PointCMC: Cross-Modal Multi-Scale Correspondences Learning for Point Cloud Understanding

Honggu Zhou
School of Media and Desgin
Hangzhou Dianzi University,China
hongguzhou@hdu.edu.cn

Xiaogang Peng
School of Media and Desgin
Hangzhou Dianzi University,China
pengxiaogang@hdu.edu.cn

Ming Zeng
School of Informatics
Xiamen University, Xiamen, China
zengming@xmu.edu.cn

Zizhao Wu*
School of Media and Desgin
Hangzhou Dianzi University,China
wuzizhao@hdu.edu.cn

Abstract

Cross-modal frameworks have achieved impressive performance in point cloud representation learning, where a 2D image encoder is employed to transfer knowledge to a 3D point cloud encoder. However, the local structures between point clouds and corresponding images are unaligned, which results in a challenge for the 3D point cloud encoder to learn fine-grained image-point cloud interactions. In this paper, we introduce a novel multi-scale training strategy (PointCMC) to enhance fine-grained cross-modal knowledge transfer in the cross-modal framework. Specifically, we design a Local-to-Local (L2L) module that implicitly learns the correspondence of local features by aligning and fusing extracted local feature sets. Moreover, we introduce the Cross-Modal Local-Global Contrastive (CLGC) loss, which enables the encoder to capture discriminative features by reasoning local structures to their corresponding cross-modal global shape. The extensive experimental results demonstrate that our approach outperforms the previous unsupervised learning methods in various downstream tasks such as 3D object classification and semantic segmentation.

Keywords: ***Self-Supervised Representation Learning, Contrastive Learning, Cross-Modal Learning, Point Cloud Understanding***

1. Introduction

Assisting machines in comprehending the 3D world is critical to numerous real-world applications, including autonomous driving, AR, VR, and other fields. In order to enhance machine comprehension of the surroundings, intermediary forms such as point clouds, meshes, and voxels have emerged. Point cloud data has gained immense popularity among these alternatives due to its ease of use, as it does not require preprocessing. Directly processing point clouds is contingent upon the ability to comprehend them, and the core challenge in this regard is to capture discriminative representations. While many supervised approaches [35][37][28][50][26][52][59] donate to tackle this issue, they rely on costly and time-consuming annotations. Self-supervised methods[1][23][56][49][64][2] have emerged to address this problem, which learns the latent semantics of point cloud data through pretext tasks. However, the scales of point cloud benchmarks constrain models from learning

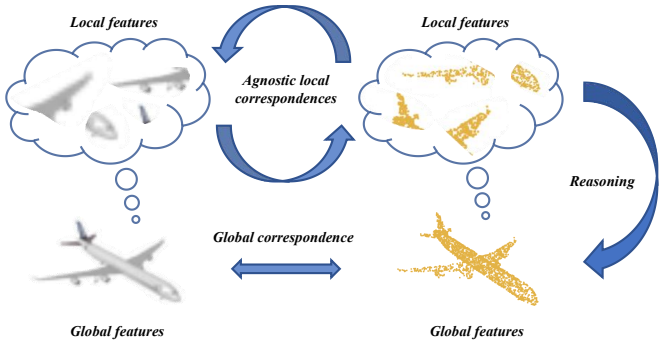


Figure 1: We present PointCMC, a method that utilizes pairs of images and point clouds with positive but agnostic correspondences among local parts. PointCMC uses intra- and inter-attention modules to explore these agnostic correspondences. Furthermore, PointCMC investigates cross-modal local-global correspondences by reasoning global features from local features in different modalities.

*Corresponding author

discriminative representations. Given the successful development of multimodal contrastive learning in domains such as audio-video and image-natural language, the corresponding 2D images of point clouds have become a promising auxiliary input for point cloud comprehension, which does not require additional annotations.

