Image-to-Lidar Relational Distillation for Autonomous Driving Data

Anas Mahmoud¹, Ali Harakeh², and Steven Waslander¹

¹ University of Toronto
 ² Mila - Quebec AI Institute

Abstract. Pre-trained on extensive and diverse multi-modal datasets. 2D foundation models excel at addressing 2D tasks with little or no downstream supervision, owing to their robust representations. The emergence of 2D-to-3D distillation frameworks has extended these capabilities to 3D models. However, distilling 3D representations for autonomous driving datasets presents challenges like self-similarity, class imbalance, and point cloud sparsity, hindering the effectiveness of contrastive distillation, especially in zero-shot learning contexts. Whereas other methodologies, such as similarity-based distillation, enhance zero-shot performance, they tend to yield less discriminative representations, diminishing few-shot performance. We investigate the gap in structure between the 2D and the 3D representations that result from state-of-the-art distillation frameworks and reveal a significant mismatch between the two. Additionally, we demonstrate that the observed structural gap is negatively correlated with the efficacy of the distilled representations on zero-shot and few-shot 3D semantic segmentation. To bridge this gap, we propose a relational distillation framework enforcing intra-modal and cross-modal constraints, resulting in distilled 3D representations that closely capture the structure of the 2D representation. This alignment significantly enhances 3D representation performance over those learned through contrastive distillation in zero-shot segmentation tasks. Furthermore, our relational loss consistently improves the quality of 3D representations in both in-distribution and out-of-distribution few-shot segmentation tasks, outperforming approaches that rely on the similarity loss.

1 Introduction

Understanding 3D scenes is pivotal for robotics applications 19,35, notably in autonomous driving, where accurate navigation and decision-making depend on precise environmental perception. Solving the perception tasks necessary for 3D scene understanding requires the point-wise labelling of LiDAR scenes, which is tedious, compounded by the sparsity of LiDAR data, and costly 28,29. These issues result in a scarcity of large-scale, diverse point cloud datasets, particularly those aligned with images or text, significantly hampering the development of foundation models for 3D tasks. This shortage is particularly problematic for few-shot or zero-shot learning approaches, which aim to achieve proficiency with



Fig. 1: We distill 2D representations from CLIP 36 to a 3D point-cloud encoder using the contrastive loss, similarity loss, and our proposed relational loss, and compute the uniformity (U), tolerance (T), and modality gap (G) of the learned 3D representations. We sample 5000 point features from each of the 16 classes defined in the nuScenes dataset 5, apply PCA and visualize the primary components. The source U and T of the CLIP image encoder are 1.54 and 0.73, respectively. Compared to the source, we see that contrastive loss learns 3D representations with higher U and lower T compared to the source, while the trends are reversed for similarity loss. Our proposed relational loss minimizes this structural mismatch and leads to the lowest modality gap.

minimal or no labelled examples [4,32]. Bridging this gap is vital for creating models adept at understanding complex 3D scenes with limited data.

Contrary to the 3D domain, the 2D and language domains benefit from the availability of large-scale, diverse, and multi-modal datasets, which facilitated the development of Vision Foundation Models (VFMs) 7.30, and Vision-Language Models (VLMs) 36,46. These models have shown remarkable label efficiency in 2D tasks like image classification and segmentation and have also been used to perform label-free 2D image classification and segmentation through language prompts during inference 26,36. Due to these advantages, recent approaches for learning 3D representations have relied on the distillation of 2D representations from VFMs 25,27,38 or VLMs 9,33,44 to point cloud encoders, and have shown encouraging results when solving few-shot and zero-shot 3D tasks.

Unfortunately, the adaptation of 2D-to-3D distillation frameworks for autonomous driving (AD) datasets reveals unique challenges, notably due to the inherent characteristics of AD data. Current frameworks, largely reliant on contrastive learning methods, face the issue of self-similarity [27], a prevalent phenomenon in AD datasets. Self-similarity arises when a significant portion of the training examples belong to a single semantic category (e.g. road, trees or sky in AD data). Under the effect of self-similarity, the contrastive loss mechanism, designed to be hardness-aware [41], inadvertently pushes away semantically similar samples, which not only disrupts the local semantic coherence of the 3D representation [24, 27, 38] but also amplifies the effects of AD datasets' severe class imbalance. As a result, while such frameworks may result in useful 3D representations for few-shot learning tasks, the misalignment induced by excessive pushing of semantically similar examples undermines their efficacy in zero-shot learning scenarios, where precise cross-modal alignment is crucial. Another approach

to distillation relies on the cosine similarity loss, which attempts to learn a 3D representation by driving the features generated for every 3D point by a point encoder to its corresponding 2D feature from VFMs or VLMs. Using the similarity loss results in 3D representations that achieve significantly better performance on the zero-shot tasks compared to ones learned with the contrastive loss. However, we argue in Section 3 that using the similarity loss under-constrains the pretraining when compared to the contrastive loss, resulting in sub-optimal 3D representations on few-shot downstream tasks.

To highlight the mismatch discussed above, we provide an example visualization of the 3D representation space (Fig. 1), which was obtained by distilling from CLIP's 36 2D representation space using both the contrastive and similarity losses. We observe that both losses result in 3D representations that diverge from the structure of the 2D representation as measured by uniformity, tolerance, and the modality gap, which we further explain in Section 3.

