CrossPoint: Self-Supervised Cross-Modal Contrastive Learning for 3D Point Cloud Understanding

Mohamed Afham† Isuru Dissanayake† Dinithi Dissanayake† Amaya Dharmasiri†
Kanchana Thilakarathna‡ Ranga Rodrigo†
†Dept. of Electronic and Telecommunication Engineering, Univeristy of Moratuwa, Sri Lanka
‡The University of Sydney
afhamaflal9@gmail.com

Abstract

*Manual annotation of large-scale point cloud dataset for varying tasks such as 3D object classification, segmentation and detection is often laborious owing to the irregular structure of point clouds. Self-supervised learning, which operates without any human labeling, is a promising approach to address this issue. We observe in the real world that humans are capable of mapping the visual concepts learnt from 2D images to understand the 3D world. Encouraged by this insight, we propose **CrossPoint**, a simple cross-modal contrastive learning approach to learn transferable 3D point cloud representations. It enables a 3D-2D correspondence of objects by maximizing agreement between point clouds and the corresponding rendered 2D image in the invariant space, while encouraging invariance to transformations in the point cloud modality. Our joint training objective combines the feature correspondences within and across modalities, thus ensembles a rich learning signal from both 3D point cloud and 2D image modalities in a self-supervised fashion. Experimental results show that our approach outperforms the previous unsupervised learning methods on a diverse range of downstream tasks including 3D object classification and segmentation. Further, the ablation studies validate the potency of our approach for a better point cloud understanding. Code and pretrained models are available at https://github.com/MohamedAfham/CrossPoint.*

1. Introduction

3D vision, which is critical in applications such as autonomous driving, mixed reality and robotics has drawn extensive attention due its ability to understand the human world. In light of that, there have been plethora of work in 3D vision research problems such as object classification [38,39,55], detection [32] and segmentation [39,49,55]

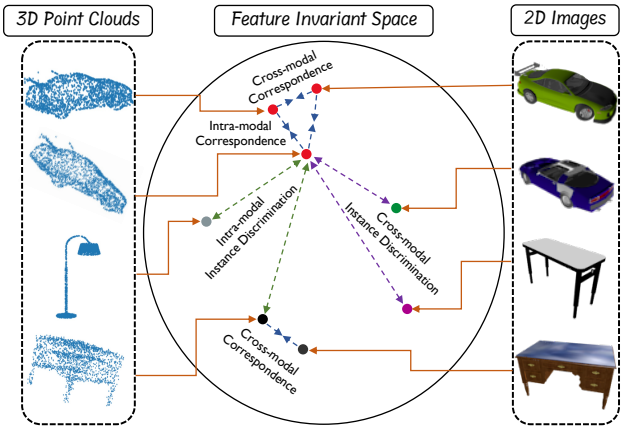

Figure 1. **An illustration of the proposed approach.** Given a 3D point cloud of an object and its rendered 2D image from a random camera view-point, *CrossPoint* enforces 3D-2D correspondence while preserving the model being invariant to affine and spatial transformations via self-supervised contrastive learning. This facilitates generalizable point cloud representations which can then be utilized for 3D object classification and segmentation. Note that the 2D images shown in the right are directly rendered from the available 3D point clouds [63].

in the recent years with point clouds as the most popularly 3D data representation method. However, the success of deep learning crucially relies on large-scale annotated data. Even though the advancements in 3D sensing technology (e.g., LIDAR) facilitates extensive collection of 3D point cloud samples, owing to the irregular structure of point clouds, manually annotating such large-scale 3D point cloud datasets is laborious. Self-supervised learning is one of the predominant approaches to address this issue and is proven to be effective in 2D domain [5,7,13,34].

Several works have explored self-supervised representation learning on point clouds and are mainly based on generative models [1,57], reconstruction [46,53] and other pre-

text tasks [36]. In addition to that, with the success in exploitation of contrastive learning for image [7,13,18,33] and video [20, 40, 54] understanding, recent works have investigated self-supervised contrastive learning for point cloud understanding as well [21,44,59,68]. However, the existing contrastive learning based approaches for point cloud understanding only rely on imposing invariance to augmentations of 3D point clouds. Learning from different modalities *i.e.,* cross-modal learning has produced substantial results in self-supervised learning. Vision + language [9, 41, 45] and video + audio [3, 34, 35] are some notable combinations for multimodal learning. Multi-modal setting has been adopted in various 2D vision tasks such as object detection [24], few-shot image classification [2,60] and visual question and answering [22, 56]. Inspired by the advancements in multimodal learning, we introduce **CrossPoint**, a simple yet effective cross-modal contrastive learning approach for 3D point cloud understanding.

The goal of our work is to capture the correspondence between 3D objects and 2D images to constructively learn transferable point cloud representations. As shown in Fig. 1, we embed the augmented versions of point cloud and the corresponding rendered 2D image close to each other in the feature space. In real world, humans are proficient at mapping the visual concepts learnt from 2D images to understand the 3D world. For example, a person would be able to recognize an object easily, if he/she has observed that object via an image. Cognitive scientists argue that 3D-2D correspondence is a part of visual learning process of children [8, 43]. Similarly, in real world applications such as robotics and autonomous driving, the model being aware of such 3D-2D correspondence will immensely facilitate an effective understanding of the 3D world. Our approach in particular, follows a joint objective of embedding the augmented versions of same point cloud close together in the feature space, while preserving the 3D-2D correspondence between them and the rendered 2D image of the original 3D point cloud.

The joint intra-modal and cross-modal learning objective enforces the model to attain the following desirable attributes: (a) relate the compositional patterns occurring in both point cloud and image modalities e.g., fine-grained part-level attribute of an object; (b) acquire knowledge on spatial and semantic properties of point clouds via imposing invariance to augmentations; and (c) encode the rendered 2D image feature as a centroid to augmented point cloud features thus promoting 3D-2D correspondence agnostic to transformations. Moreover, *CrossPoint* does not require a memory bank for negative sampling similar to SimCLR [7]. Formulation of rich augmentations and hard positive samples have been proved to boost the contrastive learning despite having memory banks [23,72]. We hypothesize that the employed transformations in intra-modal setting and cross-modal correspondence provide adequate feature augmentations. In particular, the rendered 2D image feature acts as a hard positive to formulate a better representation learning.

