Exploring Modality Guidance to Enhance VFM-based Feature Fusion for UDA in 3D Semantic Segmentation

Johannes Spoecklberger[1] Wei Lin[2] Pedro Hermosilla[3]
Sivan Doveh[4] Horst Possegger[1] M. Jehanzeb Mirza[5]
[1]Institute of Visual Computing, Graz University of Technology
[2]JKU LINZ [3]TU Wien [4]IBM Research [5]MIT CSAIL
j.spoecklberger@tugraz.at

Abstract

*Vision Foundation Models (VFMs) have become a de facto choice for many downstream vision tasks, like image classification, image segmentation, and object localization. However, they can also provide significant utility for downstream 3D tasks that can leverage the cross-modal information (*e.g.*, from paired image data). In our work, we further explore the utility of VFMs for adapting from a labeled source to unlabeled target data for the task of LiDAR-based 3D semantic segmentation. Our method consumes paired 2D-3D (image and point cloud) data and relies on the robust (cross-domain) features from a VFM to train a 3D backbone on a mix of labeled source and unlabeled target data. At the heart of our method lies a fusion network that is guided by both the image and point cloud streams, with their relative contributions adjusted based on the target domain. We extensively compare our proposed methodology with different state-of-the-art methods in several settings and achieve strong performance gains. For example, achieving an average improvement of 6.5 mIoU (over all tasks), when compared with the previous state-of-the-art.*

1. Introduction

In an ideal world with abundant resources, we could potentially manually label all the distributions present in our visual world and train neural networks to perform robustly across diverse environments. The success of supervised training is well-known and has been extensively studied in the literature [16, 29, 50]. However, in the real world, the concept of *abundance* is usually non-existent. Instead, we face several constraints – monetary, time, or human resources, among others – that demand we employ all the available resources optimally.

Efficiency and cost reduction become even more critical in domains like autonomous driving, where vehicles ideally need to operate safely in a range of environments (*i.e.,* data distributions). Labeling data for each of these distributions, particularly for dense prediction tasks like 3D semantic segmentation, can quickly become prohibitively expensive due to the need for per-point or per-voxel annotations.

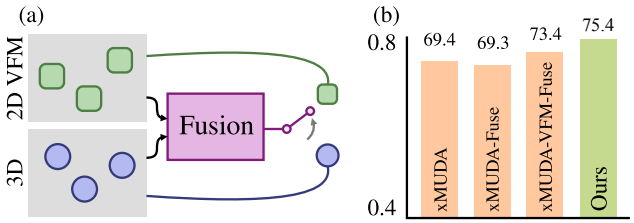

Figure 1. **(a)** Cross modal learning with frozen 2D (VFM) backbone features using a learned fusion representation. Fusion networks can lead to a suboptimal feature utilization and unwanted modality bias on the target domain. Therefore, we propose regularizing the fusion by the most effective modality in a certain environment (*e.g.,* based on lighting conditions). **(b)** mIoU Comparison of xMUDA with different fusion variants and Ours on NuScenes: USA → Singapore.

3D Unsupervised Domain Adaptation (UDA) offers a practicable solution by focusing on adapting a neural network trained in a supervised manner on a labeled source domain to an unlabeled target domain. This approach leverages information from annotated 3D (source) data to address the variability across domains without requiring manually annotated labels from each new target domain. This paper builds on the principles of 3D domain adaptation, aiming to bridge the gap between labeled and unlabeled 3D environments, particularly addressing challenges inherent in dense 3D prediction tasks, related to fine-grained understanding of the 3D scene.

Ideally, an autonomous vehicle can employ cues from both the 2D and 3D data. Although different modalities, they can offer complimentary cues that can enhance the perception abilities of autonomous driving systems. Jaritz *et*

al. [19] proposed a seminal work on using the complimentary information from the two modalities for the task of 3D domain adaptation for semantic segmentation. This method inspired a set of subsequent works [7, 43, 45], which combine data from different domains on a similar multi-modal design. However, these approaches usually rely on training a 2D feature extractor separately, thus, requiring dis-joint training of 2D and 3D backbones, increasing the computation cost significantly as compared to uni-modal setups.

In our work, motivated by the recent strong improvements offered by vision foundation models (VFM) on dense prediction tasks [34, 39], we tackle the multi-modal 3D-UDA task by employing these powerful models. However, leveraging these VFMs (*e.g.,* [39]) effectively for cross-modal learning for 3D UDA comes with the problem of optimally fusing the information from both modalities. To address this, at the heart of our approach lies a fusion refinement network, which fuses the information obtained from 2D and 3D modalities into a combined representation.

Although such fusion schemes are studied in [18, 24], our experiments show that training these fusion methods can lead to an over-reliance on features that perform well on the source domain but fail to adapt to the target domain. Moreover, despite the growing popularity of multi-modal feature fusion, few works have focused on designing fusion modules that specifically address modality-specific challenges that for example arise under degrading weather and low light conditions [4].

To address this issue, we propose to *adaptively* regularize the fusion network based on a simple intuition: in some environments, the imaging modality may be more robust while in others, reliance on the 3D modality may be more beneficial. To this end, we guide the fusion refinement network through adaptive predictive distillation from each modality, based on environmental factors (e.g., lighting conditions). This regularization effectively promotes learning to select the modality to rely on, thereby enabling more robust multi-modal domain adaptation. An overview of our approach is also laid out in Figure 1. We evaluated our proposal on the four common DA tasks, showing improved performance over traditional (fusion) methods, achieving SOTA or comparable performance on all tasks without requiring the training of an additional 2D encoder.