In this work, we investigate the impact of state-of-the-art 2D-to-3D distillation frameworks on the structure of learned 3D representations. We show that the choice of loss during pretraining can result in a significant mismatch between the structure of the 2D source representations and the distilled 3D representations. Furthermore, we demonstrate that this mismatch leads to a deterioration in performance on downstream tasks. Our contributions are as follows:

- Quantify the Gap: We quantify the mismatch in structure when performing distillation using contrastive loss 12,25,38 and similarity loss 17,33 via the uniformity 41,42, tolerance 41, and modality gap 23, revealing a significant gap between 2D and distilled 3D representations.
- Bridging the Gap using Relational Distillation: We address this mismatch by imposing structural constraints that foster the learning of a 3D representation aligned with the structure of 2D representations. To achieve this, we employ pretraining with intra-modal and cross-modal relational losses. These losses generalize the similarity loss, providing a more effective constraint on the distillation process. Our proposed losses can be applied to pixel-based [25] and superpixel-based [38] distillation frameworks.
- Bridging the Gap Improves Downstream Performance: Our proposed loss effectively minimizes the mismatch between learned 3D and 2D representations from multiple VLM and VFM image encoders, quantified by differences in the U, T, and G (Fig. 1). Consequently, the resulting 3D representations significantly outperform those learned via contrastive distillation on zero-shot segmentation tasks. Furthermore, compared to the similarity loss, our relational loss results in 3D representations that consistently improve in-distribution and out-of-distribution few-shot segmentation tasks.

2 Related Work

2.1 Cross-Modal Knowledge Distillation

Knowledge Distillation enables a student network to learn a task by mimicking the output of highly-performing teacher networks [13], achieved by using

these output as training targets for the student network 18, or by supervising the student network using the teacher's intermediate representations 37,40,43. Distilling representations rather than network predictions enables knowledge transfer across different modalities, without requiring labels. In the context of cross-modal distillation, contrastive distillation 12,40 has shown to be the most effective. However, the abundance of self-similarity 27, coupled with the hardness-aware property of contrastive losses 41 limits the effectiveness of contrastive losses for AD data. On the other hand, similarity losses only utilize positive pairs for distillation, resulting in an under-constrained loss, leading to sub-optimal performance on zero-shot tasks 12.

2.2 3D Representations for Few-Shot Learning

Few-shot learning refers to learning the underlying pattern in data from only a few training samples 32. Distilling 3D representations from 2D self-supervised models 2,6,7,10,14,15 or VFMs 30,36 has shown to be effective at significantly improving performance on 3D few-shot tasks 24,27,38. PPKT 25 proposes a pixel-based contrastive loss to distill 2D self-supervised representations to point cloud encoders. While effective in indoor settings, PPKT [25] underperforms in outdoor settings where point-to-pixel correspondences are sparse 38. Inspired by DetCon 16, SLidR 38 proposes a superpixel-based contrastive loss primarily designed for autonomous driving scenes which constructs region-level contrastive pairs suited for distilling scene-level images [16]. Due to the abundance of selfsimilarity in AD data 24, 27, 38, contrastive distillation leads to sub-optimal 3D representations 27. To address these challenges, ST-SLidR 27 proposes a semantically tolerant contrastive loss leading to improved 3D representations for 3D few-shot segmentation tasks. Finally, Seal 24 demonstrates that object priors from VFMs like SAM [21], represented as superpixels, can improve the quality of 3D representations for 3D few-shot segmentation tasks. In this paper, we demonstrate that pixel and superpixel-based contrastive distillation applied to AD data, learn 3D representations that significantly differ from the structure of 2D representations. This leads to poor performance on zero-shot tasks and unpredictable performance on few-shot tasks.

2.3 3D Representations for Zero-Shot Learning

Contrastive Language-Image Pre-training (CLIP) models are pre-trained on billions of webscale image-text pairs and have shown great success in solving zeroshot image classification tasks 36. Using frozen CLIP vision encoders, Image-Bind 12 enables zero-shot image classification by distilling image-level representations via a contrastive distillation framework. LidarCLIP 17 aligns LiDAR point features to CLIP space, demonstrating effective cross-modal retrieval and image-level zero-shot classification. PointCLIP 44 and PointCLIPv2 47 propose a distillation-free approach utilizing CLIP vision and text encoders during the inference stage to solve 3D zero-shot classification tasks. More recently, MaskCLIP 46 proposes removing the last attention pooling layer in CLIP vision encoders to enable dense feature extraction for 2D zero-shot segmentation tasks. OpenScene 33 distills image features from CLIP models fine-tuned on 2D segmentation datasets 11,22 to point cloud encoders. As demonstrated by ConceptFusion 19 and LERF 20, due to fine-tuning on closed-set 2D segmentation labels, these CLIP models have poor open-set capabilities. During inference, OpenScene fuses image and point features to solve 3D zero-shot segmentation. In this work, we do not assume access to finetuned CLIP models or access to image data during inference time. CLIP2Scene 9 assumes knowledge of class names of objects in the pre-training dataset, which leads to a contrastive loss with fewer false negatives. Requiring class names is problematic as it assumes a dataset only consists of a predefined set of classes, preventing the transfer of features associated with undefined classes and limiting open-set capabilities 34.

3 Methodology

3.1 Preliminaries

Our objective is to generate useful 3D representations by learning a point cloud encoder, $f : \mathbb{R}^{N \times (3+L)} \to \mathbb{R}^{N \times C}$, through distilling 2D representations from VFMs or VLMs (e.g., CLIP **36**], DINOv2 **30**]. Using camera-to-LiDAR calibration matrices, we create a set of positive pairs of point, $\mathbf{K}_{\mathbf{p}} \in \mathbb{R}^{N \times C}$, and pixel, $\mathbf{Q}_{\mathbf{p}} \in \mathbb{R}^{N \times C}$, features, with the latter being generated from the pre-trained image encoder of the foundation model, $g : \mathbb{R}^{N \times 3} \to \mathbb{R}^{N \times C}$. While distillation methods like PPKT **25** implement distillation losses by creating pixel-point positive pairs, techniques such as SLidR **38** optimize pair formation by harnessing boundary information from superpixels. SLidR, with M superpixels per image, employs average pooling to group points, $\mathbf{K}_{\mathbf{p}}$, and pixels, $\mathbf{Q}_{\mathbf{p}}$, into superpoint, $\mathbf{K}_{\mathbf{sp}} \in \mathbb{R}^{M \times C}$, and superpixel, $\mathbf{Q}_{\mathbf{sp}} \in \mathbb{R}^{M \times C}$, features, improving feature correspondence. Superpixels are derived from unsupervised techniques like SLIC **1**, or foundation models such as SAM **2124**. One of the most effective cross-modal distillation losses is the contrastive loss:

$$\mathcal{L}_{con}\left(\mathbf{K},\mathbf{Q}\right) = -\frac{1}{B} \sum_{i=1}^{B} \log \left[\frac{e^{\left(\langle \mathbf{k}_{i},\mathbf{q}_{i} \rangle/\tau\right)}}{\sum_{j \neq i} e^{\left(\langle \mathbf{k}_{i},\mathbf{q}_{j} \rangle/\tau\right)} + e^{\left(\langle \mathbf{k}_{i},\mathbf{q}_{i} \rangle/\tau\right)}} \right]$$
(1)

where **K** and **Q** can either be point/pixel (i.e., $\mathbf{K}_{\mathbf{p}}$, $\mathbf{Q}_{\mathbf{p}}$) or superpoint/superpixel (i.e., $\mathbf{K}_{\mathbf{sp}}$, $\mathbf{Q}_{\mathbf{sp}}$) feature vectors, *B* is the number of positive pairs in a mini-batch, τ is the temperature, and $\langle \mathbf{k}_i, \mathbf{q}_j \rangle$ is the dot product between the ℓ_2 -normalized features. The contrastive loss distills information from pre-trained image encoders by pulling the point cloud features, *K*, towards their corresponding (positive) image feature, *Q*, in representation space, simultaneously pushing them away from all the other (negative) image features. The temperature, τ , controls the strength of this push/pull mechanism by modifying the gradient's scale from the negative samples [41]. However, a significant limitation of this approach is the potential degradation in the learned point cloud representation's

quality due to false negative samples 27. These are image features incorrectly chosen as negative, despite belonging to the same semantic class as the positive point cloud feature, leading to opposing distillation signals.

To avoid both relying on negative examples while performing distillation, and tuning the temperature parameter, we can use the cosine similarity:

$$\mathcal{L}_{sim}\left(\mathbf{K},\mathbf{Q}\right) = \frac{1}{B} \sum_{i=1}^{B} \left(1.0 - \langle \mathbf{k}_{i}, \mathbf{q}_{i} \rangle\right)$$
(2)

The similarity loss does not rely on negative samples and thus has a much simpler mode of action. It focuses solely on drawing each point cloud feature, \mathbf{k}_i , nearer to its corresponding image feature, \mathbf{q}_i , effectively maximizing their dot product.

3.2 Quantifying the Quality of Distilled Representations

We investigate the quality of the distilled 3D representations as a function of the distillation loss used during training. We hypothesize that if a distilled 3D representation space closely captures the structure of the source 2D representation space, our resulting 3D encoder would: 1) possess the enhanced representation capability of the source vision/vision-language foundation model and 2) generate 3D representations that are well-aligned with the text representations of the vision-language model, enabling zero-shot downstream tasks. To assess the structural similarity between two representation spaces, we utilize uniformity and tolerance, two metrics previously proposed by Wang et al. [41], for evaluating the quality of a representation space. Uniformity measures the distribution of ℓ_2 -normalized features on a hyper-sphere. Authors in [42] have demonstrated that a high uniformity is key for high-quality representations as it quantifies that the trained encoder has successfully utilized a substantial part of the available feature space. Uniformity is formulated using a Gaussian potential function as:

$$\mathbf{U}(f(.)) = -\log \mathbb{E}_{x,y \sim p_{\text{data}}} \left[e^{-t ||f(x) - f(y)||_2^2} \right]$$
(3)

where f(.) is the point encoder and x, y are two samples from the data-distribution p_{data} . Here, x, y are point features for 3D point encoders. Similarly, we compute the uniformity of U(g(.)) for the image encoder by setting x, y as pixel features.

On the other hand, tolerance measures the semantic clustering for the ℓ_2 -normalized output representations from a given encoder, computed as:

$$T(f(.)) = \mathop{\mathbb{E}}_{x,y \sim p_{\text{data}}} \left[\left(f(x)^T f(y) \right) \cdot I_{l(x) = l(y)} \right]$$
(4)

where l(x) represents the class label of point x. $I_{l(x)=l(y)}$ is an indicator function, having the value of 1 for l(x) = l(y) and the value of 0 for $l(x) \neq l(y)$. A higher tolerance indicates that the features of all points belonging to a certain class are better clustered together on a unit sphere. Similar to uniformity, we can compute

7



Fig. 2: Blue: The source representation space, with a uniformity and tolerance of U=0.89 and T=0.66, respectively. Cyan: The predicted representation space. Here, t denotes the number of training iterations.

tolerance for the image encoder as T(g(.)), by propagating 3D point-wise labels to 2D image pixels.

Furthermore, to directly compare the alignment of the two representation spaces from different modalities, we also use the modality gap as proposed in 23:

$$G(f(.),g(.)) = \mathop{\mathbb{E}}_{x \sim p_{2\mathrm{D}}} [f(x)] - \mathop{\mathbb{E}}_{y \sim p_{3\mathrm{D}}} [g(y)]$$
(5)

The modality gap measures the difference between the mean of the representation of the point features and their corresponding pixel features. We conclude by noting that unlike G, U and T are properties of a single encoder; we show the difference in representations by comparing the values of U(f(.)) and T(f(.)) to U(g(.)) and T(g(.)), respectively. Closer values indicate our point cloud encoder well-approximates the representation space of the source image encoder, which we show in Section 4 to be beneficial for downstream tasks.