We validate the generalizability of our approach with multiple downstream tasks. Specifically, we perform shape classification in both synthetic [58] and real world [52] object datasets. Despite being pretrained on a synthetic object dataset [6], the performance of *CrossPoint* in out-of-distribution data certifies the importance of the joint learning objective. In addition, the ablation studies demonstrate the component-wise contribution of both intra-modal and cross-modal objectives. We also adopt multiple widely used point cloud networks as our feature extractors, thus proving the generic nature of our approach.

The main contributions of our approach can be summarized as follows:

- We show that a simple 3D-2D correspondence of objects in the feature space using self-supervised contrastive learning facilitates an effective 3D point cloud understanding.
- We propose a novel end-to-end self-supervised learning objective encapsulating intra-modal and cross-modal loss functions. It encourages the 2D image feature to be embedded close to the corresponding 3D point cloud prototype, thus avoiding bias towards a particular augmentation.
- We extensively evaluate our proposed method across three downstream tasks namely: object classification, few-shot learning and part segmentation on a diverse range of synthetic and real-world datasets, where *CrossPoint* outperforms previous unsupervised learning methods.
- Additionally, we perform few-shot image classification on CIFAR-FS dataset to demonstrate that finetuning the pretrained image backbone from *CrossPoint* outperforms the standard baseline.

2. Related Work

Representation Learning on Point Clouds. Learning point cloud representation is a challenging task when compared to other modalities (e.g., images). This is because of the irregular structure and also the need for permutation invariance when processing each points. Recent line of works pioneered by PointNet [38] proposed methods and architectures which directly consume 3D point cloud without any preprocessing. Since then, numerous advancements have been made in point cloud based tasks such as 3D object classification [27, 29, 39, 49, 55, 62, 70], 3D object detection [27, 32, 37, 69] and 3D point cloud synthesis [1, 57].

However, the performance of such representation learning methods depend on the annotated point cloud data which is often hard to acquire. Sharma *et al*. [47] introduced cTree, where point cloud representations can be learnt in a label efficient scenario (i.e., few-shot learning). In contrast, our approach focuses on learning transferable point cloud representations without leveraging any annotations, which can then be utilized for various downstream tasks such as classification and segmentation.

Self - Supervised Learning on Point Clouds. Several approaches have been explored to perform self-supervised representation learning on point clouds. Initial line of works exploit generative modelling using generative adversarial networks [1, 15, 57] and auto-encoders [11, 16, 26, 65, 71], which aims to reconstruct a given input point cloud with varying architectural designs. Recent line of works [17, 36, 42, 46, 48, 53, 64] introduce various pretext self-supervision tasks with the goal of learning rich semantic point attributes which eventually leads to high-level discriminative knowledge. For instance, Wang *et al*. [53] trains an encoder-decoder model to complete the occluded point clouds, and Poursaeed *et al*. [36] defines estimation of rotation angle of the point cloud as the pretext task. However, in this work we leverage contrastive learning [14] to learn an invariant mapping in the feature space. Inspired by the success of self-supervised contrastive learning for image understanding, numerous works [10,21,28,44,59,67,68] have analyzed such a setting for point cloud understanding. PointContrast [59] performs point-level invariant mapping on two transformed view of the given point cloud. STRL [21], which is a direct extension of BYOL [13] to 3D point clouds, unsupervisedly learns the representations through the interactions of online and target networks. Contrary to the existing works which leverage contrastive learning, we introduce an auxiliary cross-modal contrastive objective which captures 3D-2D correspondence yielding a better representation capability.

Cross-Modal Learning. Learning from different modalities tend to provide rich learning signals from which semantic information of the given context can be addressed easily. Recent works [3, 9, 34, 41, 45, 50] have demonstrated that pretraining in a cross-modal setting produces transferable representations which can then be deployed for various downstream tasks. CLIP [41] aims to learn a multi-modal embedding space by maximizing cosine similarity between image and text modalities. Similarly, Morgado et al. [34] combines audio and video modalities to perform a cross-modal agreement which then achieves significant gains in action recognition and sound recognition tasks. A joint learning approach with point clouds and voxels was introduced by Zhang *et al*. [68]. Further, [61] transfers the pretrained 2D image model to a point-cloud model by filter inflation. Our work is closely related to the concurrent work [30], which uses a fixed image feature extractor to perform pixel-to-point knowledge transfer. In contrast to the existing approaches, CrossPoint is designed in such a way that the 2D image feature is encouraged to be embedded close to the corresponding 3D point cloud prototype while invariance to transformations is imposed in the point cloud modality.

3. Proposed Method

In this work, we revamp the unsupervised 3D point cloud representation learning by introducing a fusion of intra-modal and cross-modal contrastive learning objectives. This section begins by introducing the network architecture details of the proposed method (Sec. 3.1). Then we describe the contrastive learning loss functions formulated in both intra-modal (Sec. 3.2) and cross-modal (Sec. 3.3) settings. Finally, we present our overall training objective (Sec. 3.4). The overview of the proposed method is shown in Fig. 2.

3.1. Preliminaries

Suppose we are given a dataset, $\mathcal{D} = \{(\mathbf{P}_i, \mathbf{I}_i)\}_{i=1}^{|\mathcal{D}|}$ with $\mathbf{P}_i \in \mathbb{R}^{N\times 3}$ and $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$ where $\mathbf{I}_i$ is a rendered 2D image of the 3D point cloud $\mathbf{P}_i$. Note that $\mathbf{I}_i$ is obtained by capturing $\mathbf{P}_i$ from a random camera view-point [6]. We aim to train a point cloud feature extractor $f_{\theta_\mathbf{P}}(.)$ in a self-supervised manner to be effectively transferable to downstream tasks. To this end, we use an image feature extractor $f_{\theta_\mathbf{I}}(.)$, multi-layer perceptron (MLP) projection heads $g_{\phi_\mathbf{P}}(.)$ and $g_{\phi_\mathbf{I}}(.)$ for point cloud and image respectively.

