Cross-Modal Contrastive Learning for Domain Adaptation in 3D Semantic Segmentation

Bowei Xing, Xianghua Ying*, Ruibin Wang, Jinfa Yang, Taiyan Chen

Key Laboratory of Machine Perception (MOE)
School of Intelligence Science and Technology, Peking University
{xingbowei, xhying, robin_wang, jinfayang}@pku.edu.cn, chenty@stu.pku.edu.cn

Abstract

Domain adaptation for 3D point cloud has attracted a lot of interest since it can avoid the time-consuming labeling process of 3D data to some extent. A recent work named xMUDA leveraged multi-modal data to domain adaptation task of 3D semantic segmentation by mimicking the predictions between 2D and 3D modalities, and outperformed the previous single modality methods only using point clouds. Based on it, in this paper, we propose a novel cross-modal contrastive learning scheme to further improve the adaptation effects. By employing constraints from the correspondences between 2D pixel features and 3D point features, our method not only facilitates interaction between the two different modalities, but also boosts feature representations in both labeled source domain and unlabeled target domain. Meanwhile, to sufficiently utilize 2D context information for domain adaptation through cross-modal learning, we introduce a neighborhood feature aggregation module to enhance pixel features. The module employs neighborhood attention to aggregate nearby pixels in the 2D image, which relieves the mismatching between the two different modalities, arising from projecting relative sparse point cloud to dense image pixels. We evaluate our method on three unsupervised domain adaptation scenarios, including country-to-country, day-to-night, and dataset-to-dataset. Experimental results show that our approach outperforms existing methods, which demonstrates the effectiveness of the proposed method.

Introduction

3D semantic segmentation is an important task for scene understanding, aiming to predict the class label for each point in the LiDAR point cloud. Like many other tasks, 3D semantic segmentation also faces the problem of domain shift when the model is trained and evaluated in datasets from different domains. Several domain adaptation methods (Wu et al. 2019; Zhao et al. 2021; Jiang and Saripalli 2021; Liu et al. 2021a; Achituve, Maron, and Chechik 2021) have been proposed to narrow the domain gap for 3D segmentation, including discrepancy-based methods, adversarial-based methods and so on.

With the development of multi-modal learning, 2D image information is proved helpful for 3D scene understanding. Specifically, Jaritz et al. (2020; 2022) propose a framework that utilizes multi-modality data to address Unsupervised Domain Adaptation (UDA) problem for 3D segmentation. They design a cross-modal learning method where 2D images and 3D point clouds learn from each other through mutually mimicking the prediction scores between modalities, thus achieving adaptation effects in target domain.

In order to further boost the adaptation effects through cross-modal learning, we introduce a contrastive learning scheme to the UDA training process to facilitate information exchange between modalities. For the cross-modal 3D segmentation task, in each sample, the given 2D image and 3D point cloud represent the same semantic content and could provide complementary information for dense label prediction. Therefore, the domain gap in one modality can be narrowed with the guidance from the other modality. For example, 3D LiDAR point clouds obtained from different sensors usually have a large domain gap, in this case, the information of more consistent 2D images is able to help the prediction. In contrast, when camera images suffer from drastic performance degradation in low light conditions, the learning can get benefits from 3D LiDAR data because of its robustness to light. To this end, we utilize cross-modal contrastive learning to maximize mutual information between modalities by enforcing correspondence between 2D pixels and 3D points. Since contrastive learning is self-supervised with no ground truth labels needed, it can be applied on both labeled source domain and unlabeled target domain, which provides a suitable learning scheme for UDA task. Compared with xMUDA that adopts KL divergence to mimicking the prediction scores, our method is directly implemented in feature space to align corresponding pixel and point features. By this means, the 2D and 3D modalities can not only learn from each other through contrastive learning, but also obtain more discriminative feature representations.

In addition to the cross-modal contrastive learning scheme, we propose a neighborhood feature aggregation module to enable more image information can be taken into consideration for cross-modal learning. In xMUDA, there is a projection from 3D point cloud to 2D image, after which only the matched image pixels are retained. However, because the number of 2D pixels is much larger than the number of 3D points, the projection leads to the loss of massive image context features. To relieve this mismatching be-

*Xianghua Ying is the corresponding author.

tween modalities, we propose to enhance the sampled pixel features with neighborhood information using an aggregation module. Specifically, it leverages transformer blocks to adaptively aggregate the nearby pixel features from the image feature map, so the final obtained pixel feature contains richer semantic information. Furthermore, referring to the idea of dilated convolution (Yu and Koltun 2015), we extend to propose dilated neighborhood attention, which is able to enlarge the reception field to aggregate information from a larger scale. With the proposed module, image data is more sufficiently exploited in cross-modal learning, bringing improvements for both 2D and 3D predictions.

We carry out experiments on three different domain adaptation scenarios for 3D semantic segmentation, including day-to-night, country-to-country and dataset-to-dataset. Compared with other uni-modal and multi-modal UDA methods, our method obtains better performance. Through ablation study and analysis, it validates the effectiveness of the proposed cross-modal contrastive learning scheme and neighborhood feature aggregation module. Besides, it is proved that the two components can improve the segmentation result jointly, in the meanwhile compatible with other methods like pseudo-labeling. Our main contributions can be summarized as follows:

1. We introduce contrastive learning to the cross-modal domain adaptive 3D semantic segmentation task. It provides extra regularization to facilitate 2D and 3D modalities to learn from each other, thus improving the adaptation effects.