2. Related Work

Our method is related to works that study UDA for 3D semantic segmentation, VFMs, and approaches that propose different fusion schemes for cross-modal learning.

UDA for 3D Semantic Segmentation. UDA for 3D semantic segmentation has been extensively studied in recent years.

A major group of methods can be categorized as learning domain-invariant feature representations. Earlier works in 3D UDA focused on minimizing statistical feature discrepancies [33, 42]. Next, adversarial training is another widely used approach to learn invariant representations [2, 13, 49]. Another line of research utilizes self-supervision to learn more domain invariant representations by constructing a label-free auxiliary optimization goal that is often modality specific [31, 48]. Domain mapping approaches can be seen as another high level category to handle domain shifts, in which the target domain is transferred to the source domain [10, 17]. Other approaches include model adaptation via adapting batch normalization statistics [25, 31, 32], pseudo-labeling [23] or self-ensembling [22].

In multimodal DA, techniques are characterized by performing adaptation across modalities. xMUDA [18] is a seminal work that introduces a dual classification head structure for predictive feature alignment across the 2D and 3D modalities and serves as the fundament for many works on 3D UDA cross-modal semantic segmentation. MM2D3D [7] extends the work by adding 3D depth to the 2D encoder, showing significant performance improvement. DsCML [35] proposes sparse-to-dense feature matching, aligning 3D point features with a dynamically selected 2D region. Further, they propose cross-modal adversarial learning to narrow the domain gap, which was also employed by Liu *et al.* [27]. Xing *et al.* [46] propose a neighborhood feature aggregation network and the usage of contrastive learning for feature alignment. Zhang *et al.* [51] propose modality-exclusive self-supervised learning tasks for the 2D and 3D branches to improve the exploitation of modality-specific characteristics. In this work, we propose a method to further enhance the 2D-3D DA capabilities enabled by VFMs. Distinct from others, we design a three-stream network architecture composed of 2D, fusion, and 3D branches, where the single-modality streams guide the fusion branch adaptively.

Vision Foundation Models (VFMs). Vision Foundation Models, such as CLIP [38], DINO [8], and SAM [21], represent general-purpose frameworks trained on large-scale visual data, capable of performing a wide range of tasks, including image classification, object detection, segmentation and even cross-modal applications. Prior works such as Vision Transformer (ViT) [11] laid the groundwork by demonstrating how transformer-based architectures could rival CNNs in vision tasks when trained on large datasets. Models like CLIP [38] and ALIGN [20] exploited multi-modal capabilities showing impressive performance by learning from paired text-image data. DINO [8] and BEiT [1] refined the application of self-supervised learning in vision, achieving high accuracy without the need for labeled data. AM-RADIO [39] distills a unified efficient backbone from multiple foundation models (in a teacher-student learning framework) which outperforms its teachers across several downstream tasks.

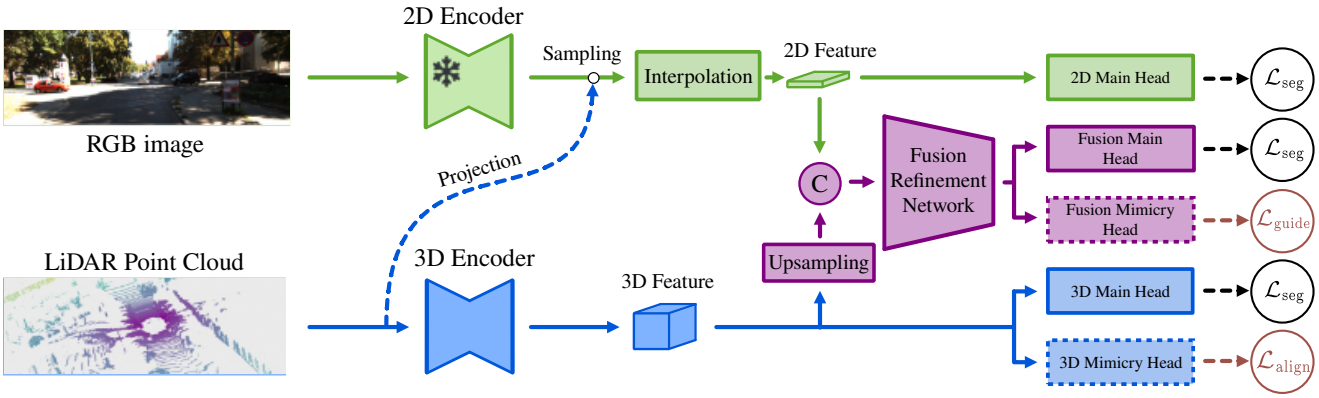

Figure 2. Our architecture for the cross-modal learning consists of a Vision Foundation Model (VFM) as the 2D encoder and a 3D SparseConvNet as the 3D encoder. We use multiple *main task heads* of semantic segmentation and *mimicry task heads* for cross-modal alignment. Besides the 2D branch (green) and 3D branch (blue), we employ a fusion branch (purple) where we concatenate the 2D and 3D feature as the fusion feature. The network is trained with the supervised loss $\mathcal{L}_{\text{seg}}$, and the self-supervised cross-modal learning losses $\mathcal{L}_{\text{align}}$ (the fusion branch guiding the 3D branch) and $\mathcal{L}_{\text{guide}}$ (the fusion branch guided by the 2D or 3D branch).