3.2. Intra-Modal Instance Discrimination

Inspired by the success of contrastive pretraining in image modality [7, 18, 33], we formulate our Intra-Modal Instance Discrimination (IMID) to enforce invariance to a set of point cloud geometric transformations $\mathbf{T}$ by performing self-supervised contrastive learning. Given an input 3D point cloud $\mathbf{P}_i$, we construct augmented versions $\mathbf{P}_i^{t_1}$ and $\mathbf{P}_i^{t_2}$ of it. We compose t_1 and t_2 by randomly combining transformations from $\mathbf{T}$ in a sequential manner. We use transformations such as rotation, scaling and translation. In addition to that we also utilize spatial transformations such as jittering, normalization and elastic distortion. Regardless of the augmentation, the corresponding transformation matrix parameters are initialized randomly.

The point cloud feature extractor $f_{\theta_\mathbf{P}}$ maps both $\mathbf{P}_i^{t_1}$ and $\mathbf{P}_i^{t_2}$ to a feature embedding space and the resulted feature vectors are projected to an invariant space $\mathbb{R}^d$ where the contrastive loss is applied, using the projection head $g_{\phi_\mathbf{P}}$. We denote the projected vectors of $\mathbf{P}_i^{t_1}$ and $\mathbf{P}_i^{t_2}$ as $\mathbf{z}_i^{t_1}$ and $\mathbf{z}_i^{t_2}$ respectively where, $\mathbf{z}_i^t = g_{\phi_\mathbf{P}}\left(f_{\theta_\mathbf{P}}\left(\mathbf{P}_i^t\right)\right)$. The goal is to maximize the similarity of $\mathbf{z}_i^{t_1}$ with $\mathbf{z}_i^{t_2}$ while minimizing the similarity with all the other projected vectors in

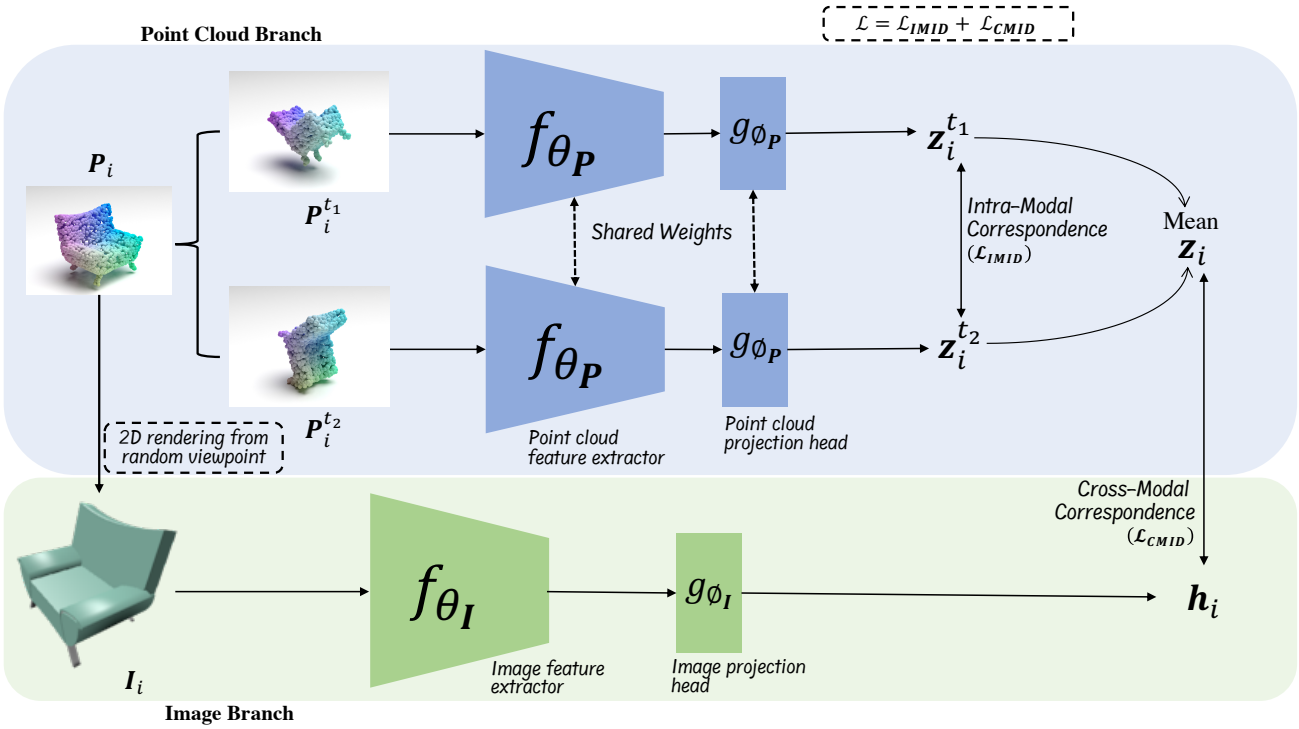

Figure 2. The overall architecture of the proposed method (CrossPoint). It comprises of two branches namely: point cloud branch which establishes an intra-modal correspondence by imposing invariance to point cloud augmentations and image branch which simply formulates a cross-modal correspondence by introducing a contrastive loss between the rendered 2D image feature and the point cloud prototype feature. CrossPoint jointly train the model combining the learning objectives of both the branches. We discard the image branch and use only the point cloud feature extractor as the backbone for the downstream tasks.

the mini-batch of point clouds. We leverage NT-Xent loss proposed in SimCLR [7] for instance discrimination at this stage. Note that our approach doesn't use any memory bank following the recent advancements in self-supervised contrastive learning [5, 13, 21]. We compute the loss function $l(i, t_1, t_2)$ for the positive pair of examples $\mathbf{z}_i^{t_1}$ and $\mathbf{z}_i^{t_2}$ as:

$$l(i,t_1,t_2) = -\log \frac{\exp(s(\mathbf{z}_i^{t_1}, \mathbf{z}_i^{t_2})/\tau)}{\sum_{\substack{k=1 \\ k\neq i}}^{N} \exp(s(\mathbf{z}_i^{t_1}, \mathbf{z}_k^{t_1})/\tau) + \sum_{k=1}^{N} \exp(s(\mathbf{z}_i^{t_1}, \mathbf{z}_k^{t_2})/\tau)} \tag{1}$$

where N is the mini-batch size, τ is the temperature co-efficient and $s(.)$ denotes the cosine similarity function. Our intra-modal instance discrimination loss function $\mathcal{L}_{imid}$ for a mini-batch can be described as:

$$\mathcal{L}_{imid} = \frac{1}{2N} \sum_{i=1}^{N} [l(i,t_1,t_2) + l(i,t_2,t_1)] \tag{2}$$

3.3. Cross-Modal Instance Discrimination

In addition to the feature alignment within point cloud modality, we introduce an auxiliary contrastive objective across point cloud and image modalities to learn discriminative features, thus yielding better representation learning capability of 3D point clouds. As discussed in Sec. 2, several works aim to learn transferable point cloud representations in a cross-modal setting. However, to the best of our knowledge, the joint learning objective of enforcing 3D-2D correspondence, while performing instance discrimination within point cloud modality has not been well explored. We empirically validate with the experimental results in Sec. 4.2 that our joint objective outperforms existing unsupervised representation methods, thus facilitating an effective representation learning of 3D point clouds.