2. We develop a feature aggregation module to enhance pixel feature representations using neighborhood attention, with which more sufficient image context information can be employed for cross-modal learning.

3. Experiments on three different adaptation settings are carried out. The results demonstrate that our method achieves better performance compared to previous methods.

Related Work

Domain Adaptation

While domain adaptation methods have made significant progress for single modality setting (Farahani et al. 2021; Wilson and Cook 2020; Vu et al. 2019), multi-modal domain adaptation is still under explored. Recently, Munro and Damen (2020) and Kim et al.(2021) investigate UDA for multi-modal action recognition, by exploiting the correspondence of RGB data and optical flow data. xMUDA (Jaritz et al. 2020) proposes a cross domain framework with the input of image and point cloud, showing the effectiveness of multimodal data for UDA. In addition, Peng et al.(2021) and Liu et al.(2021a) integrate adversarial learning with multi-modal learning to advance the model generalization ability in target domain. Shin et al.(2022) explore multi-modal extension of test-time adaptation for 3D segmentation.

Following the previous work (Jaritz et al. 2020, 2022), in this paper, we intend to further investigate the problem of multi-modal UDA for 3D segmentation. Differently, we use a cross-modal contrastive learning scheme to improve feature representation in both domains, and a neighborhood feature aggregation module is employed additionally.

Contrastive Learning

Contrastive learning has shown its powerful ability in self-supervised training (Chen et al. 2020; He et al. 2020; Oord, Li, and Vinyals 2018), which facilitates representation learning by pulling features that are semantically similar and pushing away features that are semantically different. Recently, contrastive learning has been explored under several multi-modal learning scenarios, including vision-language (Radford et al. 2021; Wen et al. 2021; Yuan et al. 2021; Zhang et al. 2021; Bakkali et al. 2022), video-text (Yang, Bisk, and Gao 2021; Zolfaghari et al. 2021), image-point cloud (Lin et al. 2021; Afham et al. 2022; Liu et al. 2021b) etc. By aligning multi-modal features, it takes advantage of modality complementary to improve model performance. Specifically, CrossPoint (Afham et al. 2022) captures the correspondence between 3D objects and 2D images, designing an intra-modal and cross-modal contrastive loss function to learn transferable point cloud representations.

Different from CrossPoint that aligns complete 3D objects and 2D images, we construct the cross-modal contrastive loss by utilizing the correspondence between 3D Lidar points and 2D image pixels, which is implemented in a more fine-grained level. Besides, the proposed cross-modal contrastive learning is applied in UDA settings to improve adaptation effects.

Attention Mechanism

Attention mechanism (Vaswani et al. 2017) has been widely used in deep learning methods, which learns to attend the most relevant regions of the input by assigning different weights to different regions. Based on attention mechanism, the transformer architecture is proposed and has made significant progress in NLP (Devlin et al. 2018; Liu et al. 2019; Lan et al. 2019; Wolf et al. 2020) and CV (Dosovitskiy et al. 2021; Liu et al. 2021c; Touvron et al. 2021; Yang et al. 2021). In computer vision field, ViT (Dosovitskiy et al. 2021) divides the image into multiple patches, then directly processes the patch features with transformer, which outperforms state-of-the-art CNN methods on many benchmarks. Liu et al. (2021c) design Shifted Window Attention mechanism and proposed Swin Transformer. Recently, Hassani et al. (2022) propose Neighborhood Attention Transformer, which localizes the receptive field for each token to its neighborhood, so as to utilize both the power of attention and the efficiency of convolution.

In this paper, we implement neighborhood attention for the multi-modal learning, in order to attentively aggregate the local image features. Besides, we extend it to consider dilated neighborhood features, which further improve the representation learning.

Problem Definition

In this multi-modal UDA for 3D segmentation task, two different modalities including 3D Lidar point cloud and 2D camera image are considered. Suppose that we have two datasets from different domains, where the source dataset is $\mathcal{S} = \{X^s_{2D}, X^s_{3D}, Y^s_{3D}\}$ and the target dataset is $\mathcal{T} = \{X^t_{2D}, X^t_{3D}\}$. For each sample in source dataset, it consists