Recently, works on DA leveraging VFMs have demonstrated strong performance for the task of 3D UDA [6, 36, 47]. Approaches to exploit VFMs can be roughly divided into works that exploit feature distillation, predictive distillation, and mask priors. Peng *et al.* [36] apply feature distillation via cosine similarity. In addition, they utilize the masking capabilities of SAM for instance-level augmentation. Xu *et al.* [47] employ the semantic-aware segmentation model SEEM [52] for pseudo label refinement and mix source and target point clouds by selecting points guided by the SAM masks. Cao *et al.* [6] exploit SAM mask priors for rare objects that are inserted in the scene tackling the long-tailed class distribution. Our method follows the predictive alignment paradigm, where we demonstrate that advancements in VFMs can have a direct correlation with the improvements obtained for the task of 3D UDA.

Feature Fusion in 3D DA. xMUDA [18] proposed a variant with late fusion of 2D and 3D features which outperforms all the unimodal baselines. A recent line of works [43, 45] also divert their attention toward distillation based on fused feature representations. Similar to our work, Wu *et al.* [43] utilizes a fusion module with a separate classifier to transfer knowledge via predictive distillation to the 2D and 3D modality. FtD++ [45] relies on a two-stream pipeline where the 2D stream is fused with 3D information, which is then used for cross-modal learning with the 3D stream. The recent UniDSeg [44] enhances the frozen VFM with learnable blocks between encoder layers allowing fine-tuning and the integration of range image information.

In our approach, we treat the fusion as a distinct network trained with supervision, offering us more degrees of freedom and better adaptability to the task of interest. As opposed to the xMUDA fusion variant, we explicitly optimize the 3D and 2D stream via classifier heads and utilize these for guiding the distribution of the fusion network. Similarly, in relation to [43], our method explicitly regularizes the fusion representation using a guiding signal to bias the fusion toward a specific modality. This regularization is a key component of our approach and is designed to enhance performance under domain shift when a modality preference is available, while still improving robustness in its absence by promoting prediction consistency on the target domain.

The recent work in [4] also identifies the need for non-uniform sensor fusion. In their work the fusion module is used as a central component to overcome sensing difficulties in a shared feature space. Their setup however requires scene attribute descriptions to learn a condition token which guides a windowed cross attention fusion over all modalities in the image space.

3. Method

We introduce our cross-modal DA approach and start with a pipeline overview in Section 3.1. Then, we present our cross-modal fusion schema in Section 3.2, provide details on cross-modal learning in Section 3.3, and conclude with details on the training objectives in Section 3.4.

3.1. Overview

2D-3D Domains. In the 2D-3D multimodal UDA setting, we have labeled source data and unlabeled target data from the 2D and 3D modalities. The task is to optimize the final segmentation prediction on the target data. We denote the source domain as $S = \{x_i^{2D}, x_i^{3D}, y_i | i \in I_S\}$ and the target domain as $T = \{x_i^{2D}, x_i^{3D} | i \in I_T\}$, where x_i^{2D} and x_i^{3D} denote the i-th image and point cloud with y_i being the 3D segmentation label, while I_S, I_T represent the set of sample indices of the source and target domain.

Network Architecture. An overview of our network architecture is depicted in Figure 2. We tackle the task of 2D-3D cross-modal DA with two encoders to process data from the 2D and 3D modalities separately. Specificially, we encode the RGB images x^{2D} via a pretrained and frozen 2D Vision Foundation Model (VFM) $F^{2D}(\cdot;\theta_F^{2D})$ which yields a coarse patch feature map. In the 3D branch, we employ a 3D SparseConvNet [15] denoted as $F^{3D}(\cdot;\theta_F^{3D})$ to encode the point clouds x^{3D} into 3D features. In the 2D-3D cross-modal task, the 2D features are computed with the guidance of the input point cloud. Specifically, we first project 3D points onto the 2D image. Based on the projected pixel positions, we compute the pixel-wise 2D feature representation from patch features via bilinear interpolation.

For the task of cross-modal semantic segmentation, we introduce several task heads. We first denote the *main segmentation heads* for the 2D and 3D branch as $C^{2D}(\cdot;\theta_C^{2D})$ and $C^{3D}(\cdot;\theta_C^{3D})$ separately. Further, we follow [18] and introduce the *mimicry heads* that are additional segmentation task heads designed only for prediction alignment across different branches. The prediction alignment between dual heads across modalities are shown to be more parameter-robust than the case of a single head per modality [19]. Specifically, we introduce the mimicry head in the 3D branch which is denoted as $C^{3D}_{\text{mmc}}(\cdot;\theta^{3D}_{C_{\text{mmc}}})$. Note that we do not have a mimicry head for the 2D branch as we keep the 2D VFM frozen during training.

3.2. Fusion Branch

In order to enable cross-modal learning with enhanced guidance, we employ a fusion branch where we concatenate the 2D and 3D feature as the fusion feature $x^{fuse} = \text{concat}(F^{2D}(x^{2D};\theta_F^{2D}), F^{3D}(x^{3D};\theta_F^{3D}))$. Further, we employ an MLP as the fusion refinement network $F^{fuse}(\cdot;\theta_F^{fuse})$ to refine the fusion feature x^{fuse}.

Correspondingly, we construct the main segmentation head and the mimicry head for the fusion branch, denoted as $C^{fuse}(\cdot;\theta_C^{fuse})$ and $C^{fuse}_{\text{mmc}}(\cdot;\theta^{fuse}_{C_{\text{mmc}}})$. The fusion branch is illustrated in the purple components in Figure 2. During training with cross-modal alignment, our fusion branch provides the guiding signal for the 3D network. In the meanwhile, the prediction in the fusion branch is also regularized by the hypothesis in the 2D and 3D modalities. We elaborate the interaction between the fusion branch and the 2D or 3D branch in Section 3.3. We also ablate the impact of the fusion branch in Table 3. To obtain the final segmentation results we take the softmax average of the predictions from the 3D main head and the fusion main head.