3.3 The Representations of Common Distillation Losses

Inspired by 39, we explore the structure of the learned representation space using contrastive loss and similarity loss through a toy example. We start with 1000 uniformly distributed points over a 3D unit sphere, representing point features before the distillation phase. Using a 2-layer MLP, we learn to align each input point with its corresponding pixel feature from the source representation space, defined by U and T levels of 0.89 and 0.66, respectively. We use the Adam optimizer with a learning rate of 10^{-4} and train our model for 50,000 iterations (refer to Appendix A for detailed analysis).

Figure 2 shows a visualization of the source representation space (blue) and the predicted representation space (cyan) at various stages of training. The top row highlights a significant difference: the contrastive losses generate a representation space with a high U of 1.38, but with a very low T of 0.02, differing significantly from the source space's U (0.89) and T (0.66). This is attributed

to self-similarity [27], a phenomenon that occurs when points from the source representation space are close to one another, as in our toy example. Although the contrastive loss moves a particular prediction to its corresponding point in the source, its hardness-awareness property [41] pushes all other predictions, particularly close ones, away from that point. Combined with self-similarity, this yields a uniform predicted space with a low T, and a significant modality gap. As demonstrated in Section [4], this discrepancy hinders zero-shot task performance due to inadequate cross-modal alignment.

On the other hand, learning with similarity loss leads to representations with lower U (0.59) and higher T (0.80) compared to the source. This stems from the tendency of similarity loss to form compact clusters in the predicted space, as shown with autonomous driving data in Figure 1 and the toy example in Figure 2. Notably, early training (t=1000) exhibits noticeable clustering that disperses with additional epochs. Additionally, the non-uniqueness of the dot product between the learnable point vector, \mathbf{k}_i , and the fixed pixel vector, \mathbf{q}_i , within the framework of similarity loss (Eq. 2) results in a significant increase in the number of learnable vectors that can achieve the same cosine similarity with the fixed vector, \mathbf{q}_i . This growth can result in slow convergence and suboptimal predicted representation spaces. Combined with neural network optimization complexities, these issues lead to the observed disparity in uniformity and tolerance from the source representation space in distillation tasks using similarity loss.

3.4 Relational Loss

Inspired by distilling relations in model compression [31], this section presents two relational losses that impose structural constraints on the 3D representation space. This modification disrupts the symmetry inherent in the similarity loss. It drives the network to select solutions that, while possessing equivalent similarity loss values to alternatives, yield a 3D representation space that more accurately mirrors the structure of the image representation space.

Cross-modal Relational Loss We propose imposing a constraint on the structure of the learned 3D representation space to ensure that the similarities between a given predicted point feature, \mathbf{k}_i , and all source pixel features align with the similarities between its corresponding pixel feature, \mathbf{q}_i , and all other source pixel features in the same batch. The cross-modal relation loss is defined as:

$$\mathcal{L}_{cross}\left(\mathbf{K},\mathbf{Q}\right) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{C}_{ii} + \frac{1}{N^{2} - N} \sum_{i=1}^{N} \sum_{\substack{j=1\\i \neq j}}^{N} \mathbf{C}_{ij}$$
$$= \underbrace{\frac{1}{N} \sum_{i=1}^{N} \left(1.0 - \langle \mathbf{k}_{i}, \mathbf{q}_{i} \rangle\right)}_{\mathcal{L}_{sim}(\mathbf{K},\mathbf{Q})} + \frac{1}{N^{2} - N} \sum_{i=1}^{N} \sum_{\substack{j=1\\i \neq j}}^{N} \left| \langle \mathbf{k}_{i}, \mathbf{q}_{j} \rangle - \langle \mathbf{q}_{i}, \mathbf{q}_{j} \rangle \right|$$
(6)

where $\mathbf{C} = |\mathbf{K}\mathbf{Q}^T - \mathbf{Q}\mathbf{Q}^T|$ represents the matrix of differences, capturing the discrepancy between the predicted point-to-pixel and the source pixel-to-pixel similarities. First, the matrix's diagonal components, \mathbf{C}_{ii} , are directly linked to the similarity loss, underscoring the necessity of aligning each point feature with its corresponding source pixel feature. The second term of Equation 6 facilitates learning a point feature, \mathbf{k}_i , that is not only close to its corresponding source pixel feature, \mathbf{q}_i , but is also consistent with \mathbf{q}_i 's similarities to other pixel features, $\langle \mathbf{q}_i, \mathbf{q}_j \rangle$, where $j \neq i$. These cross-modal constraints foster learning point features that maintain the relational structure within the image representation space.

Intra-modal Relational Loss Another strategy to ensure structural similarity between the 3D and 2D representation spaces involves directly penalizing differences in their relational graphs. A relational graph of a representation space can be understood as a graph with the nodes representing point or pixel features, and edges indicating the similarity between node features. We represent the relational graph of the point features and the pixel features using \mathbf{KK}^T and \mathbf{QQ}^T , respectively. Here, $\langle \mathbf{k}_i, \mathbf{k}_j \rangle$ represents the similarity between the i^{th} and the j^{th} predicted point feature. We denote the discrepancy of the relational graphs as $\mathbf{U} = |\mathbf{KK}^T - \mathbf{QQ}^T|$. Since this matrix is symmetric, and its diagonal elements, $|\langle \mathbf{k}_i, \mathbf{k}_i \rangle - \langle \mathbf{q}_i, \mathbf{q}_i \rangle|$, degenerate to 0, the intra-modal relational loss is expressed as:

$$\mathcal{L}_{intra}\left(\mathbf{K},\mathbf{Q}\right) = \frac{2}{N^2 - N} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{U}_{ij}$$
(7)