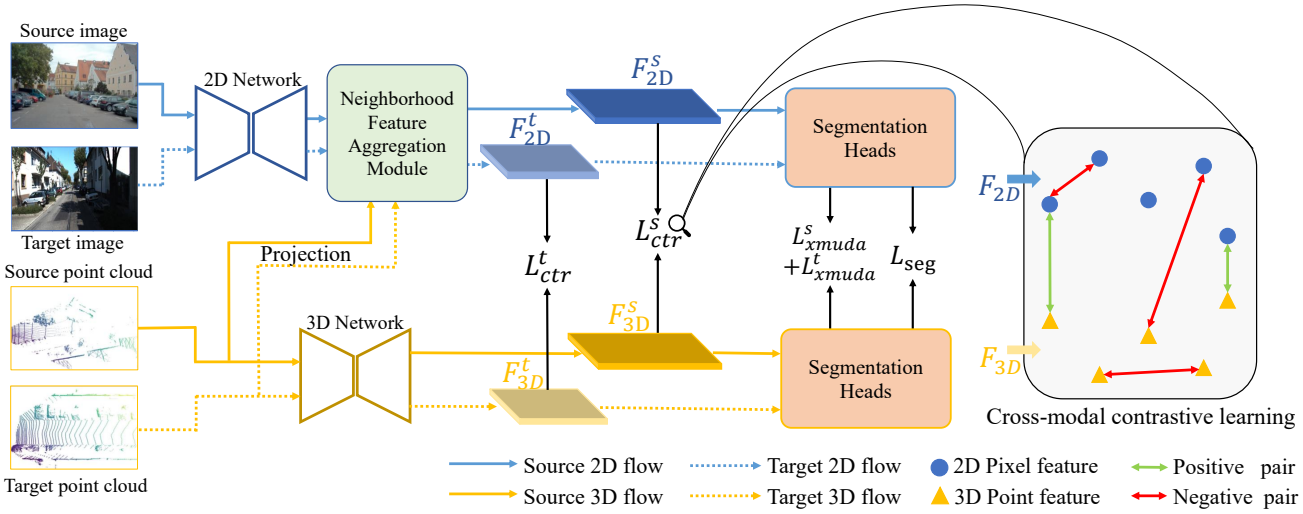

Figure 1: Overall training process for UDA 3D segmentation. We use a two-stream network for camera image and point cloud. After processed by backbone network and the proposed neighborhood feature aggregation module, we obtain features for each domain and modality. In both source domain and target domain, the corresponding 2D features and 3D features are utilized to calculate cross-modal contrastive loss, where the matched pixel and point constitute a positive pair while those unmatched are seen as negative pairs. Following is the segmentation heads, whose outputs are used for segmentation loss and xmuda loss.

of an image x^s_{2D}, a point cloud x^s_{3D} and its point-wise 3D segmentation label y^s_{3D}. For the target dataset, there are no available labels. The size of 2D image is $(H, W, 3)$ and the size of 3D point cloud is $(N, 3)$, where N is the number of 3D points inside the camera view. It is noted that the given segmentation label y^s_{3D} is of size N, corresponding to the 3D point cloud. There are no direct 2D image labels provided, but we can obtain the image pixel labels by projecting the labeled point cloud on the image. After projection, a part of image pixels are retained and obtain their labels, denoted as y^s_{2D}, which is also of size N.

We intend to learn a model with the labeled source dataset $\mathcal{S}$ and unlabeled target dataset $\mathcal{T}$, utilizing the multi-modal data from both datasets. The goal is the trained model can adapt well and correctly predict the point labels in target domain, even though there are not any labels provided in target dataset.

Method

Overview of the Framework

The architecture of our proposed method is illustrated in Figure 1. We use a two-stream network for image and point cloud. With the input from source and target dataset, we first extract features using 2D network and 3D network respectively. Then image features are sampled based on the projection from point cloud to image, and enhanced by the neighborhood feature aggregation module. After obtaining pixel features F^s_{2D}, F^t_{2D} and point features F^s_{3D}, F^t_{3D}, the features of each domain are enforced to interact with each other via cross-modal contrastive learning. Finally, we use segmentation heads to get prediction scores for each modality, which are used for the calculation of segmentation loss and xmuda loss. Details will be introduced in following sections.

Cross-Modal Contrastive Learning

In each domain, suppose that we have obtained the 2D pixel features $F_{2D} = \{f^i_{2D}\}^N_{i=1}$ and 3D point features $F_{3D} = \{f^i_{3D}\}^N_{i=1}$, where pixels and points are one-to-one corresponding after projection. The matched pixel-point pair not only belongs to the same category, but also represents similar semantic information. Based on this observation, we introduce a cross-modal contrastive learning objective for each domain, in order to utilize complementary 2D and 3D data to facilitate modalities learning from each other. By aligning the corresponding pixel features and point features, we aim to enforce the consistency between modalities and jointly improve 2D and 3D feature representations in both source and target domains.

Specifically, we first map the feature vectors to a joint feature space R^d using a 2D projection head ϕ_{2D} and a 3D projection head ϕ_{3D}. Following previous contrastive learning works (Chen et al. 2020; He et al. 2020), we measure the similarity between feature vectors by calculating cosine distance of the projected features in R^d, where cosine distance function is defined as $cos(u, v) = u^T v / ||u|| \, ||v||$. So the similarity score between f^i_{2D} and f^i_{3D} is:

$$s(f^i_{2D}, f^i_{3D}) = cos(\phi_{2D}(f^i_{2D}), \phi_{3D}(f^i_{3D})) \quad (1)$$

We would like to enforce the consistency between modalities by pulling close the 2D pixel feature with its corresponding 3D point feature while pushing away other features. Therefore, by implementing the NT-Xent loss (Chen et al. 2020), our cross-modal contrastive loss between f^i_{2D} and f^i_{3D} can be formulated as:

$$l(f^i_{2D}, f^i_{3D}) = -\log \frac{\exp\left(s\left(f^i_{2D}, f^i_{3D}\right)/\tau\right)}{\sum_{\substack{k=1 \\ k\neq i}}^{N} \exp\left(s\left(f^i_{2D}, f^k_{2D}\right)/\tau\right) + \sum_{k=1}^{N} \exp\left(s\left(f^i_{2D}, f^k_{3D}\right)/\tau\right)} \quad (2)$$