3.3. Cross-Modal Learning with Fusion

We employ the fusion network as a cross-modal learner, trained under supervision to integrate 2D and 3D representations. To mitigate the susceptibility to domain shift, we introduce additional regularization through the 2D and 3D branches, resulting in a three-branch architecture that enables robust cross-modal learning.

Following the practice of cross-modal UDA [7, 36], we apply the KL divergence to encourage a mimicry distribution P_{mmc} (predictions from a mimicry head) to mimic a main head guiding distribution P_{main} (predictions from a main head). This is shown to be more parameter-robust than the cross-modal learning with only one task head per modality [19]. This self-supervision is applied on both source and target data, *i.e.,*

$$\mathcal{L}^{(D)}_{\text{KLD}}(P_{\text{main}}, P_{\text{mmc}}) = \sum_{i\in I_D} P_{\text{main}}(x_i)\log\frac{P_{\text{main}}(x_i)}{P_{\text{mmc}}(x_i)}, \quad (1)$$

where I_D is the corresponding domain sample index set.

In our empirical study, we realize that the fusion branch shows consistent improved prediction performance over the 2D branch, and therefore perform cross-modal alignment between the fusion branch and the 3D branch. We first apply the KL divergence encouraging the 3D mimicry distribution p^{3D}_{mmc} to mimic the fusion main head distribution p^{fuse}_{main}, *i.e.,*

$$\mathcal{L}^{(D)}_{\text{align}} = \sum_{i\in I_D} p^{fuse}_{\text{main}}(x_i)\log\frac{p^{fuse}_{\text{main}}(x_i)}{p^{3D}_{\text{mmc}}(x_i)}. \quad (2)$$

As the fusion branch is only supervised on the source domain, it is susceptible to domain shift from both modalities. To address this, we further regularize the fusion prediction to encourage consistency with the hypothesis of either the 2D or 3D modality on the fusion branch. *i.e.,*

$$\mathcal{L}^{(D)}_{\text{guide}} = \lambda\cdot\mathcal{L}^{(D)}_{\text{KLD}}(p^{2D}_{\text{main}}, p^{fuse}_{\text{mmc}}) + (1-\lambda)\cdot\mathcal{L}^{(D)}_{\text{KLD}}(p^{3D}_{\text{main}}, p^{fuse}_{\text{mmc}}) \quad (3)$$

Here, λ is the coefficient which biases the fusion towards the 2D or 3D branch and is chosen depending on the type of environmental conditions that may be expected in the task we aim to solve. We empirically demonstrate that this can lead to a more modality-biased fusion representation in Section 4.

3.4. Overall Training Objectives

For clarity, we denote the predicted logits from a main head as p^m_{main} and logits from a mimicry head as p^m_{mmc} where $p^m = C^m(F^m(x^m;\theta^m_F);\theta^m_C), m\in\{2D, 3D, fuse\}$.

Supervised Learning. For each of the 2D, 3D, and fusion branches, we perform the supervised segmentation task on the three main heads, *i.e.,*

$$\mathcal{L}^{(D)}_{\text{seg}}(x^m, y) = \sum_{i\in I_D} -y_i\log\left(p^m_{\text{main}(i)}\right), \quad (4)$$

where $m \in \{2D, 3D, fuse\}$. In the following, we denote the domain as $D \in \{S, T\}$. For the supervised learning on ground truth labeled source data, we have D = S.

Overall. The overall objective is the sum of the three supervised segmentation losses on source (Eq. (4)), and the two cross-modal losses on source and target (Eq. (2), (3)):

$$\min_{\theta} \frac{1}{|I_T|}\lambda_T(\mathcal{L}^{(T)}_{\text{align}} + \mathcal{L}^{(T)}_{\text{guide}}) + \frac{1}{|I_S|}\lambda_S(\mathcal{L}^{(S)}_{\text{align}} + \mathcal{L}^{(S)}_{\text{guide}}) + \frac{1}{|I_S|}\sum_{m\in M}\mathcal{L}^{(S)}_{\text{seg}}(x^m, y) \quad (5)$$

where $\theta = \{\theta^{3D}_F, \theta^{fuse}_F, \theta^{2D}_C, \theta^{3D}_C, \theta^{fuse}_C, \theta^{3D}_{C_{\text{mmc}}}, \theta^{fuse}_{C_{\text{mmc}}}\}$ and M = {2D, 3D, fuse}.

Additional Stage with Self-Training. To further boost the adaptation performance, we follow the practice of self-training in domain adaptation [9, 26, 30, 53], and conduct the second stage by adding the supervised loss on pseudo-labeled target data. Specifically, we compute the pseudo labels by averaging the softmax scores from the fusion and the 3D branch in the first-stage model:

$$\hat{y}_i = \frac{1}{2}\Big(\text{softmax}\left(p^{fuse}_{\text{main}}(x_i)\right) + \text{softmax}\left(p^{3D}_{\text{main}}(x_i)\right)\Big), i \in I_T. \quad (6)$$

In the second stage, the overall loss is the objective from stage 1 (Eq. (5)) together with the additional supervised loss on target $\frac{1}{|I_T|}\lambda_{PL}\sum_{m\in M}\mathcal{L}^{(T)}_{\text{seg}}(x^m, \hat{y})$.

4. Experiments

In this section we describe our conducted experiments along with the setup following the commonly used 3D UDA evaluation based on [19].