To this end, we first embed the rendered 2D image $\mathbf{I}_i$ of $\mathbf{P}_i$ to a feature space using the *visual backbone* $f_{\theta_\mathbf{I}}$. We opt for the commonly used ResNet [19] architecture as $f_{\theta_\mathbf{I}}$. We then project the feature vectors to the invariant space $\mathbb{R}^d$ using the image projection head $g_{\phi_\mathbf{I}}$. The projected image feature is defined as $\mathbf{h}_i$ where $\mathbf{h}_i = g_{\phi_\mathbf{I}}(f_{\theta_\mathbf{I}}(\mathbf{I}_i))$. In contrast to previous cross-modal approaches [34, 68], we do not explicitly perform IMID on both the modalities (point cloud and image). Instead, we implement IMID on point cloud and leverage image modality for a better point cloud understanding. We propose a learning objective which specifi-

cally induces a bias towards 3D point cloud understanding when compared to image understanding. To this end, we compute the mean of the projected vectors $\mathbf{z}_i^{t_1}$ and $\mathbf{z}_i^{t_2}$ to obtain the projected prototype vector $\mathbf{z}_i$ of $\mathbf{P}_i$.

$$\mathbf{z}_i = \frac{1}{2}\left(\mathbf{z}_i^{t_1} + \mathbf{z}_i^{t_2}\right) \tag{3}$$

In the invariance space, we aim to maximize the similarity of $\mathbf{z}_i$ with $\mathbf{h}_i$ since they both correspond to same objects. Our cross-modal alignment enforces the model to learn from harder positive and negative samples, thus enhances the representation capability than learning only from intra-modal alignment. We compute the loss function $l(i, \mathbf{z}, \mathbf{h})$ for the positive pair of examples $\mathbf{z}_i$ and $\mathbf{h}_i$ as:

$$c(i, \mathbf{z}, \mathbf{h}) = -\log \frac{\exp(s(\mathbf{z}_i, \mathbf{h}_i)/\tau)}{\sum\limits_{\substack{k=1 \\ k\neq i}}^{N} \exp(s(\mathbf{z}_i, \mathbf{z}_k)/\tau) + \sum\limits_{k=1}^{N} \exp(s(\mathbf{z}_i, \mathbf{h}_k)/\tau)} \tag{4}$$

where s, N, τ refers to the same parameters as in Eq. 1. The cross-modal loss function $\mathcal{L}_{cmid}$ for a mini-batch is then formulated as:

$$\mathcal{L}_{cmid} = \frac{1}{2N}\sum_{i=1}^{N}[c(i, \mathbf{z}, \mathbf{h}) + c(i, \mathbf{h}, \mathbf{z})] \tag{5}$$

3.4. Overall Objective

Finally, we obtain the resultant loss function during training as the combination of $\mathcal{L}_{imid}$ and $\mathcal{L}_{cmid}$ where $\mathcal{L}_{imid}$ imposes invariance to point cloud transformation while $\mathcal{L}_{cmid}$ injects the 3D-2D correspondence.

$$\mathcal{L} = \mathcal{L}_{imid} + \mathcal{L}_{cmid} \tag{6}$$

4. Experiments

4.1. Pre-training

Dataset. We use ShapeNet [6] as the dataset for pretraining *CrossPoint*. It originally consists of more than 50,000 CAD models from 55 categories. We obtain the rendered RGB images from [63], which has 43,783 images from 13 object categories. For a given point cloud, we randomly select a 2D image out of all the rendered images, which is captured from an arbitrary viewpoint. We use 2048 points for each point cloud while we resize the corresponding rendered RGB image to 224×224. In addition to the augmentations applied for point cloud as described in Sec. 3.2, we perform random crop, color jittering and random horizontal flip for rendered images as data augmentation.

Implementation Details. To draw fair comparison with the existing methods, we deploy PointNet [38] and DGCNN [55] as the point cloud feature extractors. We use ResNet-50 [19] as the image feature extractor. We employ a 2-layer MLP as the projection heads which yield a 256-dimensional feature vector projected in the invariant space $\mathbb{R}^d$. We use Adam [25] optimizer with weight decay 1×10^{-4} and initial learning rate 1×10^{-3}. Cosine annealing [31] is employed as the learning rate scheduler and the model is trained end-to-end for 100 epochs. After pre-training we discard the image feature extractor $f_{\theta_\mathbf{I}}(.)$ and projection heads $g_{\phi_\mathbf{P}}(.)$ and $g_{\phi_\mathbf{I}}(.)$. All downstream tasks are performed on the pre-trained point cloud feature extractor $f_{\theta_\mathbf{P}}(.)$.

Table 1. **Comparison of ModelNet40 linear classification results with previous self-supervised methods.** A linear classifier is fit onto the training split of ModelNet40 using the pretrained model and overall accuracy for classification in test split is reported. Our method CrossPoint, surpasses existing works in both PointNet and DGCNN backbones.