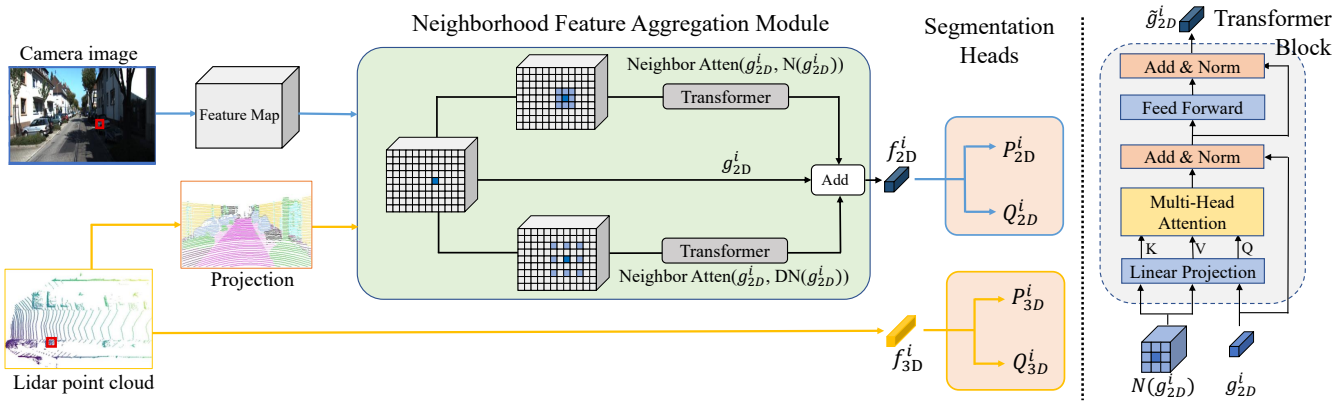

Figure 2: Illustration of neighborhood feature aggregation module. After the projection from point cloud to image, we use transformer blocks that leverage neighborhood attention to aggregate information from neighbors and dilated neighbors, which relieves the information loss in xMUDA (Jaritz et al. 2020, 2022). Then dual segmentation heads are used to obtain predictions for each modality. The small red squares on input data refer to the matched pixel and point.

where τ is a temperature co-efficient. For a 2D pixel feature vector f^i_{2D}, the corresponding (f^i_{2D}, f^i_{3D}) is regarded as positive pair, and the rest $2N$–2 feature vectors constitute negative examples. We intend to maximize the similarity of positive pair, in the meanwhile minimizing the similarity of negative pairs in the joint feature space.

The above loss function is asymmetric, which regards the 2D pixel feature as anchor, then attracts or separates other 3D features to boost 2D learning. On the contrary, we can regard the 3D point feature as the anchor and add another loss item, so as to leverage 2D information for 3D learning. In order to enable each modality can utilize the complementary information from the other modality, the final cross-modal contrastive loss is formulated as:

$$L_{ctr}(F_{2D}, F_{3D}) = \frac{1}{2N}\sum_{i=1}^{N}[l(f^i_{2D}, f^i_{3D}) + l(f^i_{3D}, f^i_{2D})] \tag{3}$$

In consideration of empirical GPU memory concern, we are not able to use total $2N$ features to calculate the contrastive loss, so we need to select a portion of total point and pixel features. For the sampling strategy, assuming the selected feature number is $2N_c$, we will select top N_c pixel features that have the highest prediction confidence from 2D stream and their corresponding 3D point features from 3D stream. The sampled features are discriminative because of the high prediction confidence, which is helpful for cross-modal contrastive learning. This sampling strategy is also analyzed in ablation experiment. There is no memory bank needed due to the large number of pixel and point features.

Neighborhood Feature Aggregation Module

With the input x_{2D} and x_{3D}, after feature extraction by two independent 2D and 3D backbone networks, we obtain the image feature map $G_{2D} = \{g_{2D}\}$ which is of size (H, W, C_{2D}) and point cloud feature $F_{3D} = \{f_{3D}\}$ of size (N, C_{3D}), where C is the feature channel. Based on the projection from point cloud to image. N pixel features are retained. Since the number of 2D features in feature map is much bigger than that of 3D features ($H \times W$ is quite larger than N), a lot of useful image information is lost in this process. To relieve the above issue, we propose a neighborhood feature aggregation module to enhance the sampled pixel features, as shown in Figure 2.

To be specific, for point feature f^i_{3D}, assume its corresponding pixel feature is g^i_{2D} in the feature map. We denote the neighbor features of g^i_{2D} as $\mathcal{N}(g^i_{2D})$, which consists of the nearest pixels around g^i_{2D} in image. Consider a neighbor region of size $k \times k$, the obtained neighbor features $\mathcal{N}(g^i_{2D})$ is of size (k^2, C_{2D}), including g^i_{2D} itself. We regard g^i_{2D} as a query token, while $\mathcal{N}(g^i_{2D})$ constitutes the key-value pair. They serve as the input for a transformer block, in which we implement the neighborhood attention mechanism.

Concretely, taking g^i_{2D} and $\mathcal{N}(g^i_{2D})$ as input, we first use linear projections to compute a set of query, key and value (Q, K and V). Then we employ multi-head attention mechanism, where we compute the attention weights using the scaled dot products between Q and K and then multiply the weights with V, obtaining the feature vector h^i_{2D}:

$$Q = g^i_{2D}W_q,\ K = \mathcal{N}(g^i_{2D})W_k,\ V = \mathcal{N}(g^i_{2D})W_v \tag{4}$$

$$h^i_{2D} = softmax(\frac{QK^T}{\sqrt{C_{2D}}})V \tag{5}$$

where $W_q, W_k, W_v \in R^{C_{2d} \times C_{2d}}$ are weighting matrices.