Cross-modal contrastive learning approaches[2] [18][30] can help encoders learn 2D-3D correspondences for understanding point clouds. These promising strategies mainly rely on two key advantages: 1) **Complementary**: The 2D and 3D learning signals exhibit a certain degree of complementary. 2) **Agnostic transformations**: Enforcing 2D and 3D correspondence promotes agnostic transformations that can improve 3D representation learning. Some methods[2][18] explore object-level cross-modal correspondences, whereas others[30][29][47] investigate point-pixel level correspondences. Object-level methods often lose local features, as the encoders depend on consuming points or pixels to acquire global features. Besides, point-pixel level methods predominantly require annotated point-pixel pairs or image-reconstructed point clouds, which may result in a loss of 2D semantics and local geometric perception in the reconstruction process. These approaches generally lack the ability to model cross-modal local correspondences, as local features contain rich learning signals. Therefore, it is natural to raise a question: *Could we enhance point cloud shape analysis by modeling multi-scale correspondences across modalities, particularly correspondences between local representations?*

The first challenge is to establish direct local correspondences across modalities. Directly annotating the correspondences between local regions is a time-consuming and complex task. Nevertheless, the intermediate representations captured by the encoder are collections of local representations, which have the potential to facilitate the transfer of cross-modal local representations. The second challenge is to establish local and global correspondences across modalities. Inspired by [39], we can design a reasoning task to implicitly build cross-modal local-global interactions to overcome the lack of correspondence.

We present PointCMC, a novel framework designed to address the aforementioned challenges. We utilize two independent encoders as general setting as other methods [2][30][19], one for image data and the other for point cloud data, to extract multi-scale representations from their respective modalities. However, unlike these methods, we introduce a local-to-local (L2L) module that aims to perform cross-modal alignment and fusion on intermediate representations to investigate cross-modal local correspondences. To further explore the potential correlation between cross-modal local and global features, we propose a Cross-Modal Local-Global Contrastive (CLGC) loss that enables the point cloud encoder to capture the shared semantics of cross-modal global and local features. Moreover, a conventional cross-modal global Contrastive (CGC) loss is also applied in PointCMC to facilitate knowledge transfer from the image encoder to the point cloud encoder.

We have conducted to validate the effectiveness of our method with a series of downstream tasks on several generic point cloud backbone networks. Furthermore, a sequence of ablation experiments has validated the impact of the proposed three modules to our method.

The main contributions of our approach are as follows:

- We propose a novel L2L module that does not require the annotation of local correspondences between point clouds and images. This module investigates cross-modal local correspondences by aligning and fusing intermediate representations across modalities.
- We propose a CLGC loss function, which is used to drive a local-to-global reasoning task to promote the learning of shared semantics across modalities.
- We evaluate our method in three downstream tasks: 1) 3D object classification, 2) few-shot object classification, and 3) 3D Object part segmentation. Quantitative and qualitative results on real-world and synthetic benchmarks have shown that our methods capture discriminative representation and outperform the state-of-the-art methods.

2. Related Work

2.1. Supervised methods for point cloud learning

Point cloud understanding is a difficult task compared to NLP and 2D. On the one hand, point clouds lack a highly regular structure: there are images for 2D and word embeddings for NLP. On the other hand, point clouds also need to be permutation invariant when processing. Unlike 2D images, which have a unified architecture like convolutional neural networks, there are many network structures to handle point clouds from different perspectives. Point-based networks [35][37] directly consume raw point clouds, and the pioneering work is Pointnet [35], which proposes an architecture that stacks MLP layers to extract point-wise features independently, then aggregates them by max pooling. However, PointNet fails to capture local information. To address this issue, Qi et al. [37] then propose PointNet++ to learn global and local information through the hierarchical aggregation of neighborhood points. Graph-based networks model the relationship between points as a graph.For example, Wang et al. [50] propose pioneering work dubbed DGCNN, which use a graph convolution named EdgeConv to capture local features for each point from its k nearest neighbor points. Voxel-based networks [36][21] voxelize irregular point clouds into regular 3D grids and use 3D

convolution for feature learning. Spatial CNN-based networks [44][28] directly apply Spatial-convolutions on irregular point clouds. RSCNN [28] is proposed to explicitly encode the geometric relation of points and achieve contextual shape-aware learning of point clouds.