Method	ModelNet40
3D-GAN [57]	83.3
Latent-GAN [1]	85.7
SO-Net [26]	87.3
FoldingNet [65]	88.4
MRTNet [12]	86.4
3D-PointCapsNet [71]	88.9
DepthContrast [68]	85.4
ClusterNet [67]	86.8
VIP-GAN [15]	90.2
PointNet + Jigsaw [46]	87.3
PointNet + STRL [21]	88.3
PointNet + Rotation [36]	88.6
PointNet + OcCo [53]	88.7
PointNet + CrossPoint (Ours)	**89.1**
DGCNN + Multi-Task [17]	89.1
DGCNN + Self-Contrast [10]	89.6
DGCNN + Jigsaw [46]	90.6
DGCNN + STRL [21]	90.9
DGCNN + Rotation [36]	90.8
DGCNN + OcCo [53]	89.2
DGCNN + CrossPoint (Ours)	**91.2**

4.2. Downstream Tasks

We evaluate the transferability of *CrossPoint* on three widely used downstream tasks in point cloud representation learning namely: (i) 3D object classification (synthetic and real-world), (ii) Few-shot object classification (synthetic and real-world) and (iii) 3D object part segmentation.

(i) 3D Object classification. We perform our classification experiments on both ModelNet40 [58] and ScanObjectNN [52] to demonstrate the generalizability of our approach in both synthetic and real-world 3D shape representation learning. ModelNet40 is a synthetic dataset, where

the point clouds are obtained by sampling 3D CAD models. It contains 12,331 objects (9,843 for training and 2,468 for testing) from 40 categories. ScanObjectNN [52] is more realistic and challenging 3D point cloud classification dataset which comprises of occluded objects extractred from real-world indoor scans. It contains 2,880 objects (2304 for training and 576 for testing) from 15 categories.

We follow the standard protocol [21, 53] to test the accuracy of our model in object classification. We freeze the pretrained point cloud feature extractor and fit a simple linear SVM classifier on the train split of the classification datasets. We randomly sample 1024 points from each object for both training and testing the classification results. Our CrossPoint also delivers a consistent performance in different backbones. We perform experiments in both PointNet [38] and DGCNN [55] where PointNet is an MLP based feature extractor while DGCNN is built on graph convolutional networks. Table. 1 reports the linear classification results on ModelNet40. It is clear that CrossPoint outperforms previous state-of-the-art unsupervised methods in both of the feature extractors, thus establishing a new benchmark in self-supervised learning for point clouds. In particular, our model outperforms Depth-Contrast [68], which also employ a cross-modal setting for point cloud representation learning, by a significant margin of 5.8%. While our approach surpasses prior works which utilizes self-supervised contrastive learning [10, 17, 21] by considerable margins, some other methods [28, 42, 59] cannot be compared in a fair manner for varying reasons such as different pretraining mechanisms and discrepancy in feature extractors.

Table 2. **Comparison of ScanObjectNN linear classification results with previous self-supervised methods.** CrossPoint shows consistent improvements over prior works in both PointNet and DGCNN backbones. This illustrates the effectiveness of our approach in real-world setting.

Method	Backbone	
	PointNet	DGCNN
Jigsaw [46]	55.2	59.5
OcCo [53]	69.5	78.3
STRL [21]	74.2	77.9
CrossPoint (Ours)	**75.6**	**81.7**

Table. 2 demonstrates the linear evaluation results on ScanObjectNN. Significant accuracy gains of 1.3% in PointNet backbone and 3.4% in DGCNN backbone, over previous state-of-the-art unsupervised methods show that our proposed joint learning approach generalizes to out-of-distribution data as well.

(ii) Few-shot object classification. Few-shot learning (FSL) aims to train a model that generalizes with limited data. We conduct experiments on conventional few-shot task (N-way K-shot learning), where the model is evaluated on N classes, and each class contains N samples. Similar to standard 3D object classification, we use ModelNet40 and ScanObjectNN datasets to carry out FSL experiments. While there is no standard split for FSL in both of the datasets, for a fair comparison with previous methods [47, 53], we randomly sample 10 few-shot tasks and report the mean and standard deviation. Table. 3 shows the FSL results on ModelNet40, where CrossPoint outperforms prior works in all the FSL settings in both PointNet and DGCNN backbones. It is noticeable that our method with DGCNN backbone performs poorly in some of the FSL settings compared to that with PointNet backbone. Similar pattern is also observed in previous methods [47, 53] as well. We attribute to the fact that complex backbones might degrade the few-shot learning performance which has been observed consistently in the FSL literature in images [51].

Table 3. **Few-shot object classification results on ModelNet40.** We report mean and standard error over 10 runs. Top results of each backbone is colored in red and blue. Proposed CrossPoint improves the few-shot accuracy in all the reported settings. Table is an extended version from [53]

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
3D-GAN [57]	55.8±3.4	65.8±3.1	40.3±2.1	48.4±1.8
FoldingNet [65]	33.4±4.1	35.8±5.8	18.6±1.8	15.4±2.2
Latent-GAN [1]	41.6±5.3	46.2±6.2	32.9±2.9	25.5±3.2
3D-PointCapsNet [71]	42.3±5.5	53.0±5.9	38.0±4.5	27.2±4.7
PointNet++ [39]	38.5±4.4	42.4±4.5	23.1±2.2	18.8±1.7
PointCNN [27]	65.4±2.8	68.6±2.2	46.6±1.5	50.0±2.3
RSCNN [29]	65.4±8.9	68.6±7.0	46.6±4.8	50.0±7.2
PointNet + Rand	52.0±3.8	57.8±4.9	46.6±4.3	35.2±4.8
PointNet + Jigsaw [46]	66.5±2.5	69.2±2.4	56.9±2.5	66.5±1.4
PointNet + cTree [47]	63.2±3.4	68.9±3.0	49.2±1.9	50.1±1.6
PointNet + OcCo [53]	89.7±1.9	92.4±1.6	83.9±1.8	89.7±1.5
PointNet + CrossPoint	90.9±4.8	93.5±4.4	84.6±4.7	90.2±2.2
DGCNN + Rand	31.6±2.8	40.8±4.6	19.9±2.1	16.9±1.5
DGCNN + Jigsaw [46]	34.3±1.3	42.2±3.5	26.0±2.4	29.9±2.6
DGCNN + cTree [47]	60.0±2.8	65.7±2.6	48.5±1.8	53.0±1.3
DGCNN + OcCo [53]	90.6±2.8	92.5±1.9	82.9±1.3	86.5±2.2
DGCNN + CrossPoint	92.5±3.0	94.9±2.1	83.6±5.3	87.9±4.2

We report the FSL results on ScanObjectNN dataset in Table. 4. CrossPoint produces significant accuracy gains in most of the settings in both PointNet and DGCNN feature extractors, proving the capability of generalizing with a limited data even in an out-of-distribution setting.