4.1. Datasets

We evaluate our method on the widely used nuScenes [5], SemanticKITTI [3] (SK), VirtualKITTI [12] (VK) and the A2D2 dataset [14]. All datasets provide a synchronized and calibrated setup that allows point to pixel projection. For simplicity and comparability, only front image sensor data is utilized. The DA tasks feature four domain shift scenarios, aiming to evaluate diverse scenarios: *USA* $\rightarrow$ *Singapore* evaluates adaptation between geographic regions using a similar sensor setup. The *Day* $\rightarrow$ *Night* scenario evaluates adaptation to a low light scenario. Both tasks are extracted from the nuScenes dataset, which provides a significant challenge for 3D modeling due to its low point density.

The *A2D2* $\rightarrow$ *SemanticKITTI* (A2D2 $\rightarrow$ SK) task is a challenging dataset adaptation task, since both 2D and 3D sensor systems significantly differ in terms of resolution, field of view, and LiDAR beam structure. The *VirtualKITTI* $\rightarrow$ *SemanticKITTI* (VK $\rightarrow$ SK) explores a virtual-to-real adaptation, aiming to study the adaptability from generated 2D and 3D data towards real data, where the generated point clouds are randomly sampled points from depth maps of the generated scenes. The different classes are merged to six classes (ten for the A2D2-SemanticKITTI task).

4.2. Implementation Details

In general, we follow the setup and hyperparameters provided by xMuda [19], however, we adjusted several important parameters, as detailed in the following paragraph. The pipeline consists of a pre-trained, frozen vision foundation model with a trainable linear head, a fusion network and a 3D encoder based on a U-Net [40] style 3D SparseConvNet [15]. We used batch size 24 for training. Further, for the training of the linear 2D classifier and the fusion network, we reduce the learning rate to 1×10^{-3}. We set the learning rate for the 3D model to 3×10^{-3}. For the pseudo-label loss, we used $\lambda_{PL} = 1$ for all datasets. We set the source and target alignment coefficients λ_S and λ_T to 1 and 0.1 for the NuScenes tasks and 0.5 and 0.5 for the A2D2-SemanticKITTI and VirtualKITTI-SemanticKITTI task, respectively. The fusion modality guidance λ is set to 1 for adaptation in daylight target domain tasks and 0 for night tasks to bias the fusion on modalities that are more robust in these lighting conditions.

Vision Foundation Models. In our experiments we employ the AM-Radio [39] VFM version 2.5-L as our primary model (patch size 16), chosen for its strong linear probing capabilities. To show the generalization of our approach, we also ablate with DINOv2 [34]. we observe minimal performance fluctuation.

High Resolution 2D Features. Since ViT Transformers operate on image patches, we apply two general strategies to obtain higher resolution pixel-level features, which are essential for dense prediction tasks such as semantic segmentation. We follow [28, 41] and apply cropout-resize with higher resolution in order to maximize the resolution for the patch-wise feature extraction for the ViT-encoder. Second, we add bilinear interpolation following [36, 37] to retrieve interpolated pixel-level features.

Fusion Network. We use an MLP with two hidden layers, each followed by batch normalization, GeLU nonlinearity, and a dropout layer. We choose the hidden dimensions of the same dimension as the 2D input, i.e., 1024 for the AM-

Method	VFM	USA → Singapore			Day → Night			VK → SK			A2D2 → SK			Avg
		2D	3D	2D3D	2D	3D	2D3D	2D	3D	2D3D	2D	3D	2D3D	
Source		58.4	62.8	68.2	47.8	68.8	63.3	26.8	42.0	42.2	34.2	35.9	40.4	49.2
Target		75.4	76.0	79.6	61.5	69.8	69.2	66.3	78.4	80.1	59.3	71.9	73.6	71.8
DsCML [35]		65.6	56.2	66.1	50.9	49.3	53.2	38.4	38.4	45.5	39.6	45.1	44.5	49.4
DsCML$_{PL}$ [35]		65.6	57.5	66.9	51.4	49.8	53.8	39.6	41.8	42.2	46.8	51.8	52.4	51.6
xMUDA [19]		64.4	63.2	69.4	55.5	69.2	67.4	42.1	46.7	48.2	38.3	46.0	44.0	54.5
BFtD [43]		63.7	62.2	69.4	57.1	70.4	68.3	41.5	45.5	51.5	40.5	44.4	48.7	55.3
SSE [51]		64.9	63.9	69.2	62.8	69.0	68.9	45.9	40.0	49.6	44.5	46.8	48.4	56.2
XMUDA$_{PL}$ [19]		67.0	65.4	71.2	57.6	69.9	64.4	45.8	51.0	52.0	41.2	49.8	47.5	56.9
SSE$_{PL}$ [51]		66.9	64.4	70.6	59.1	67.0	66.3	47.2	53.5	55.2	45.9	51.5	52.5	58.3
FtD++ [45]		69.7	64.6	69.8	68.8	69.6	71.0	51.0	44.0	52.6	48.8	46.2	51.1	58.9
BFtD$_{PL}$ [43]		65.9	66.0	71.3	60.6	70.0	66.6	48.6	55.4	57.5	42.6	53.7	52.7	59.2
MM2D3D [7]		71.7	66.8	72.4	70.5	70.2	72.1	53.4	50.3	56.5	42.3	46.1	46.2	59.9
FtD++$_{PL}$ [45]		71.7	65.5	72.3	68.9	70.3	71.8	52.9	51.2	57.8	51.4	49.7	54.8	61.5
MM2D3$_{PL}$ [7]		74.3	68.3	74.9	**71.3**	69.6	**72.2**	55.4	55.0	59.7	46.4	48.7	50.7	62.2
LTA-SAM$_{PL}$ [36]	✓	-	**73.6**	-	-	**70.5**	-	-	64.9	-	-	52.1	-	-
VFM-BOOST$_{PL}$ [47]	✓	70.0	65.6	72.3	60.6	**70.5**	66.5	57.2	52.0	61.0	45.0	52.3	50.0	60.3
UniDSeg [44]	✓	67.2	67.6	72.9	63.2	71.2	71.2	60.5	50.9	62.2	50.7	55.4	57.5	62.5
Ours	✓	74.4	67.9	75.4	67.9	68.1	70.3	70.1	64.5	70.7	60.3	54.9	63.1	67.2
Ours$_{PL}$	✓	**76.1**	70.5	**76.2**	69.2	69.0	70.4	**72.1**	**68.6**	**72.1**	**62.3**	**57.8**	**63.3**	**69.0**