Next, the original input g^i_{2D} is added to h^i_{2D} with a skip connection, and a feed forward network is applied. We denote the final output as $\tilde{g}^i_{2D}$, which can be seen as the updated feature representation of g^i_{2D}.

$$\tilde{h}^i_{2D} = h^i_{2D} + g^i_{2D} \tag{6}$$

$$\tilde{g}^i_{2D} = FFN(\tilde{h}^i_{2D}) + \tilde{h}^i_{2D} \tag{7}$$

To further enlarge the reception filed of local regions, we adopt the idea of dilated convolution (Yu and Koltun 2015), and propose dilated neighborhood attention. The dilated neighbors refer to the neighbor pixels around g^i_{2D}

where each pixel has a distance of r with the nearest one, and r is called dilated rate. Suppose that we would like to select k^2 dilated neighbors with dilated rate 2, the attended neighbor region can be $(2k-1)\times(2k-1)$, which provides a larger reception filed for feature learning. We denote the dilated neighbors of g^i_{2D} as $\mathcal{DN}(g^i_{2D})$. Using g^i_{2D} and $\mathcal{DN}(g^i_{2D})$ as input, we can use the transformer block to update g^i_{2D} with the information from dilated neighbors, leading to the output $\hat{g}^i_{2D}$.

Finally, we add $g^i_{2D}, \tilde{g}^i_{2D}, \hat{g}^i_{2D}$to get the aggregated feature f^i_{2D}, which is a more representative 2D pixel feature. The above process can be formulated as:

$$\tilde{g}^i_{2D} = Neighborhood\_Attention(g^i_{2D}, \mathcal{N}(g^i_{2D})) \quad (8)$$

$$\hat{g}^i_{2D} = Neighborhood\_Attention(g^i_{2D}, \mathcal{DN}(g^i_{2D})) \quad (9)$$

$$f^i_{2D} = g^i_{2D} + \tilde{g}^i_{2D} + \hat{g}^i_{2D} \quad (10)$$

This operation is repeated for every pixel that matches 3D point in projection. Since the weight coefficients that aggregate neighbor features are learned by attention, we can focus on those more important neighbor features for each pixel respectively. Compared with the original feature g^i_{2D}, our aggregated feature f^i_{2D} enables more sufficient image information can be utilized for cross-modal learning.

Segmentation Heads

With the 3D point feature f^i_{3D} and the aggregated 2D pixel feature f^i_{2D}, we use dual segmentation heads to predict class label probabilities. In each modality, there are two segmentation heads, where the main one is used for best possible class label prediction, and the second one serves as the supervision for the other modality's prediction. We denote the output of 2D stream as P^i_{2D} and Q^i_{2D}, and output of 3D stream as P^i_{3D} and Q^i_{3D}. The segmentation loss is calculated by classical cross-entropy using P^i and ground-truth label y^i, and the xmuda loss (Jaritz et al. 2020) is calculated by estimating the other modality's main prediction with Q^i. In formulation, the loss functions can be written as :

$$L_{seg} = -\frac{1}{N}\sum_{i=1}^{N}(y^i_{2D} log P^i_{2D} + y^i_{3D} log P^i_{3D}) \quad (11)$$

$$L_{xmuda} = \frac{1}{N}\sum_{i=1}^{N}(D_{KL}(P^i_{2D}||Q^i_{3D}) + D_{KL}(P^i_{3D}||Q^i_{2D})) \quad (12)$$

where D_{KL} is KL divergence.

It is worth mentioning that in above functions, P^i_{2D} and Q^i_{2D} are the segmentation outputs of the enhanced 2D feature f^i_{2D}. Since f^i_{2D} is attentively aggregated by neighborhood information from different reception fields, it has a direct lifting effect for the prediction in 2D stream. Besides, on account of the usage of more pixel features, it is able to provide better guidance for 3D modality learning, which is also beneficial for the prediction in 3D stream.

Overall Objective

For this multi-modal UDA 3D segmentation task, the 2D and 3D network are trained jointly and optimized with data from source and target in each iteration.

Considering cross-modal learning in both source and target domains, the total loss during training can be written as:

$$\mathcal{L} = L^s_{seg} + \lambda_1(L^s_{xmuda} + L^t_{xmuda}) + \lambda_2(L^s_{ctr} + L^t_{ctr}) \quad (13)$$

where the superscript s, t respectively refer to source and target dataset, λ_1 and λ_2 are hyperparameters used to balance the loss components. Since the contrastive learning is self-supervised and does not need labels, our proposed learning scheme is able to provide extra regularization for feature learning in both source and target domains, which can improve adaptation ability of the model.

Besides, we can also train the model with pseudo-labeling (PL). When trained with PL, we need to add an additional segmentation loss item from the target training set:

$$\mathcal{L}_{PL} = \mathcal{L} + \lambda_{PL} L^t_{seg} \quad (14)$$

where λ_{PL} is coefficient, L^t_{seg} is calculated by the cross-entropy of predictions and pseudo-labels in target dataset.

Experiment

Datasets

Following previous work (Jaritz et al. 2020, 2022), we evaluate the performance of our method in three real-to-real adaption settings, including country-to-country, day-to-night and dataset-to-dataset. Three autonomous driving datasets: nuScenes (Caesar et al. 2020), A2D2 (Geyer et al. 2020) and SemanticKITTI (Behley et al. 2019) are adopted. In each dataset, LiDAR and RGB-camera are synchronized and calibrated, so that we can obtain the correspondence between 3D point and 2D image pixel from projection.