The consequence of this loss on the structure of the point representation space is simple; the similarity between a predicted point feature and other predicted point features in a batch should match the similarity between its corresponding source pixel feature and other source pixel features in the same batch. Our proposed relational loss is a combination of the two losses:

$$\mathcal{L}_{rel}\left(\mathbf{K}, \mathbf{Q}\right) = \mathcal{L}_{intra}\left(\mathbf{K}, \mathbf{Q}\right) + \mathcal{L}_{cross}\left(\mathbf{K}, \mathbf{Q}\right) \tag{8}$$

Unlike the similarity loss, the proposed loss ensures learned point cloud features align with image features in a structured manner, preserving the data's inherent relationships. Figure 2 reveals two key advantages of relational loss over similarity loss: closer alignment of the predicted space's U and T with those of the source space, alongside a reduced G between the two spaces. Additionally, the relational loss achieves faster convergence, as the visual appearance of the predicted representation space resulting from learning with the relational loss remains consistent from the mid-point of training (t=25000) to its conclusion (t=50000). In Section 4, we further illustrate how these attributes of the relational loss contribute to enhanced performance in downstream tasks.

4 Experiments

4.1 Pre-training

Backbones We focus on distilling 2D representations from two vision foundation models, CLIP <u>36</u>, pretrained on the WIT dataset containing 400 million image-text pairs, and DINOv2 <u>30</u>, pretrained on the LVD-142 dataset containing 142 million images. For CLIP encoders, we experiment with different pre-trained architectures available from OpenAI <u>36</u>. For the 3D backbone, we use a randomly initialized Minkowski U-Net <u>38</u>. For details on the design of the 2D and 3D projection layers, refer to Appendix C.

Dataset In the pre-training phase we use the nuScenes dataset **5**, consisting of 700 scenes. In line with previous works **24**,**27**,**38**, these scenes are divided into two subsets: 600 scenes for pre-training and 100 scenes for the determination of the best hyper-parameters. Throughout this pre-training process, we exclusively employ keyframes from the 600 scenes to train all models. For pre-training hyperparameters and data-augmentation details, refer to Appendix B.

4.2 Evaluation

To study the effect of the distillation framework on the structure of the 3D representations, we evaluate the average U, T, and G between the distilled 3D and source 2D representations. For both pixel-based and superpixel-based losses, U, T, and G are evaluated using point or pixel features. To assess the relationship between the performance on downstream tasks and the difference in the structure of the 2D and the 3D representation spaces, we use the in-distribution 3D few-shot segmentation task, where we learn a classifier by finetuning the 3D representations on 1% of the labels of the nuScenes training set. Similar to 27,38, we also evaluate the utility of the distilled representations in the out-of-distribution setting by fine-tuning on 1% of the SemanticKITTI 3 dataset. The nuScenes and SemanticKITTI datasets contain 16 and 19 classes, respectively. We present our results using the official validation sets for these datasets.

For DINOv2, we follow 27,38 and evaluate the Linear Probing performance on nuScenes by freezing the backbone of the point cloud encoder and training a linear classifier on 100% of the labels. For CLIP models, we evaluate the 3D zero-shot segmentation tasks which can be performed with language prompts. To enable openset scene understanding, during pretraining, we assume we have no access to nuScenes class labels. Similar to 46, we apply prompt engineering during inference using 85 hand-crafted prompts for each class label, and then use the CLIP text encoder to compute an average text embedding for each class. Each point is then assigned the label that corresponds to the highest cosine similarity, computed between the point features and the CLIP text embeddings.

Table 1: Evaluation of 3D representations from CLIP image encoder with source U of 1.54, and source T of 0.73. Best results are bolded and second best are underlined.

2D	Distillation			Modality	nuScenes		KITTI
Encoder	Loss	Uniformity	Tolerance	Gan	Zero-shot	Finetuning	Finetuning
Lincoder	1033			Gap	Leto shot	1%	1%
CLIP 36	PPKT 25]	3.5210	0.5217	0.00158	14.53	45.31	45.77
	Sim_{pl}	0.5519	0.9477	0.00045	20.84	44.39	45.60
	Rel_{pl} (ours)	0.9089	0.9145	0.00033	23.53	45.67	46.06
	SLidR 38	3.5090	0.4472	0.00153	16.82	46.76	46.53
	ST-SLidR 27	3.5000	0.4544	0.00150	18.54	47.13	46.84
	Sim_{spl}	0.5919	0.9333	0.00040	23.93	45.63	46.42
	Rel_{spl} (ours)	1.0990	0.8700	0.00019	26.03	47.22	47.52

Table 2: Evaluation of 3D representations from DINOv2 image encoder with source U of 2.8, and source T of 0.51. Best results are bolded and second best are underlined.

2D	Distillation			Modality	nuScenes		KITTI
Encoder	Loss	Uniformity	Tolerance	Gan	Lin. Probing	Finetuning	Finetuning
Lincouci	1055			Gup	100%	1%	1%
	PPKT 25	3.625	0.4451	0.00060	48.50	43.30	43.62
	Sim_{pl}	2.176	0.6926	0.00030	49.72	45.85	48.32
	Rel_{pl} (ours)	2.393	0.6588	0.00020	49.20	46.90	49.00
DINOv2 30	SLidR 38	3.655	0.3739	0.00044	49.60	44.81	42.23
	ST-SLidR 27	3.589	0.4326	0.00042	53.00	47.11	45.61
	Sim_{spl}	2.286	0.6522	0.00030	52.30	47.23	49.01
	Rel_{spl} (ours)	2.504	0.6312	0.00023	52.92	48.42	49.10

4.3 Results

Baselines In Tab. 1 and Tab. 2 we present the results of distilling CLIP and DINOv2 models using our proposed relational loss. We compare against multiple state-of-the-art contrastive losses including PPKT [25], SLidR [38], and ST-SLidR [27]. For superpixel-based losses, masks are generated using SAM [21]. As an additional baseline, we report results for both the pixel-based and superpixel-based similarity loss, which we denote as \lim_{pl} and \lim_{spl} , respectively. We report the average over 3 runs for all metrics.