Table 1. Quantitative (mIoU) results (**best**, second). P_L denotes two-stage training with pseudo-labels. For 2D we report the result of our fusion network. 2D3D denotes the softmax average of the fusion and 3D head. Source and Target: The baseline xMUDA implementation is trained either on source data or on target data, which serves as the lower and upper bound.

RADIO VFM. For the input we linearly project the 3D features to the 2D VFM dimension to reduce the bias from the large dimension gap between the 2D and 3D feature dimensions.

4.3. Experimental Results

We evaluate our method on the four commonly used 3D UDA tasks for semantic segmentation. Following previous works in the field [7, 18, 35], we evaluate the performance on the test set using the checkpoint that achieved the best IoU score on the validation set.

In Table 1 we report the results for our method and comparison with all other baselines and state-of-the-art methods. Generally, our method improves in most scenarios when compared with other approaches. We observe an impressive improvement of 6.5% on average (over all tasks), when compared with the strongest baseline UniDSeg [44], which prompt-tunes a VLM. On individual tasks, we also find that our method generally performs well. When comparing with the other approaches that also employ the recently proposed VFMs, we see a positive trend and we outperform the state-of-the-art VFM-BOOST [47] by an average of 7.5%, whereas obtaining up to 14.2% gains on the adaptation task of VirtualKITTI to SemanticKITTI. Similarly, we find that our method obtains an improvement over the other VFM-guided method (LTA-SAM [36]) of over 5% and 3% in the adaptation scenarios of A2D2 → SemanticKITTI and VirtualKITTI → SemanticKITTI, while remaining competitive on the other two tasks. These results highlight the benefits of our proposed fusion scheme and biasing of this fusion according to the downstream task of interest. In Table 1 we observe that our method fares better on most tasks but remains competitive on adaptation from Day → Night scenario, where we rank second by a small margin of 0.8% on average, when comparing to FtD++. FtD++, uses additional distillation to preserve the domain-specific attributes which helps them to perform well on the task of Day → Night adaptation, whereas our method outperforms them on all other scenarios, without requiring an additional distillation. Overall, we outperform FtD++ by an average of 7.5%. These strong results highlight the two main aspects of our method: the effective VFM utilization and the proposed modality guided fusion.

4.4. Ablation Study

In this subsection, we provide additional experiments validating the effectiveness of our method.

Component ablation study. We ablate our fusion adaptation method starting from a lightweight fusion proposed in [18], which consists of 2D and 3D feature concatenation followed by a projection layer and ReLU nonlinearity and evaluated the effectiveness of our proposed modality guidance. In addition, we ablate the utilization of an MLP instead of a single layered fusion and also report the impact of the modality guidance (MG) on the MLP. For comparison, we also provide results for a symmetric alignment (SymAl) where we align the fusion from both the VFM and 3D net-

Figure 3. Qualitative comparison of our method on an example from each dataset. We show the softmax average of our fusion and 3D head. Boxes mark locations of interest with zoom-in below. Multiple traffic participants are not detected by xMUDA-VFM-Fuse; **VK** $\rightarrow$ **SK**: A car is incorrectly identified as nature; **A2D2** $\rightarrow$ **SK**: Two persons are missed; **USA** $\rightarrow$ **Sing.** A bus is wrongly identified as a manmade structure. Our method correctly identifies these traffic participants likely due to our stronger reliance on the well-generalizing VFM features. **Day** $\rightarrow$ **Night** xMUDA-VFM-Fuse detects false positive vehicles, a potential sign of overreliance on visual features in low-light conditions which can be avoided with our proposed fusion regularization.

work in a symmetric way. This means, we added another mimicry head to the fusion network such that each modality (VFM, 3D) has their respective student on the fusion network.

The results in Table 2 show practically no gains when applying the guidance on the vanilla fusion, as its simple structure may prevent it from effectively responding to the modality guidance. Discernible improvements can be achieved when employing an MLP instead of a single layered fusion. The MLP fusion is improved when aligning the fusion from both the VFM and the 3D teacher, suggesting that a symmetric regularization is worthy in the absence of any priors regarding the more robust modality. However, for our evaluated tasks the full potential can be harnessed when the fusion head is guided toward the VFM on daylight target data and for the night task toward the 3D network.