2.2. Self-supervised methods for point cloud learning

Motivated by the success of self-supervised representation learning [6][65][3][5][15][40][11] in the field of 2D, some recent works have investigated self-supervised methods for 3D representation learning, which can be roughly divided into two categories.

Classical learning methods. Classical methods typically rely on autoencoders [20] and generative adversarial networks (GANs)[10]. Autoencoder networks aim to train an encoder to extract the representation of input point clouds, while a decoder reconstructs the point cloud from the captured representation. This simple process enables the encoder to learn shape information, making autoencoder networks popular for many self-supervised learning methods. For example, self-reconstruction[61][8][23][67][13], a classical self-supervised learning task based on autoencoder networks, forces the final output to be the same as the input. Autoencoder-based point cloud up-sampling[24][63][25] and completion[48][17][42][55] emphasize learning fine-grained representations, which enables the encoder to capture more comprehensive representations. In addition to autoencoder networks, GAN-based methods[1][45][22] learn representations by adversarially training generators and discriminators without annotations. After training, the learned discriminator can be applied into various downstream tasks.

Contrastive learning methods. Contrastive learning uses predefined positive and negative samples as inputs, maximizing agreements between positive pairs and minimizing agreements between positive and negative samples. Xie et al. [56] propose to learn representations by performing point-level discriminations from two viewpoints. Zhang et al. [64] propose a method based on deep graph convolutional neural networks, which learns representations by verifying whether two random parts belong to the same object. Du et al. [7] propose a method based on the non-local self-similarity of point clouds to learn representation. Rao et al. [39] combine contrastive learning and self-reconstruction to formulate a task focusing on global and local representations reasoning. Huang et al. [16] introduce BYOL [11] into the point cloud and extract spatial and temporal representation from point clouds. Unlike prior works, we attempt to exploit contrastive learning in modeling 2D-3D correspondence and further explore the multi-scale correspondences across modalities.

2.3. Cross-modal learning for point cloud learning

In recent years, cross-modal learning [68][38][2][53][60] has shown that additional training signals, such as data from other modalities, can help encoders to capture representations. Some recent works have shown the success of cross-modal learning between 2D images and natural language. However, as pointed out in [54] , few methods use 2D images as auxiliary inputs to enhance 3D representation. Xu et al. [57] directly transfer 2D convolutional models to the point cloud model by inflating 2D convolutions and replicating parameters. Liu et al. [30] introduce a cross-modal knowledge distillation approach that enforces point correspondences by reconstructing 2D images into point clouds, but this process may lead to the loss of 2D semantics. Jing et al. [18] and Afham et al. [2] learn 3D point cloud representation by predicting correspondences between cross-modal global representations and introducing auxiliary cross-modal instance discrimination, respectively. Inspired by the above cross-modal learning methods, we propose a training strategy that establishes multi-scale cross-modal correspondences to transfer 2D image knowledge to the point cloud encoder.

3. Method

The proposed PointCMC method aims to model cross-modal multi-scale correspondences in a self-supervised way, and its key component is the L2L module. This section provides a detailed overview of the method and its key component. The overall framework is shown in Figure 2,which includes several components that work together to achieve the desired results.

3.1. Method Overview

The input can be represented as $\{P_i,\ I_i\}$, with $P_i \in \mathbb{R}^{N\times 3}$ and $I_i \in \mathbb{R}^{H\times W\times 3}$, where I_i is a random-view rendered image from point cloud P_i. Our architecture consists of two Encoders $\{E^p,\ E^{img}\}$ and the L2L module. We use E^P to extract multi-scale point cloud features from augmented P_i and use E^{img} to extract multi-scale features from augmented I_i. Using the extracted features as inputs, our method first models the correspondences of cross-modal local features through the L2L modules. This module employs intra-attention and inter-attention blocks to consider the correspondences between cross-modal local patches and applies a feature alignment block (FAB) to constrain fused features. In addition, we model the cross-modal local-global relationships by utilizing the CLGC loss. To learn global correspondences, we also use the CGC loss to enforce the global features of images as a centroid to the global features of point clouds.