Uniformity, tolerance, and closing the modality gap Looking at distillation losses from the 2D CLIP encoder in Tab. \square , we first observe that U of 3D representations distilled using contrastive loss (i.e., PPKT, SLidR and ST-SLidR) is higher than the source U of 1.54, while U of similarity loss (i.e., Sim_{pl} , Sim_{spl}) is lower than the U of the source. This is also observed when distilling from DINOv2. The relational constraints (i.e., Rel_{pl} , Rel_{spl}) effectively close the gap between 3D representations distilled by similarity loss (i.e., Sim_{pl} , Sim_{spl}) and 2D representations. For instance, when the pixel-based loss Rel_{pl} is applied to CLIP, relational constraints close the gap in U resulting in a 3D representation with a U of 0.9089, instead of 0.5519, thus closer to the U of the source (1.54). Similarly, for the superpixel-based loss, Rel_{spl} , relational constraints close the gap in U resulting in a 3D representation with a U of 1.099, instead of 0.5919, thus closer to the U of the source (1.54). In addition, relational losses result in semantic clusters that are closer to the T of the source. By closing the gap

-TC	Distillation Loss	Uniformity	Tolerance	Modality - Gap	nuScenes		KITTI
Encoder					Zoro shot	Finetuning	Finetuning
					Zero-shot	1%	1%
ViT-B32	PPKT	3.5110	0.5203	0.00020	14.21	43.29	45.89
	Sim_{pl}	0.5804	0.9389	0.00010	22.84	44.22	44.30
	Rel_{pl} (ours)	0.9605	0.8989	0.00010	23.85	44.67	45.62
	SLidR	3.5130	0.4706	0.00030	15.55	43.38	43.82
	Sim_{spl}	0.6009	0.9323	0.00011	25.46	44.83	45.69
	Rel_{spl} (ours)	1.2020	0.8648	0.00007	25.66	45.60	47.26
	PPKT	3.5210	0.5217	0.00158	14.53	45.31	45.77
	Sim_{pl}	0.5519	0.9477	0.00045	20.84	44.39	45.59
VIT B16	Rel_{pl} (ours)	0.9089	0.9145	0.00033	23.53	45.67	46.06
VII-DI0	SLidR	3.5090	0.4472	0.00153	16.82	46.76	46.53
	Sim_{spl}	0.5919	0.9333	0.00040	23.93	45.63	46.42
	Rel_{spl} (ours)	1.0990	0.8700	0.00019	26.03	47.22	47.52
ViT-L14	PPKT	3.5020	0.5203	0.00089	16.28	44.69	46.69
	Sim_{pl}	0.6909	0.9257	0.00016	27.91	44.87	44.80
	Rel_{pl} (ours)	0.8775	0.9027	0.00013	27.74	45.92	45.86
	SLidR	3.4800	0.4403	0.00065	18.31	47.34	46.86
	Sim_{spl}	0.7311	0.9102	0.00020	30.77	45.76	45.29
	Rel_{spl} (ours)	0.9019	0.8873	0.00017	<u>30.11</u>	47.07	46.81

Table 3: Performance for different CLIP backbones using pixel-based and superpixel-based relational loss. Best results are bolded and second best are underlined.

in U and T, relational constraints result in the lowest G for pixel-based and superpixel-based losses. These results align well with the ones presented on the toy example in Section 3 which further supports the validity of our analysis.

Zero-shot performance Looking at the utility of the 3D representations for 3D zero-shot segmentation in Tab. 1, we observe that the contrastive loss (i.e., PPKT, SLidR and ST-SLidR) has significantly worse 3D zero-shot mean IOU when compared to the similarity loss. This is contrary to the conclusion reached by ImageBind 12, which observed that the contrastive loss achieves 6% improvement on 2D zero-shot tasks compared to similarity losses. We hypothesize that the abundance of self-similarity in AD datasets, coupled with the hardnessaware property of contrastive loss [27], leads to pushing away semantically similar point and pixel features. This, in turn, results in poor alignment between point features and CLIP image features. Notably, we observe that ST-SLidR improves zero-shot performance compared to SLidR (18.54% vs 16.82%) as it utilizes the superpixel feature similarities to exclude a portion of the false negative samples from the pool of negative samples. Without resorting to negative samples, relational constraints minimize the 2D-to-3D structural gap compared to similarity loss, leading to improvements in zero-shot performance from 20.84% to 23.53%for pixel-based losses, and from 23.93% to 26.03% for superpixel-based losses.

In Tab. 4, we compare pixel and superpixel-based relational losses to SOTA methods. All methods distill from CLIP [36], with the performance on nuScenes provided by [8]. Methods are further categorized based on the prior knowledge required during the distillation phase. Some methods require class names defined for a dataset [8, 9], while others utilize SAM to refine CLIP predictions [8]. Requiring class names assumes a dataset only consists of a predefined set of classes, preventing the transfer of features associated with undefined classes, thus

Table 4: Zero-shot segmentation per-formance of relational loss compared tostate-of-the-art methods.

Method	Publication	Class Names Required	Uses SAM	3D mIoU
MaskCLIP 46	ECCV 2022	×	X	12.80
OpenScene 33	CVPR 2023	×	X	14.60
CLIP2Scene 9	CVPR 2023	1	X	20.80
Rel_{pl} (ours)	-	×	X	23.53
TLF 8]	NeurIPS 2023	1	1	26.80
Rel_{spl} (ours)	-	×	1	26.03

Table 5: Effect of relational loss components. Distilling from CLIP image encoder with source U of 1.54 and T of 0.73.