There are two adaptation scenarios generated by nuScenes: USA/Singapore and Day/Night. The objects are classified into 5 classes, $i.e.$ vehicle, pedestrian, bike, traffic boundary and background. For USA/Singapore, in some cases the 3D domain has a larger domain gap than 2D and vice versa. In Day/Night, 2D images show considerable difference because of the lighting condition, while 3D data is more robust. As for dataset-to-dataset UDA setting, we generate A2D2/SemanticKITTI adaptation scenario, where ten shared classes are considered. Since the sensors for LiDAR data collection are different in the two datasets, the 3D domain gap is larger. Details are introduced in Appendix.

Implementation

Backbone Network. For the sake of a fair and direct comparison, we use the same backbone network as xMUDA. Specifically, we use a modified version of U-Net (Ronneberger, Fischer, and Brox 2015) as 2D backbone and SparseConvNet (Graham, Engelcke, and Van Der Maaten 2018) as 3D backbone. The 3D voxel size is set as 5cm, which is small enough to ensure each voxel only contains one 3D point.

Parameter Setting. In training process, the learning rate is set to 0.001 at initial and is divided by 10 at 80k and 90k iterations. We totally train the model for 100k iterations on each adaptation scenario. For the neighborhood features, we

Modality	Method	USA → Singapore			Day → Night			A2D2 → SemanticKITTI		
		2D	3D	2D+3D	2D	3D	2D+3D	2D	3D	2D+3D
	Source only	53.2	46.8	61.2	41.8	41.4	47.6	36.4	37.3	42.2
Uni-Modal	Deep logCORAL(2018)	52.6	47.1	59.1	41.4	42.8	51.8	35.8	39.3	40.3
	MinEnt(2019)	53.4	47.0	59.7	44.9	43.5	51.3	38.8	38.0	42.7
	PL(2019)	55.5	51.8	61.5	43.7	45.1	48.6	37.4	44.8	47.7
	CLAN(2019)	57.8	51.2	62.5	45.6	43.7	49.2	39.2	44.7	44.5
Multi-Modal	xMUDA(2020)(Baseline)	59.3	52.0	62.7	46.2	44.2	50.0	38.3	46.0	44.0
	AUDA(2021a)	59.8	52.0	63.1	49.0	47.6	54.2	43.0	43.6	46.8
	DsCML(2021)	61.3	53.3	63.6	48.0	45.7	51.0	39.6	45.1	44.5
	DsCML+CMAL(2021)	63.4	55.6	64.8	49.5	48.2	52.7	46.3	50.7	51.0
	xMUDA+PL(2020)	61.1	54.1	63.2	47.1	46.7	50.8	41.2	49.8	47.5
	AUDA+PL(2021a)	61.9	54.8	65.6	50.3	49.7	52.6	46.8	48.1	50.6
	DsCML+CMAL+PL(2021)	**63.9**	56.3	65.1	50.1	48.7	53.0	46.8	**51.8**	52.4
	Ours	62.0	54.2	64.5	49.0	46.6	51.6	42.8	47.7	48.0
	Ours+PL	63.3	**57.1**	**66.7**	**50.6**	**49.9**	**54.5**	**47.6**	51.3	**52.8**
	Target only (Oracle)	66.4	63.8	71.6	48.6	47.1	55.2	58.3	71.0	73.7

Table 1: Comparison results with other methods in different cross-modal UDA scenarios. The best results are marked in bold.

adopt the nearby region of 5×5. For dilated neighbor features, we sample the features from the nearby 9×9 region with dilated rate 2, which also leads to a number of 25 features for each pixel. The batch size is set as 8 for USA/Singapore and Day/Night, and 6 for the A2D2/SemanticKITTI. Due to GPU memory limitation, 30% features in each mini-batch are sampled to calculate contrastive loss in the former two scenarios and 20% features are sampled in the last scenario. The 2D and 3D network are trained jointly and optimized on source and target for each iteration.

Evaluation Metric. For evaluation, we adopt the commonly used segmentation evaluation metric, mean Intersection-over-Union (mIoU) to evaluate the performance. In addition to the direct prediction results of 2D and 3D stream, we also evaluate with the '2D+3D' results, which is the average of predicted 2D and 3D probabilities after softmax.

Comparison Results

In this section, we evaluate the proposed method on the above three adaption scenarios, comparing with uni-modal and multi-modal UDA methods. To be specific, the compared uni-modal methods include MinEnt (Vu et al. 2019), pseudo-labeling(Li, Yuan, and Vasconcelos 2019), Deep logCORAL(Morerio, Cavazza, and Murino 2018), and CLAN(Luo et al. 2019). The multi-modal methods include xMUDA(Jaritz et al. 2020), DsCML(Peng et al. 2021) and AUDA(Liu et al. 2021a), which are designed for domain adaptation with multi-modal data. In the above multi-modal methods, the former one uses dual segmentation heads and proposes xmuda cross-modal loss, and the other two methods introduce adversarial learning to the multi-modal learning. As other works did, we also carry out experiments to prove the effectiveness of our model with pseudo-labeling (PL). The comparison results are shown in Table 1.

From Table 1 we can observe that, with our proposed method, the segmentation results significantly outperform the source only method and uni-modal UDA methods in all three adaptation scenarios. It proves the effectiveness of utilizing multi-modality data. For the comparison with xMUDA, which can be seen as the baseline of our method, we achieve great improvements in both training settings that with PL and without PL. Since we use the same backbone as xMUDA, the improvement can be directly attributed to the cross-modal contrastive loss and aggregation module. Compared with DsCML and AUDA that use adversarial learning, our method employs contrastive learning scheme for multi-modal domain adaptation and obtains better performance in most metrics. It is also noted that in Day/Night scenario, we even outperform the model directly trained on target data, which proves the effectiveness of using source data and introducing cross-modal domain adaptation method. Besides, it is worth mentioning that our method is compatible with the adversarial training used in DsCML/AUDA. The combination with adversarial training can relieve the distribution difference between domains and is likely to further boost the performance, which can be explored in future.