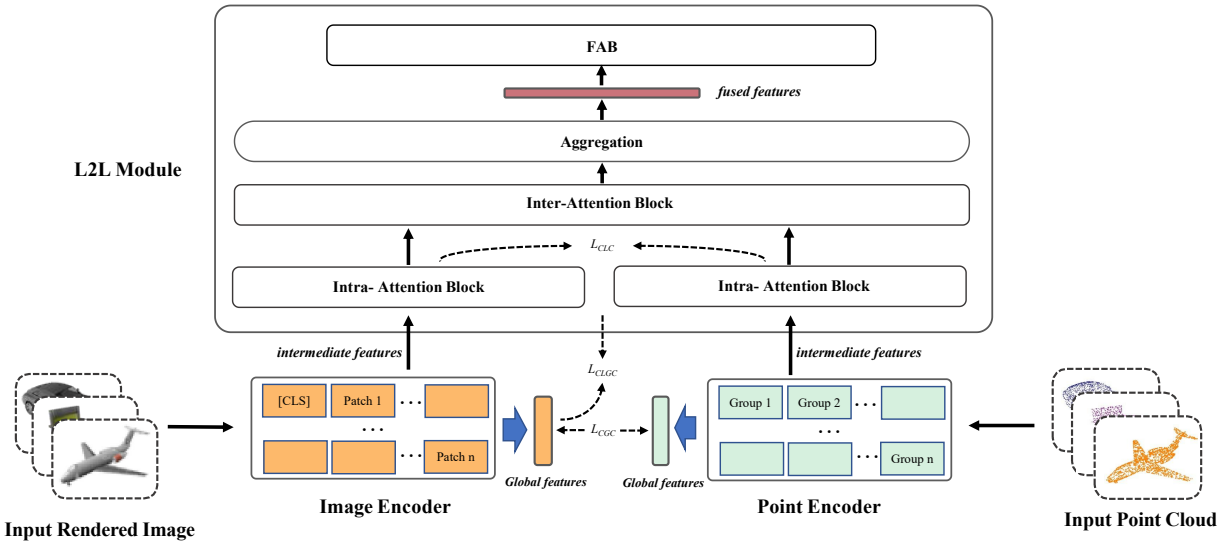


Figure 2: PointCMC leverages an additional module, known as L2L, which inserts between the 3D and 2D backbones. This module consists of intra-attention blocks and inter-attention blocks. Specifically, the CLC loss is employed to explicitly learn local correspondences within the optimized unimodal representations of intra-attention modules. Moreover, inter-attention modules further capture local correspondences implicitly through the use of the FAB module. Additionally, the CLGC loss is applied to explore the potential connections between representations across different levels of modalities. The CGC loss is also applied to learn the connections between global representations across modalities. After training, the L2L module is discarded, and only the point cloud encoders are retained for downstream tasks.

3.2. L2L Module

This subsection introduces how the L2L module utilizes intra-attention and inter-attention blocks to learn local correspondences for image and point cloud. Specifically, the L2L module takes the cross-modal intermediate representations into intra-attention blocks to procure their optimized versions. The CLC loss is then applied to these optimized versions. Subsequently, we fuse these optimized representations and continue their refinement using inter-attention blocks. Finally, the representations are input into the FAB block to facilitate local correspondences.

Intra-attention block. The intra-attention block, illustrated in Figure 3, utilizes the self-attention mechanism that takes three specific inputs: query (Q), key (K), and value (V). This block selectively weighs the importance of various positions in a sequence by computing the similarity between the representation of each position (Q) and that of all other positions (K and Q). The resulting attention scores are used to compute a weighted sum of the values, representing the self-attention mechanism's final output. In our work, we employ intra-attention blocks to facilitate interaction within the modality.