(iii) 3D Object part segmentation. We perform object part segmentation in the widely used ShapeNetPart dataset [66]. It contains 16881 3D objects from 16 categories, annotated with 50 parts in total. We initially pretrain the backbone proposed in DGCNN [55] for part segmentation using our approach in ShapeNet dataset and fine tune in an end-to-end manner in the train split of ShapeNetPart dataset. We

Table 4. **Few-shot object classification results on ScanObjectNN.** We report mean and standard error over 10 runs. Top results of each backbone is colored in red and blue. Proposed CrossPoint improves the few-shot accuracy in all the reported settings. Table is an extended version from [53]

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
PointNet + Rand	57.6±2.5	61.4±2.4	41.3±1.3	43.8±1.9
PointNet + Jigsaw [46]	58.6±1.9	67.6±2.1	53.6±1.7	48.1±1.9
PointNet + cTree [47]	59.6±2.3	61.4±1.4	53.0±1.9	50.9±2.1
PointNet + OcCo [53]	70.4±3.3	72.2±3.0	54.8±1.3	61.8±1.2
PointNet + CrossPoint	68.2±1.8	73.3±2.9	58.7±1.8	64.6±1.2
DGCNN + Rand	62.0±5.6	67.8±5.1	37.8±4.3	41.8±2.4
DGCNN + Jigsaw [46]	65.2±3.8	72.2±2.7	45.6±3.1	48.2±2.8
DGCNN + cTree [47]	68.4±3.4	71.6±2.9	42.4±2.7	43.0±3.0
DGCNN + OcCo [53]	72.4±1.4	77.2±1.4	57.0±1.3	61.6±1.2
DGCNN + CrossPoint	74.8±1.5	79.0±1.2	62.9±1.7	73.9±2.2

report the mean IoU (Intersection-over-Union) metric, calculated by averaging IoUs for each part in an object before averaging the obtained values for each object class in Table. 5. Part segmentation using the backbone pretrained via CrossPoint outperforms the randomly initialised DGCNN backbone by 0.4%. This shows that CrossPoint provides a better weight initialization to the feature extractors. Accuracy gains over the previous self-supervised learning frameworks indicates that CrossPoint, by imposing intra-modal and cross-modal correspondence in a joint manner, tends to capture fine-grained part-level attributes which is crucial in part segmentation.

Table 5. **Part segmentation results on ShapeNetPart dataset.** We report the mean IoU across all the object classes. *Supervised* indicates the models trained with randomly initialising feature backbones, while *Self-Supervised* models are the ones initialised with pretrained feature extractors.

Category	Method	Mean IoU
Supervised	PointNet [38]	83.7
	PointNet++ [39]	85.1
	DGCNN [55]	85.1
Self-Supervised	Self-Contrast [10]	82.3
	Jigsaw [46]	85.3
	OcCo [53]	85.0
	PointContrast [59]	85.1
	Liu *et al*. [28]	85.3
	CrossPoint (Ours)	**85.5**

4.3. Ablations and Analysis

Impact of joint learning objective. As described in Sec. 3, our approach aims to train the model with a joint learning objective. We hypothesize that, addressing both intra-modal and cross-modal correspondence in a joint manner contribute to a better representation learning than the individual learning objectives. Intra-modal correspondence encourages the model to capture the fine-grained part semantics via imposing invariance to transformations and cross-modal correspondence establishes hard positive feature samples for contrastive learning to make the learning even more challenging, thus yielding better results. We empirically test this hypothesis by training the model in all possible settings and evaluating a linear SVM classifier in both ModelNet40 and ScanObjectNN datasets. Figure. 3 graphically illustrates that in all the learning settings, the proposed joint learning paradigm performs better than the individual objectives. In particular, the combination of intra-modal and cross-modal learning objectives obtains accuracy gains of 1.2% and 0.7% over the second best approach in ModelNet40 and ScanObjectNN respectively with the DGCNN feature extractor.

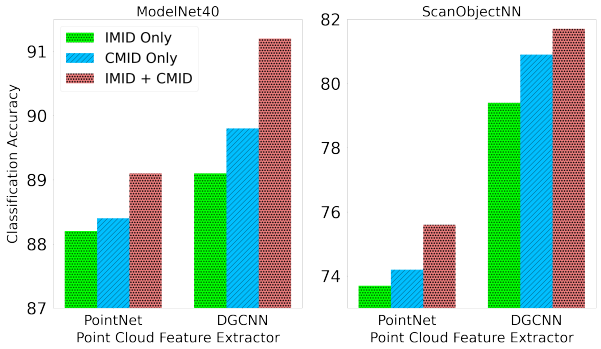

Figure 3. Impact of the joint learning objective when compared to individual intra-modal and cross-modal objectives. Classification results with Linear SVM on the pretrained embedding on ModelNet40 (left) and ScanObjectNN (right) dataset.

We specifically observe that linear evaluation with cross-modal learning objective marginally outperforms that with intra-modal learning objective in the classification accuracy metrics. We believe that the cross-modal learning objective facilitates a part semantic understanding, by embedding the image feature in the close proximity to the features of both the augmented point clouds by leveraging the point cloud prototype feature. Figure. 4 visualizes the t-SNE plot of the features obtained from the test split of ModelNet10 dataset. It is visible that both CMID and IMID settings provide a good discrimination to classes even without explicitly trained with labeled data. However, the class boundaries of some classes (e.g., desk, table) are not precise and compact. Joint learning objective is able to create a better discrimination boundaries in such classes.

