez, P.Z., Salti, S., Di Stefano, L.: 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. pp. 98–109 (2023)
6. Chang, W.L., Wang, H.P., Peng, W.H., Chiu, W.C.: All about structure: Adapting structural information across domains for boosting semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1909 (2019)
7. Chen, R., Liu, Y., Kong, L., Chen, N., Zhu, X., Ma, Y., Liu, T., Wang, W.: Towards label-free scene understanding by vision foundation models. In: Advances in Neural Information Processing Systems (2023)
8. Chen, R., Liu, Y., Kong, L., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y., Wang, W.: Clip2scene: Towards label-efficient 3d scene understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7020–7030 (2023)
9. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3075–3084 (2019)
10. Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., Yu, J., Ma, Y.: Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 461–469 (2023)
11. Contributors, M.: MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. `https://github.com/open-mmlab/mmdetection3d` (2020)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Fong, W.K., Mohan, R., Hurtado, J.V., Zhou, L., Caesar, H., Beijbom, O., Valada, A.: Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. IEEE Robotics and Automation Letters **7**(2), 3795–3802 (2022)
14. Gaidon, A., Wang, Q., Cabon, Y., Vig, E.: Virtual worlds as proxy for multi-object tracking analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4340–4349 (2016)

15. Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A.S., Hauswald, L., Pham, V.H., Mühlegg, M., Dorn, S., et al.: A2d2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320 (2020)
16. Graham, B., Engelcke, M., Van Der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9224–9232 (2018)
17. Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., Bennamoun, M.: Deep learning for 3d point clouds: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(12), 4338–4364 (2020)
18. He, D., Abid, F., Kim, J.H.: Multimodal fusion and data augmentation for 3d semantic segmentation. In: IEEE International Conference on Control, Automation and Systems. pp. 1143–1148 (2022)
19. Jaritz, M., Vu, T.H., Charette, R.d., Wirbel, E., Pérez, P.: xmuda: Cross-modal unsupervised domain adaptation for 3d semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12605–12614 (2020)
20. Jaritz, M., Vu, T.H., De Charette, R., Wirbel, É., Pérez, P.: Cross-modal learning for domain adaptation in 3d semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(2), 1533–1544 (2022)
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)
22. Kim, M., Byun, H.: Learning texture invariant representation for domain adaptation of semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12975–12984 (2020)
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
24. Kong, L., Ren, J., Pan, L., Liu, Z.: Lasermix for semi-supervised lidar semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21705–21715 (2023)
25. Krispel, G., Opitz, M., Waltner, G., Possegger, H., Bischof, H.: Fuseseg: Lidar point cloud segmentation fusing multi-modal data. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1874–1883 (2020)
26. Li, M., Zhang, Y., Ma, X., Qu, Y., Fu, Y.: Bev-dg: Cross-modal learning under bird's-eye view for domain generalization of 3d semantic segmentation. arXiv preprint arXiv:2308.06530 (2023)
27. Li, M., Zhang, Y., Xie, Y., Gao, Z., Li, C., Zhang, Z., Qu, Y.: Cross-domain and cross-modal knowledge distillation in domain adaptation for 3d semantic segmentation. In: Proceedings of the ACM International Conference on Multimedia. pp. 3829–3837 (2022)
28. Liu, Y., Kong, L., Cen, J., Chen, R., Zhang, W., Pan, L., Chen, K., Liu, Z.: Segment any point cloud sequences by distilling vision foundation models. arXiv preprint arXiv:2306.09347 (2023)
29. Mei, J., Zhu, A.Z., Yan, X., Yan, H., Qiao, S., Chen, L.C., Kretzschmar, H.: Waymo open dataset: Panoramic video panoptic segmentation. In: European Conference on Computer Vision. pp. 53–72. Springer (2022)
30. Morerio, P., Cavazza, J., Murino, V.: Minimal-entropy correlation alignment for unsupervised deep domain adaptation. arXiv preprint arXiv:1711.10288 (2017)

31. Paszke, A., Gross, S., Massa, e.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
32. Peng, D., Lei, Y., Li, W., Zhang, P., Guo, Y.: 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. pp. 7108–7117 (2021)
33. Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 815–824 (2023)
34. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017)
35. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in Neural Information Processing Systems **30** (2017)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
37. Saltori, C., Galasso, F., Fiameni, G., Sebe, N., Poiesi, F., Ricci, E.: Compositional semantic mix for domain adaptation in point cloud segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
38. Saltori, C., Galasso, F., Fiameni, G., Sebe, N., Ricci, E., Poiesi, F.: Cosmix: Compositional semantic mix for domain adaptation in 3d lidar segmentation. In: European Conference on Computer Vision. pp. 586–602 (2022)
39. Shaban, A., Lee, J., Jung, S., Meng, X., Boots, B.: Lidar-uda: Self-ensembling through time for unsupervised lidar domain adaptation (2023)
40. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
41. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., et al.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14408–14419 (2023)
42. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Segmenting everything in context. arXiv preprint arXiv:2304.03284 (2023)
43. Xiao, A., Huang, J., Guan, D., Cui, K., Lu, S., Shao, L.: Polarmix: A general data augmentation technique for lidar point clouds. Advances in Neural Information Processing Systems **35**, 11035–11048 (2022)
44. Xu, Y., Cong, P., Yao, Y., Chen, R., Hou, Y., Zhu, X., He, X., Yu, J., Ma, Y.: Human-centric scene understanding for 3d large-scale scenarios. arXiv preprint arXiv:2307.14392 (2023)
45. Yan, X., Gao, J., Zheng, C., Zheng, C., Zhang, R., Cui, S., Li, Z.: 2dpass: 2d priors assisted semantic segmentation on lidar point clouds. In: European Conference on Computer Vision. pp. 677–695. Springer (2022)
46. Yi, L., Gong, B., Funkhouser, T.: 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. pp. 15363–15373 (2021)

47. Zhang, Y., Wang, Z.: Joint adversarial learning for domain adaptation in semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 6877–6884 (2020)
48. Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D.: Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9939–9948 (2021)
49. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. arXiv preprint arXiv:2304.06718 (2023)
50. Zou, Y., Yu, Z., Kumar, B., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: European Conference on Computer Vision. pp. 289–305 (2018)