Inter-attention block. In our inter-attention module, we have integrated two cross-attention mechanisms: the co-attention and merge-attention mechanisms. As illustrated in Figure 3, under the co-attention mechanism, the optimized representations of the point cloud and image are individually inputted into separate transformers, and cross-modal interactions are enabled through cross-attention techniques. In addition, the optimized representations of the point cloud and image are merged under the merge-attention mechanism and subsequently fed into the same transformer.

The co-attention mechanism is implemented by acquiring the K and Q values from another modality, thereby suppressing unimodal interactions and enhancing cross-modal interactions. On the other hand, the merge-attention mechanism shares Q, K and V between modalities, further promoting cross-modal interactions. Our ablation results demonstrate that the joint use of these two mechanisms is a superior choice.

FAB block. The FAB module is a nonlinear transformation that maps cross-modal representations to the same feature space, which can significantly reduce negative transfer caused by directly aligning features, as demonstrated in the results presented in[60] . To enhance the alignment of cross-modal representations, we pre-train the FAB module using image features and subsequently freeze it during the PointCMC training stage.

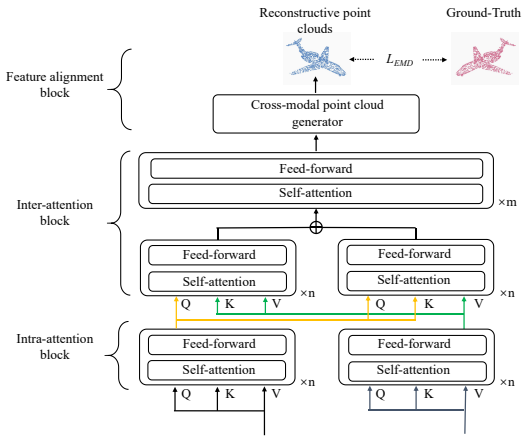


Figure 3: L2L module employs the transformer architecture, which consists of intra-attention blocks, inter-attention blocks, and the feature align block. Note that the representations captured by inter-attention blocks are applied to CLC loss and CLGC loss.

3.3. Pre-training Objectives

In this subsection, we provide an in-depth discussion on the training of PointCMC. To be more precise, we introduce the multi-scale contrastive losses, which aim to learn the multi-scale correspondences between cross-modal representations. Subsequently, we present the commonly used feature align loss in point cloud analysis tasks, which aims to optimize the reconstruction effect of fused features, thereby promoting implicit alignment between features.

CLC Loss. By leveraging the intra-attention blocks mentioned earlier, we can allocate more significant attention to salient local regions in cross-modal intermediate representations, resulting in optimized local representations denoted as $\{f_l^{img}, f_l^{p}\}$. Our primary objective is to ensure that f_l^{img} is more similar to f_l^{p} of the same object than to other objects. Inspired by instance discrimination, we introduce a novel CLC loss that maximizes the similarity between f_l^{img} and f_l^{p} while simultaneously minimizing the similarity between f_l^{p} and representations of other objects. Mathematically, this loss can be expressed as:

$$s(a,b) = exp(sim(a,b)/\tau), \tag{1}$$

$$c(x,y,i) = \frac{s(x_i,y_i)}{\sum_{k=1,k\neq i}^{N} s(x_i,x_k) + \sum_{k=1}^{N} s(x_i,y_k)}, \tag{2}$$

$$L_{CLC} = -\frac{1}{2N}\left[\sum_{i=1}^{N} log(c(f_l^{img}, f_l^{p}, i) + c(f_l^{p}, f_l^{img}, i))\right], \tag{3}$$

where N and τ are the numbers of local representations and the temperature coefficient, $sim(\cdot)$ denotes the cosine similarity function.