			Modelity	nuScenes		KITTI
Loss	Uniformity	Tolerance	Can	75	Ft	Ft
			Gap	20	1%	1%
Sim_{pl}	0.5519	0.9477	0.00045	20.84	44.39	45.59
+ cross	0.6426	0.9386	0.00042	21.39	45.95	45.45
+intra	0.9089	0.9145	0.00033	23.53	45.67	46.06
Sim_{spl}	0.5919	0.9333	0.00040	23.93	45.63	46.42
+ cross	0.7656	0.9086	0.00033	25.05	46.18	46.99
+intra	1.0990	0.8700	0.00019	26.03	47.22	47.52

Table 6: Improvement of finetuned models with respect to majority versus minority classes. Similar to ST-SLidR [27], We group classes based on whether their superpixels occupy more than 5% of the superpixels in nuScenes training set.

2D Encoder	Loss	Majority (mIoU)	Minority (mIoU)	
	Sim_{pl}	69.16	33.35	
	Rel_{pl}	69.59	34.80	
CUIP	Gain	+0.43	+1.45	
OLII	Sim_{spl}	67.58	36.06	
	Rel_{spl}	67.75	38.77	
	Gain	+0.17	+2.71	
	Sim_{pl}	70.89	34.24	
	Rel_{pl}	71.22	35.85	
DINOv9	Gain	+0.33	+1.61	
DINOVZ	Sim_{spl}	68.15	38.53	
	Rel_{spl}	68.80	40.20	
	Gain	+0.65	+1.67	

limiting openset capabilities 34. Without using class information, we observe in Tab. 4 that Rel_{pl} performs the best compared to other pixel-based losses, while our superpixel-based relational loss, Rel_{spl} , is within 1% from state-of-the-art 8. Few-shot performance We evaluate the effectiveness of 3D representations in both in-distribution (namely, nuScenes) and out-of-distribution (namely, SemanticKITTI) settings. For representations distilled from CLIP in Tab. 1, we observe that the limited U of representations distilled through similarity losses, as opposed to the source, leads to comparatively weaker performance on indistribution tasks with few-shot learning when compared with contrastive loss. For instance, ST-SLidR representations enhance performance on the nuScenes dataset by +1.5% compared to Sim_{spl} . Furthermore, by applying relational constraints and without the need for negative samples, we can bridge the U gap. This results in Rel_{spl} outperforming Sim_{spl} by +1.59% and +1.1% on nuScenes and SemKITTI, respectively. When examining 3D representations distilled from DI-NOv2 in Tab. 2, we note a source U of 2.8. Here, the similarity loss closely aligns with the source's U more than the contrastive loss. The contrastive loss (PPKT and SLidR) shows significantly poorer performance compared to the similarity loss. Additionally, 3D representations distilled by ST-SLidR attain a U of 3.589 and a T of 0.4326, aligning more closely with the source U and T of 2.8 and 0.51 than those distilled by the SLidR loss. Lastly, we find that relational constraints narrow the G more effectively than similarity losses, thereby enhancing the fewshot performance by +1.05% and +1.19% on pixel and superpixel-based losses, respectively, on the nuScenes dataset compared to similarity losses.

CLIP Backbones Tab. 3 depicts the performance of contrastive, similarity,

and relational losses when distilling from different CLIP backbones. Relational losses achieve the lowest modality gap, surpassing both contrastive and similarity loss across all 2D encoders. In zero-shot tasks, relational losses either match or exceed the performance of similarity losses, whereas contrastive distillation consistently underperforms. The performance of contrastive distillation on few-shot tasks proves unpredictable. For instance, distilling from ViT-B32 using similarity loss Sim_spl surpasses distilling using contrastive loss SLidR in mIOU by +1.45% and +1.87% on nuScenes and SemanticKITTI, respectively. Conversely, distilling from ViT-L14, SLidR achieves competitive performance, outperforming similarity loss Sim_{spl} by +1.58% and +1.57% on the same datasets. Notably, relational losses bridge the gap in few-shot learning while maintaining robust zero-shot task performance. This indicates that relational losses utilize inherent 2D representation relationships, avoiding ungrounded negative samples common in contrastive loss, which facilitates learning representations more aligned with the source's U, thereby enhancing few-shot task performance.

Class Imbalance We investigate the performance of finetuned models on semantic segmentation using 3D representations distilled using similarity and relational losses. Similar to ST-SLidR 27, we distinguish between majority and minority classes based on the percentage of superpixels they occupy in the nuScenes dataset. The 11 classes occupying less than 5% of the superpixels are considered to be in the minority set, while the remaining classes are in the majority set. In Tab. 6, compared to similarity losses, relational losses learn representations that significantly improve performance on minority classes for pixel and superpixel-based losses without degrading performance on majority classes.

Relational Loss Ablation We investigate the contribution of cross-modal and intra-modal constraints for pixel-based and superpixel-based losses. In Tab. ⁵, we observe that both constraints lead to learning 3D representations that are closer to the source U and T, and thus result in a lower G compared to similarity losses. Moreover, both constraints lead to improved performance on zero-shot and few-shot segmentation tasks. Interestingly, the superpixel-based relational loss with both constraints (last row) results in the smallest gap in U and T compared to other losses, leading to the best performance on semantic segmentation.

5 Conclusion

In this work, we study the impact of the state-of-the-art 2D-to-3D distillation frameworks when applied to AD datasets on the structure of the learned 3D representation. We reveal a significant structural gap between the 2D and the 3D representations and show that this gap is negatively correlated with the utility of the learned 3D representations for solving 3D zero-shot and few-shot segmentation tasks. Our proposed relational loss bridges this structural gap, resulting in well-aligned 3D representations that outperform representations learned via contrastive loss on zero-shot segmentation tasks. In addition, compared to the similarity loss, our relational loss results in 3D representations that consistently improve in-distribution and out-of-distribution few-shot segmentation tasks.