Baseline	NFAM	Ctr	PL	2D	3D	2D+3D
✓				59.3	52.0	62.7
	✓			61.3	53.6	63.8
		✓		60.1	53.2	63.0
	✓	✓		62.0	54.2	64.5
			✓	61.1	54.1	63.2
	✓		✓	61.9	55.1	64.8
	✓	✓	✓	**63.3**	**57.1**	**66.7**

Table 2: Ablation study on USA/Singapore.

Ablation Studies

Next, we conduct ablation studies on USA/Singapore dataset to analyze the effectiveness of each component.

Effectiveness of Proposed Components. Beginning with baseline method (xMUDA), we gradually add the aggregation module and the contrastive loss to validate the performance improvement. The ablation study results are shown in Table 2. From the results, we can observe that the two components are able to improve the performance over the baseline method respectively. By combining both, it provides a

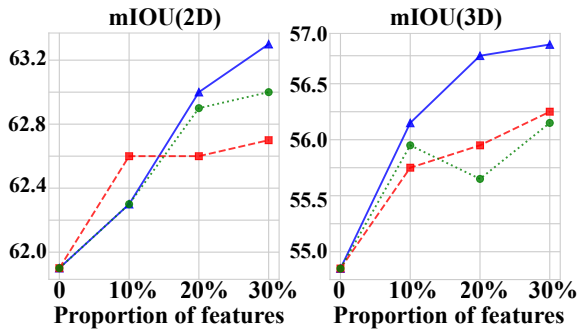

Figure 3: Analysis of cross-modal contrastive learning.

Method	2D	3D	2D+3D
xMUDA (2020; 2022)	59.3	52.0	62.7
sCML(2021)	60.6 (↑ 1.3)	52.5 (↑ 0.5)	63.2 (↑ 0.5)
DsCML(2021)	**61.3** (↑ **2.0**)	53.3 (↑ 1.3)	63.6 (↑ 0.9)
NFAM w/o $\mathcal{DN}$	61.0 (↑ 1.7)	53.5 (↑ 1.5)	63.6 (↑ 0.9)
NFAM w/. $\mathcal{DN}$	**61.3** (↑ **2.0**)	**53.6** (↑ **1.6**)	**63.8** (↑ **1.1**)

Table 3: Comparison of neighborhood feature aggregation module (NFAM) with xMUDA, sCML and DsCML. $\mathcal{DN}$ refers to dilated neighbors.

higher accuracy in all three evaluation metrics. In the last several lines of Table 2, the experiments are conducted with pseudo-labeling (PL). When combined with our proposed method, it achieves the best performance, showing that our method is complementary with PL, and they can contribute to better adaptation ability in a joint manner.

Cross-Modal Contrastive Learning. In this part, we investigate the contribution of cross-modal contrastive learning by varying the number of features used in calculating contrastive loss. Besides, we implement three different sampling strategies, *i.e.* select features randomly, select features with high prediction confidence, and select features with low prediction confidence. The results of different experiment settings are reported in Figure 3. The x-axis refers to the proportion of sampled features taken from total features in a mini-batch. When the proportion is 0, it is equivalent to training the model without contrastive loss, then we gradually add the proportion to 10%, 20% and 30%. From Figure 3 we can observe that, with the increasing number of sampled features, the performance presents a general upward trend. It can be explained as, with more features considered, there are more negative pairs, which improve the effect of contrastive learning. For the sampling strategy, selecting features with high prediction confidence achieves better performance, empirically validating our sampling strategy. When sampling 30% features with the highest prediction confidence, our method obtains the best result.

Neighborhood Feature Aggregation Module. The aggregation module attentively aggregates the nearby features

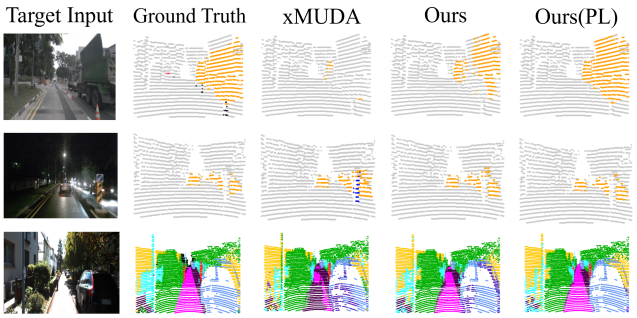

Figure 4: Visualization of our method on three adaptation scenarios. Different colors refer to different object classes.

with neighborhood attention, which is able to obtain more representative features. To demonstrate the module's effect, we specifically compare it with other methods, including xMUDA, sCML and DsCML. xMUDA (Jaritz et al. 2020, 2022) is original method that uses the only projected 2D pixel feature. Peng et al. (2021) propose sCML and DsCML, which utilize multiple probability scores from 2D patch to implement sparse-to-dense loss. The comparison results are listed in Table 3. It validates the effectiveness of the proposed aggregation module. By attentively attending to neighborhood regions and employing the aggregated features to exchange information with 3D features, the module is proved to be beneficial for both 2D and 3D modality learning. Furthermore, with the dilated neighborhood features considered, better performance is obtained.