Generalization beyond RADIO. In Figure 4 we evaluate our method by using DINOv2. We see that our method can generalize across VFMs. With the DINOv2 backbone, we observe a improvement of 1.3% for the USA $\rightarrow$ Singapore task, while we observe a degradation of 1.9% on the VirtualKITTI $\rightarrow$ Semantic KITTI adaptation. These re-

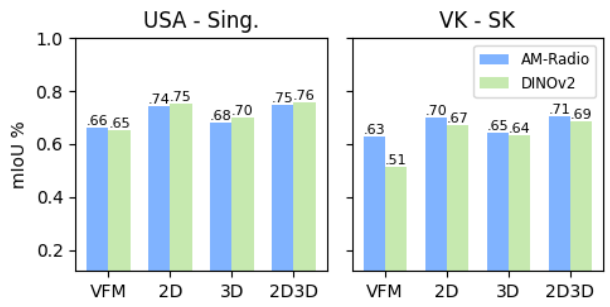

Figure 4. Comparison of current SOTA VFMs on USA → Sing. and VK → SK. We report the mIoU % for our main heads including the VFM head utilized for the fusion regularization.

Fusion	Day → Night			VK → SK		
	2D	3D	2D3D	2D	3D	2D3D
Vanilla	64.8	68.9	68.3	68.3	61.1	68.5
+ MG	64.3	67.8	67.8	68.8	64.2	69.7
MLP	65.9	67.5	68.8	69.7	63.2	69.2
MLP + SymAl	66.1	67.5	68.9	69.9	64.0	69.9
MLP + MG (ours)	67.9	68.1	70.3	70.1	64.5	70.7

Table 2. Fusion Ablation Study. Modality-guided (MG) fusion with an MLP is compared against vanilla fusion and symmetric alignment (SymAl) from both the 2D and 3D networks. Note, that the vanilla fusion here is conducted slightly differently than the xMUDA-Fuse method since we first project the 3D features to the dimension of the 2D features.

sults highlight the generalization ability of our method beyond AM-RADIO. Further, as VFMs advance, our method can also directly benefit and obtain further performance improvements.

xMUDA Fusion Ablation. In this experiment, we compare our method with xMUDA variants including their fusion variant that we implemented with our utilized VFM. The xMUDA fusion variant has only a single classifier on top of a linear layer and a ReLu nonlinearity. As input serves the feature concatenation of the 2D and 3D features. The supervision is only applied to the fusion classifier. Further, 2D and 3D stream encoders are aligned by the fusion classifier via a mimicry head. The results of this ablation are presented in Table 3. We find that our proposed fusion schema fares better than the fusion proposed by xMUDA. A notable result is obtained by comparing the xMUDA method and replacing the 2D feature extractor with the Radio VFM (employed in our work). We find that we outperform their method by 2% and 1.7% on the two adaptation scenarios we test.

2D3D	USA → Sing.	A2D2 → SK
xMUDA [19]	69.2	44.0
xMUDA-Fuse [19]	69.3	42.6
xMUDA-VFM-Fuse	73.4	61.4
Ours	75.4	63.1

Table 3. Comparison of xMUDA-Fuse variants with and without VFM backbone. Our method improves over other methods outlining the need for fusion adaptation strategies for VFM utilization.

Qualitative Results. Figure 3 depicts illustrative examples of our method and a comparison to xMUDA-VFM-Fuse. While the comparison implies similar performance for USA → Singapore and A2D2 → SK task, our method overall produces smoother segmentation masks compared to xMUDA-VFM-Fuse, while xMUDA-VFM-Fuse more often correctly classifies occluded points (resulting from the sensor system layout) which are especially present in the SemanticKITTI dataset and can be observed on most borders of close objects, well observable in the left vehicle in the VK → SK and the A2D2 → SK example.

Limitations The fusion guidance relies on a predefined modality preference informed by environmental conditions (e.g., prioritizing LiDAR under low-light and RGB in daylight). As such, it may be less effective in ambiguous conditions. In future work, we plan to explore mechanisms for estimating modality reliability at a per-point level, enabling flexible and fine-grained guidance based on spatial and semantic context.

5. Conclusion

We present a method that effectively leverages vision foundation models (VFMs) for 3D unsupervised domain adaptation (UDA) by introducing an adaptive fusion strategy informed by environmental conditions. Specifically, we show that biasing the fusion toward the more reliable modality, based on lighting conditions, can enhance adaptation performance. We extensively evaluate our proposed method on four commonly employed UDA benchmarks and demonstrate strong improvements over existing state-of-the-art methods. These results offer insights into how VFMs can be effectively integrated into multi-modal learning and highlight the potential of adaptive fusion schemes. We believe that fusion modules remain a promising direction for addressing challenges such as sensor misalignment, failures, and varying environmental conditions.

References

[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. In *ICLR*, 2022. 2

[2] Alejandro Barrera, Jorge Beltrán, Carlos Guindel, Jose Antonio Iglesias, and Fernando García. Cycle and semantic consistent adversarial domain adaptation for reducing simulation-to-real domain shift in lidar bird's eye view. In *ITSC*, pages 3081–3086, 2021. 2

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jürgen Gall. SemantickITTI: A dataset for semantic scene understanding of LiDAR sequences. In *CVPR*, pages 9297–9307, 2019. 5

[4] Tim Brödermann, Christos Sakaridis, Yuqian Fu, and Luc Van Gool. Condition-aware multimodal fusion for robust semantic perception of driving scenes. *arXiv preprint arXiv:2410.10791*, 2024. 2, 3

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 5

[6] Haozhi Cao, Yuecong Xu, Jianfei Yang, Pengyu Yin, Shenghai Yuan, and Lihua Xie. Mopa: Multi-modal prior aided domain adaptation for 3d semantic segmentation. In *ICRA*, pages 9463–9470, 2024. 3