References

- Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE TPAMI 34(11), 2274–2282 (2012)
- Bardes, A., Ponce, J., Lecun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In: ICLR (2022)
- Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: SemanticKITTI: A Dataset for Semantic Scene Understanding of LiDAR Sequences. In: ICCV (2019)
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901 (2020)
- Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR. pp. 11621–11631 (2020)
- Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. NeurIPS 33, 9912– 9924 (2020)
- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650– 9660 (2021)
- Chen, R., Liu, Y., Kong, L., Chen, N., Xinge, Z., Ma, Y., Liu, T., Wang, W.: Towards label-free scene understanding by vision foundation models. In: Thirtyseventh Conference on Neural Information Processing Systems (2023)
- Chen, R., Liu, Y., Kong, L., Zhu, X., Ma, Y., Li, Y., Hou, Y., Qiao, Y., Wang, W.: Clip2scene: Towards label-efficient 3d scene understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7020–7030 (2023)
- Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
- Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European Conference on Computer Vision. pp. 540–557. Springer (2022)
- Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180– 15190 (2023)
- Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision 129, 1789–1819 (2021)
- 14. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020)
- He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
- Hénaff, O.J., Koppula, S., Alayrac, J.B., Van den Oord, A., Vinyals, O., Carreira, J.: Efficient visual pretraining with contrastive detection. In: ICCV. pp. 10086– 10096 (2021)

- 16 A. Mahmoud et al.
- Hess, G., Tonderski, A., Petersson, C., Åström, K., Svensson, L.: Lidarclip or: How i learned to talk to point clouds. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7438–7447 (January 2024)
- Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
- Jatavallabhula, K., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Li, S., Iyer, G., Saryazdi, S., Keetha, N., Tewari, A., Tenenbaum, J., de Melo, C., Krishna, M., Paull, L., Shkurti, F., Torralba, A.: Conceptfusion: Open-set multimodal 3d mapping. Robotics Science and Systems (2023)
- Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19729–19739 (2023)
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
- 22. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: ICLR (2022)
- Liang, V.W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.Y.: Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. Advances in Neural Information Processing Systems 35, 17612–17625 (2022)
- Liu, Y., Kong, L., Cen, J., Chen, R., Zhang, W., Pan, L., Chen, K., Liu, Z.: Segment any point cloud sequences by distilling vision foundation models. In: Advances in Neural Information Processing Systems (2023)
- Liu, Y.C., Huang, Y.K., Chiang, H.Y., Su, H.T., Liu, Z.Y., Chen, C.T., Tseng, C.Y., Hsu, W.H.: Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. arXiv preprint arXiv:2104.04687 (2021)
- Mahmoud, A., Elhoushi, M., Abbas, A., Yang, Y., Ardalani, N., Leather, H., Morcos, A.S.: Sieve: Multimodal dataset pruning using image captioning models. In: CVPR. pp. 22423–22432 (2024)
- Mahmoud, A., Hu, J.S.K., Kuai, T., Harakeh, A., Paull, L., Waslander, S.L.: Selfsupervised image-to-point distillation via semantically tolerant contrastive loss. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7102–7110 (June 2023)
- Mahmoud, A., Hu, J.S., Waslander, S.L.: Dense voxel fusion for 3d object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 663–672 (2023)
- Mahmoud, A., Waslander, S.L.: Sequential fusion via bounding box and motion pointpainting for 3d objection detection. In: 18th Conference on Robots and Vision (CRV) (2021)
- Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
- Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3967–3976 (2019)
- Parnami, A., Lee, M.: Learning from few examples: A summary of approaches to few-shot learning. arXiv preprint arXiv:2203.04291 (2022)

- Peng, S., Genova, K., Jiang, C., Tagliasacchi, A., Pollefeys, M., Funkhouser, T., et al.: Openscene: 3d scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 815–824 (2023)
- Puy, G., Gidaris, S., Boulch, A., Simeoni, O., Sautier, C., Perez, P., Bursuc, A., Marlet, R.: Revisiting the distillation of image representations into point clouds for autonomous driving. arXiv preprint arXiv:2310.17504 (2023)
- Qian, J., Chatrath, V., Yang, J., Servos, J., Schoellig, A.P., Waslander, S.L.: Pocd: Probabilistic object-level change detection and volumetric mapping in semi-static scenes. Robotics: Science and Systems (2022)
- Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
- Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
- Sautier, C., Puy, G., Gidaris, S., Boulch, A., Bursuc, A., Marlet, R.: Image-to-lidar self-supervised distillation for autonomous driving data. In: CVPR. pp. 9891–9901 (2022)
- Shi, P., Welle, M.C., Björkman, M., Kragic, D.: Towards understanding the modality gap in CLIP. In: ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls (2023)
- Tian, Y., Krishnan, D., Isola, P.: Contrastive representation distillation. arXiv preprint arXiv:1910.10699 (2019)
- 41. Wang, F., Liu, H.: Understanding the behaviour of contrastive loss. In: CVPR. pp. 2495–2504 (2021)
- Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020)
- 43. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928 (2016)
- 44. Zhang, R., Guo, Z., Zhang, W., Li, K., Miao, X., Cui, B., Qiao, Y., Gao, P., Li, H.: Pointclip: Point cloud understanding by clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8552–8562 (2022)
- Zhang, Z., Girdhar, R., Joulin, A., Misra, I.: Self-supervised pretraining of 3d features on any point-cloud. In: ICCV. pp. 10252–10263 (October 2021)
- 46. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European Conference on Computer Vision. pp. 696–712. Springer (2022)
- 47. Zhu, X., Zhang, R., He, B., Zeng, Z., Zhang, S., Gao, P.: Pointclip v2: Adapting clip for powerful 3d open-world learning. arXiv preprint arXiv:2211.11682 (2022)