Visualization

In Figure 4, we visualize the segmentation results on the three adaptation settings. As the figure shows, compared to baseline, our method predicts the category labels more accurately, in the meanwhile with less false predictions.

Conclusion

In this paper, we propose cross-modal contrastive learning and neighborhood feature aggregation module to investigate domain adaptive 3D semantic segmentation task. By enforcing the consistency between 2D and 3D modalities and considering more sufficient context information for cross-modal learning, the two components can jointly improve adaptation effects. Experiment results show that our method establishes a marked improvement over previous methods on three different unsupervised domain adaptation scenarios.

Specifically, here we give a brief discussion about cross-modal contrastive learning. Although contrastive learning has achieved excellent performance in many fields, it is still challenging to apply contrastive learning in cross-modal tasks. The difficulty lies in the inhomogeneity and non-correspondence of different modalities, and the construction of positive and negative pairs can be complicated and time-consuming. In this paper, using the calibration between image and point cloud, we ingeniously construct contrastive learning pairs based on the one-to-one correspondence of pixel and point features, which is intuitive and effective.

Acknowledgements

This work was supported in part by National Key R&D Program of China Grant No. 2020YFB1708002, and NNSFC Grant No. 61971008.

References

Achituve, I.; Maron, H.; and Chechik, G. 2021. Self-supervised learning for domain adaptation on point clouds. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 123–133.

Afham, M.; Dissanayake, I.; Dissanayake, D.; Dharmasiri, A.; Thilakarathna, K.; and Rodrigo, R. 2022. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9902–9912.

Bakkali, S.; Ming, Z.; Coustaty, M.; Rusiñol, M.; and Terrades, O. R. 2022. VLCDoC: Vision-Language Contrastive Pre-Training Model for Cross-Modal Document Classification. *arXiv preprint arXiv:2205.12029*.

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

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

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations*, 1–12.

Farahani, A.; Voghoei, S.; Rasheed, K.; and Arabnia, H. R. 2021. A brief review of domain adaptation. *Advances in data science and information engineering*, 877–894.

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

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

Hassani, A.; Walton, S.; Li, J.; Li, S.; and Shi, H. 2022. Neighborhood Attention Transformer. *arXiv preprint arXiv:2204.07143*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

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

Jaritz, M.; Vu, T.-H.; Charette, R. d.; Wirbel, E.; and Pérez, P. 2022. Cross-modal Learning for Domain Adaptation in 3D Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Kim, D.; Tsai, Y.-H.; Zhuang, B.; Yu, X.; Sclaroff, S.; Saenko, K.; and Chandraker, M. 2021. Learning cross-modal contrastive features for video domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13618–13627.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Li, Y.; Yuan, L.; and Vasconcelos, N. 2019. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6936–6945.

Lin, M.-X.; Yang, J.; Wang, H.; Lai, Y.-K.; Jia, R.; Zhao, B.; and Gao, L. 2021. Single Image 3D Shape Retrieval via Cross-Modal Instance and Category Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11405–11415.

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

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y.-C.; Huang, Y.-K.; Chiang, H.-Y.; Su, H.-T.; Liu, Z.-Y.; Chen, C.-T.; Tseng, C.-Y.; and Hsu, W. H. 2021b. Learning from 2D: Contrastive Pixel-to-Point Knowledge Transfer for 3D Pretraining. *arXiv preprint arXiv:2104.04687*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021c. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Luo, Y.; Zheng, L.; Guan, T.; Yu, J.; and Yang, Y. 2019. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2507–2516.

Morerio, P.; Cavazza, J.; and Murino, V. 2018. Minimal-Entropy Correlation Alignment for Unsupervised Deep Domain Adaptation. In *International Conference on Learning Representations*, 1–10.

Munro, J.; and Damen, D. 2020. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 122–132.

Oord, A. V. D.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241.

Shin, I.; Tsai, Y.-H.; Zhuang, B.; Schulter, S.; Liu, B.; Garg, S.; Kweon, I. S.; and Yoon, K.-J. 2022. MM-TTA: Multi-Modal Test-Time Adaptation for 3D Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16928–16937.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 10347–10357.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30: 6000–6010.

Vu, T.-H.; Jain, H.; Bucher, M.; Cord, M.; and Pérez, P. 2019. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2517–2526.

Wen, K.; Xia, J.; Huang, Y.; Li, L.; Xu, J.; and Shao, J. 2021. Cookie: Contrastive cross-modal knowledge sharing pre-training for vision-language representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2208–2217.

Wilson, G.; and Cook, D. J. 2020. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5): 1–46.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 38–45.

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

Yang, J.; Bisk, Y.; and Gao, J. 2021. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11562–11572.

Yang, J.; Li, C.; Zhang, P.; Dai, X.; Xiao, B.; Yuan, L.; and Gao, J. 2021. Focal attention for long-range interactions in vision transformers. *Advances in Neural Information Processing Systems*, 34: 30008–30022.

Yu, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Yuan, X.; Lin, Z.; Kuen, J.; Zhang, J.; Wang, Y.; Maire, M.; Kale, A.; and Faieta, B. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6995–7004.

Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 833–842.

Zhao, S.; Wang, Y.; Li, B.; Wu, B.; Gao, Y.; Xu, P.; Darrell, T.; and Keutzer, K. 2021. ePointDA: An end-to-end simulation-to-real domain adaptation framework for LiDAR point cloud segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 3500–3509.

Zolfaghari, M.; Zhu, Y.; Gehler, P.; and Brox, T. 2021. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1450–1459.