[7] 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 *CVPR Workshop*, pages 98–109, 2023. 2, 4, 6

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *CVPR*, pages 9650–9660, 2021. 2

[9] Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift. In *NeurIPS*, pages 21061–21071, 2020. 5

[10] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018. 2

[11] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[12] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, pages 4340–4349, 2016. 5

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016. 2

[14] 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. 5

[15] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, pages 9224–9232, 2018. 4, 5

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *ICML*, pages 1989–1998, 2018. 2

[18] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation. In *CVPR*, 2020. 2, 3, 4, 6

[19] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Pérez. Cross-modal learning for domain adaptation in 3D semantic segmentation. In *PAMI*, 2022. 2, 4, 5, 6, 8

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 2

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *CVPR*, pages 4015–4026, 2023. 2

[22] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *ICLR*, 2017. 2

[23] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*, page 896, 2013. 2

[24] Jiale Li, Hang Dai, Hao Han, and Yong Ding. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In *CVPR*, pages 21694–21704, 2023. 2

[25] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018. 2

[26] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation. *NeurIPS*, pages 22968–22981, 2021. 5

[27] 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. 2

[28] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *NeurIPS*, pages 37193–37229, 2023. 5

[29] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 1

[30] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. Instance adaptive self-training for unsupervised domain adaptation. In *ECCV*, pages 415–430, 2020. 5

[31] Björn Michele, Alexandre Boulch, Gilles Puy, Tuan-Hung Vu, Renaud Marlet, and Nicolas Courty. Saluda: Surface-based automotive lidar unsupervised domain adaptation. In *3DV*, pages 421–431, 2024. 2

[32] M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *CVPR*, pages 14765–14775, 2022. 2

[33] Pietro Morerio, Jacopo Cavazza, and Vittorio Murino. Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In *ICLR*, 2018. 2

[34] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 5

[35] Duo Peng, Yinjie Lei, Wen Li, Pingping Zhang, and Yulan Guo. Sparse-to-dense feature matching: Intra and inter domain cross-modal learning in domain adaptation for 3d semantic segmentation. In *CVPR*, pages 7108–7117, 2021. 2, 6

[36] Xidong Peng, Runnan Chen, Feng Qiao, Lingdong Kong, Youquan Liu, T Wang, X Zhu, and Y Ma. Learning to adapt sam for segmenting cross-domain point clouds. In *ECCV*, 2024. 3, 4, 5, 6

[37] Gilles Puy, Spyros Gidaris, Alexandre Boulch, Oriane Siméoni, Corentin Sautier, Patrick Pérez, Andrei Bursuc, and Renaud Marlet. Three pillars improving vision foundation model distillation for lidar. In *CVPR*, pages 21519–21529, 2024. 5

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[39] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, pages 12490–12500, 2024. 2, 5

[40] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 5

[41] Corentin Sautier, Gilles Puy, Spyros Gidaris, Alexandre Boulch, Andrei Bursuc, and Renaud Marlet. Image-to-lidar self-supervised distillation for autonomous driving data. In *CVPR*, pages 9891–9901, 2022. 5

[42] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV Workshop*, pages 443–450, 2016. 2

[43] Yao Wu, Mingwei Xing, Yachao Zhang, Yuan Xie, Jianping Fan, Zhongchao Shi, and Yanyun Qu. Cross-modal unsupervised domain adaptation for 3d semantic segmentation via bidirectional fusion-then-distillation. In *ACMMM*, pages 490–498, 2023. 2, 3, 6

[44] Yao Wu, Mingwei Xing, Yachao Zhang, Xiaotong Luo, Yuan Xie, and Yanyun Qu. Unidseg: Unified cross-domain 3d semantic segmentation via visual foundation models prior. In *NeurIPS*, pages 101223–101249, 2024. 3, 6

[45] Yao Wu, Mingwei Xing, Yachao Zhang, Yuan Xie, and Yanyun Qu. Fusion-then-distillation: Toward cross-modal positive distillation for domain adaptive 3d semantic segmentation, 2024. arXiv preprint. 2, 3, 6

[46] Bowei Xing, Xianghua Ying, Ruibin Wang, Jinfa Yang, and Taiyan Chen. Cross-modal contrastive learning for domain adaptation in 3d semantic segmentation. In *AAAI*, pages 2974–2982, 2023. 2

[47] Jingyi Xu, Weidong Yang, Lingdong Kong, Youquan Liu, Rui Zhang, Qingyuan Zhou, and Ben Fei. Visual foundation models boost cross-modal unsupervised domain adaptation for 3d semantic segmentation, 2024. arXiv preprint. 3, 6

[48] Li Yi, Boqing Gong, and Thomas Funkhouser. Complete & label: A domain adaptation approach to semantic segmentation of lidar point clouds. In *CVPR*, pages 15363–15373, 2021. 2

[49] Zhimin Yuan, Ming Cheng, Wankang Zeng, Yanfei Su, Weiquan Liu, Shangshu Yu, and Cheng Wang. Prototype-guided multitask adversarial network for cross-domain lidar point clouds semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, pages 1–13, 2023. 2

[50] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks, 2017. 1

[51] Yachao Zhang, Miaoyu Li, Yuan Xie, Cuihua Li, Cong Wang, Zhizhong Zhang, and Yanyun Qu. Self-supervised exclusive learning for 3d segmentation with cross-modal unsupervised domain adaptation. In *ACMMM*, pages 3338–3346, 2022. 2, 6

[52] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *NeurIPS*, 2024. 3

[53] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *CVPR*, pages 5982–5991, 2019. 5