Number of corresponding 2D images. We investigate the contribution of the image branch by varying the number of rendered 2D images (n). We select the rendered 2D im-

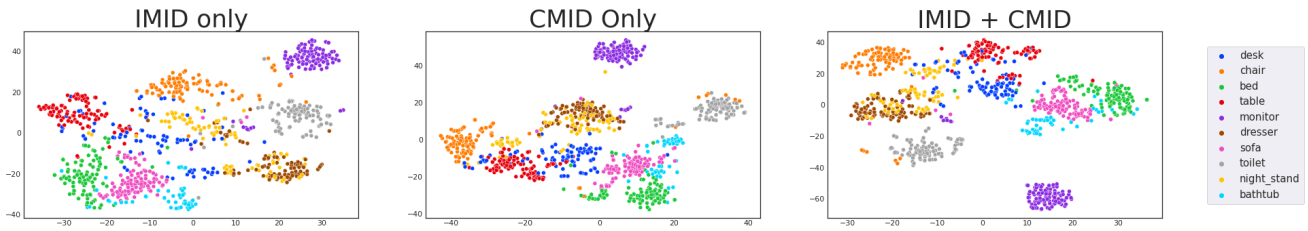

Figure 4. t-SNE visualization of features on the test split of **ModelNet10** dataset after training the DGCNN backbone in a self-supervised manner. The proposed joint learning approach provide better discrimination of classes (e.g., *desk, table*) when compared to models learnt with individual objectives.

age captured from different random directions. In case of more than one rendered 2D image, we compute the mean of all the projected features of rendered images to perform the cross-modal instance discrimination (CMID). Table. 6 reports the linear SVM classification results on ModelNet40 dataset. Our approach, even with a single rendered 2D image captures the cross-modal correspondence to yield better linear classification results. It is clear that, when using more than 2 rendered images, the information gathered from 2D image modality might have become redundant, hence the drop in accuracy.

Table 6. **Linear classification results on ModelNet40 with varying number of rendered 2D images (n).** CrossPoint with single corresponding image performs better than or equal to multiple rendered images. We choose n=1 for all the experiments.

No. of rendered 2D images (n)	**1**	**2**	**3**	**4**	**5**
Linear Accuracy	**91.2**	91.2	90.9	91.0	90.5

Few-shot image classification on CIFAR-FS. Even though we discard the image feature extractor during point cloud downstream tasks, we perform a simple few-shot image classification to investigate the image understanding capability of it. We use CIFAR-FS [4], which is a widely used dataset for few-shot image classification, that contains 100 categories with 64, 16, and 20 train, validation, and test splits. Table. 7 reports the results in comparison with the standard baseline, RFS [51] in 5-way 1-shot and 5-way 5-shot settings. It is to be noticed that CrossPoint, without any supervised fine-tuning, fails to generalize well in the few-shot image classification setting. We believe that this is because, there is a considerable discrepancy between the rendered 2D images from point clouds and the images in CIFAR-FS which are real world images. Hence, CrossPoint fails to generalize to such an out-of-distribution data which is a limitation of our work. However, initializing the backbone with the unsupervisedly trained image feature extractor in CrossPoint and finetuning using the method proposed in RFS outperforms the baseline results by significant margins in both the few-shot settings.

Table 7. **Few-shot image classification results on CIFAR-FS.** Fine-tuning CrossPoint with RFS improves the performance.

Method	Backbone	5-way 1-shot	5-way 5-shot
CrossPoint	ResNet-50	24.12±0.48	28.18±0.54
RFS [51]	ResNet-50	60.20±0.87	76.79±0.71
CrossPoint + RFS	ResNet-50	**64.45±0.86**	**80.14±0.65**

5. Conclusion

In this paper, we propose CrossPoint, a simple self-supervised learning framework for 3D point cloud representation learning. Even though our approach is trained on synthetic 3D object dataset, experimental results in downstream tasks such as 3D object classification and 3D object part segmentation in both synthetic and real-world datasets demonstrate the efficacy of our approach in learning transferable representations. Our ablations empirically validate our claim that joint learning of imposing intra-modal and cross-modal correspondences leads to more generic and transferable point cloud features. Additional few-shot image classification experiment provides a powerful insight for cross-modal understanding which can be explored in future researches.

Acknowledgments

The authors would like to thank Salman Khan (MBZUAI, UAE) and Sadeep Jayasumana (Google Research, NY) for their valuable comments and suggestions on the manuscript. The computational resources of this research were supported by the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education, Sri Lanka funded by the World Bank. We also thank Facebook Reality Labs for partially funding this work through Facebook Research Awards for Explorations of Trust in AR, VR, and Smart Devices.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3D point clouds. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 40–49, 2018. 1, 2, 3, 5, 6

[2] Mohamed Afham, Salman Khan, Muhammad Haris Khan, Muzammal Naseer, and Fahad Shahbaz Khan. Rich semantics improve few-shot learning. *32nd British Machine Vision Conference*, 2021. 2

[3] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3

[4] Luca Bertinetto, Joao F. Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2019. 8

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 1, 4

[6] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *arXiv:1512.03012*, 2015. 2, 3, 5

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607, 2020. 1, 2, 3, 4

[8] Judy S. DeLoache, Mark S. Strauss, and Jane Maynard. Picture perception in infancy. *Infant Behavior and Development*, 2:77–89, 1979. 2

[9] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11162–11173, 2021. 2, 3

[10] Bi'an Du, Xiang Gao, Wei Hu, and Xin Li. Self-contrastive learning with hard negative sampling for self-supervised point cloud learning. In *ACM Multimedia Conference*, pages 3133–3142, 2021. 3, 5, 6, 7

[11] Benjamin Eckart, Wentao Yuan, Chao Liu, and Jan Kautz. Self-supervised learning on 3D point clouds by learning discrete generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8248–8257, 2021. 3

[12] Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3D point cloud processing. In *ECCV*, 2018. 5

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284, 2020. 1, 2, 3, 4

[14] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1735–1742, 2006. 3

[15] Zhizhong Han, Mingyang Shang, Yu-Shen Liu, and Matthias Zwicker. View inter-prediction gan: Unsupervised representation learning for 3D shapes by learning global shape memories to support local view predictions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:8376–8384, 2019. 3, 5

[16] Zhizhong Han, Xiyang Wang, Yu-Shen Liu, and Matthias Zwicker. Multi-angle point cloud-vae: Unsupervised feature learning for 3D point clouds from multiple angles by joint self-reconstruction and half-to-half prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[17] Kaveh Hassani and Mike Haley. Unsupervised multi-task feature learning on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3, 5, 6

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 2, 3

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 5

[20] Lianghua Huang, Yu Liu, Bin Wang, Pan Pan, Yinghui Xu, and Rong Jin. Self-supervised video representation learning by context and motion decoupling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13886–13895, 2021. 2

[21] Siyuan Huang, Yichen Xie, Song-Chun Zhu, and Yixin Zhu. Spatio-temporal self-supervised representation learning for 3D point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6535–6545, 2021. 2, 3, 4, 5, 6

[22] Drew Arad Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. In *International Conference on Learning Representations*, 2018. 2

[23] Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21798–21809, 2020. 2

[24] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr - modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1780–1790, 2021. 2

[25] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 5

[26] Jiaxin Li, Ben M. Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9397–9406, 2018. 3, 5

[27] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. In *Advances in Neural Information Processing Systems*, volume 31, 2018. 2, 6

[28] Fayao Liu, Guosheng Lin, and Chuan-Sheng Foo. Point discriminative learning for unsupervised representation learning on 3D point clouds. *arXiv preprint arXiv:2108.02104*, 2021. 3, 6, 7

[29] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8895–8904, 2019. 2, 6

[30] Yueh-Cheng Liu, Yu-Kai Huang, Hung-Yueh Chiang, Hung-Ting Su, Zhe-Yu Liu, Chin-Tang Chen, Ching-Yu Tseng, and Winston H. Hsu. Learning from 2D: Contrastive pixel-to-point knowledge transfer for 3D pretraining. *arXiv preprint arXiv:2104.04687*, 2021. 3

[31] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016. 5

[32] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2906–2917, 2021. 1, 2

[33] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3

[34] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12475–12486, 2021. 1, 2, 3, 4

[35] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[36] Omid Poursaeed, Tianxing Jiang, Quintessa Qiao, Nayun Xu, and Vladimir G. Kim. Self-supervised learning of point clouds via orientation estimation. *2020 International Conference on 3D Vision (3DV)*, pages 1018–1028, 2020. 2, 3, 5

[37] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3D object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[38] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 5, 6, 7

[39] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 1, 2, 6, 7

[40] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6964–6974, 2021. 2

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021. 2, 3

[42] Yongming Rao, Jiwen Lu, and Jie Zhou. Global-local bidirectional reasoning for unsupervised representation learning of 3D point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 6

[43] Susan A. Rose. Infants' transfer of response between two-dimensional and three-dimensional stimuli. *Child Development*, 48:1086–1091, 1977. 2

[44] Aditya Sanghi. Info3D: Representation learning on 3D objects using mutual information maximization and contrastive learning. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[45] Mert Bulent Sariyildiz, Julien Perez, and Diane Larlus. Learning visual representations with caption annotations. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3

[46] Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32:12962–12972, 2019. 1, 3, 5, 6, 7

[47] Charu Sharma and Manohar Kaul. Self-supervised few-shot learning on point clouds. In *Advances in Neural Information Processing Systems*, volume 33, pages 7212–7221, 2020. 3, 6, 7

[48] Chao Sun, Zhedong Zheng, Xiaohan Wang, Mingliang Xu, and Yi Yang. Point cloud pre-training by mixing and disentangling. *arXiv preprint arXiv:2109.00452*, 2021. 3

[49] Hugues Thomas, Charles R. Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, Francois Goulette, and Leonidas J. Guibas. KPConv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 2

[50] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 3

[51] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *European Conference on Computer Vision (ECCV)*, 2020. 6, 8

[52] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *2019 IEEE/CVF International*

Conference on Computer Vision (ICCV), pages 1588–1597, 2019. 2, 5, 6

[53] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matthew J. Kusner. Unsupervised point cloud pre-training via occlusion completion. In *International Conference on Computer Vision, ICCV*, 2021. 1, 3, 5, 6, 7

[54] Jinpeng Wang, Yiqi Lin, Andy J. Ma, and Pong C. Yuen. Self-supervised temporal discriminative learning for video representation learning. *arXiv preprint arXiv:2008.02129*, 2020. 2

[55] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 2019. 1, 2, 5, 6, 7

[56] Zhonghao Wang, Mo Yu, Kai Wang, Jinjun Xiong, Wen-mei Hwu, Mark Hasegawa-Johnson, and Humphrey Shi. Interpretable visual reasoning via induced symbolic space. *arXiv preprint arXiv:2011.11603*, 2020. 2

[57] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3D generative-adversarial modeling. In *Advances in Neural Information Processing Systems*, volume 29, 2016. 1, 2, 3, 5, 6

[58] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3D ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2, 5

[59] Saining Xie, Jiatao Gu, Demi Guo, Charles Qi, Leonidas Guibas, and Or Litany. PointContrast: Unsupervised pre-training for 3D point cloud understanding. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 6, 7

[60] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O. Pinheiro. Adaptive cross-modal few-shot learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 2

[61] Chenfeng Xu, Shijia Yang, Bohan Zhai, Bichen Wu, Xiangyu Yue, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2Point: 3D point-cloud understanding with pretrained 2d convnets. *arXiv preprint arXiv:2106.04180*, 2021. 3

[62] Mutian Xu, Runyu Ding, Hengshuang Zhao, and Xiaojuan Qi. Paconv: Position adaptive convolution with dynamic kernel assembling on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3173–3182, 2021. 2

[63] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. DISN: Deep implicit surface network for high-quality single-view 3D reconstruction. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 1, 5

[64] Juyoung Yang, Pyunghwan Ahn, Doyeon Kim, Haeil Lee, and Junmo Kim. Progressive seed generation auto-encoder for unsupervised point cloud learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6413–6422, 2021. 3

[65] Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 206–215, 2018. 3, 5, 6

[66] L. Yi, Vladimir G. Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas J. Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM Transactions on Graphics (TOG)*, 35:1–12, 2016. 6

[67] Ling Zhang and Zhigang Zhu. Unsupervised feature learning for point cloud understanding by contrasting and clustering using graph convolutional neural networks. In *International Conference on 3D Vision (3DV)*, pages 395–404, 2019. 3, 5

[68] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3D features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 2, 3, 4, 5, 6

[69] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3DNet: 3D object detection using hybrid geometric primitives. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[70] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H.S. Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16259–16268, 2021. 2

[71] Yongheng Zhao, Tolga Birdal, Haowen Deng, and Federico Tombari. 3D point capsule networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 5, 6

[72] Rui Zhu, Bingchen Zhao, Jingen Liu, Zhenglong Sun, and Chang Wen Chen. Improving contrastive learning by visualizing feature transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2