CLGC Loss. In this subsection, we introduce a novel loss for modeling local-to-global correspondence across modalities. As noted by [39], a semantic correlation exists between local and global features of 3D point clouds. We extend this observation to cross-modal learning, leveraging the CLGC Loss. This loss relies on two perspectives: Firstly, we can consider the local features and their corresponding cross-modal global features as positive samples and utilize the CLGC loss constraint to promote the transfer of latent semantics. Secondly, we can view the constraint process of the CLGC loss as reasoning from local to global, thereby compelling the model to acquire shared semantic information across modalities. Further, Our ablation experiments provide evidence for the efficacy of the CLGC Loss.

Specifically, the intermediate representations of the point-cloud modality are constructed in the same manner described above, which we denote as f_l^{p}. Meanwhile, the global representations are extracted from the image encoders and denoted as g^{img}. We then aim to maximize the local-to-global similarity within the same batch while minimizing the similarity between batches to learn shared attributions. Therefore, our method enables the modeling of cross-modal local-to-global correspondence, which can be expressed as follows:

$$L_{CLGC} = -\frac{1}{2N}\sum_{i=1}^{N} log\,(c\,(\,g^{img}, f_l^{p}, i)\,). \tag{4}$$

CGC Loss. In addition to the multi-scale correspondence learning discussed in the previous sections, we introduce a general auxiliary contrastive objective promoting high-level semantic correspondence across modalities. As mentioned before, the global representations $\{g^{p}, g^{img}\}$ of point-cloud and image modalities are utilized. We define the cross-modal global representation of the same object in the shared space as a positive sample. By incorporating harder negative samples (the representations from different modalities) in our approach, we aim to enhance the representational capability, as demonstrated in previous works[2][66]. Consequently, our cross-modal global-to-global correspondence can be expressed as follows:

$$L_{CGC} = -\frac{1}{2N}\sum_{i=1}^{N} log\,(c\,(\,g^{img}, g^{p}, i)\,). \tag{5}$$

Feature Align Loss. We note that while the CLGC loss can ensure the similarity of local features with global representations of another modality, it may not be sufficient to guarantee the quality of the local features. To further enhance the quality of cross-modal local representations, imposing additional constraints on cross-modal local representations are necessary. We utilize self-reconstruction tasks as a reliable solution, as they are general low-level generation tasks that optimize the quality of local representations during the reconstruction process and can implicitly align

cross-modal local representations in PointCMC.

After reconstructing the fused representations into the 3D space using the cross-modal point generator, we obtain reconstructive point clouds P_{recon}. To ensure high-quality local features, we impose the Earth Mover's Distance (EMD) constraint between P_{recon} and the ground-truth point clouds P_{gt}. This constraint can be expressed as follows:

$$L_{recon} = \min_{\phi: S_1 \to S_2} \sum_{p \in S_1}^{N} ||p - \phi(p)||_2, \tag{6}$$

where $S_1 = P_{recon}$, $S_2 = P_{gt}$ and $\phi : S_1 \to S_2$ is a bijection.

Overall Objective. Our ultimate joint learning objective combines the losses from the preceding four sections to facilitate the multi-scale 2D-3D correspondences, where $\alpha = 1,\ \beta = 1,\ and\ \gamma = 10$ are the weights to adjust the ratios of each loss, respectively..

$$L_{final} = L_{CLC} + \alpha L_{CLGC} + \beta L_{CGC} + \gamma L_{recon}. \tag{7}$$

4. Experiments

To assess the effectiveness of our model in learning point cloud representations, we conduct experiments on three widely adopted downstream tasks. In addition, we investigate the impact of our proposed module and losses in ablation studies. Specially, we introduce our pre-training setting (Sec 4.1), followed by the presentation of our experimental results on the downstream tasks (Sec 4.2), which demonstrates the efficacy of our proposed model. Lastly, we report the results of our ablation studies (Sec 4.3).

4.1. Pre-Training

Dataset. For pre-training, we utilize the ShapeNet[4], which consists of over 50,000 CAD models across 55 classes, as our point cloud dataset. We also employ an image dataset[58] that contained 43,783 images of 13 classes rendered from ShapeNet at random angles. To maintain point cloud-image pairs, we limite our selection to the 13 classes in common between the two datasets. Our input point clouds are sampled to contain 2048 points, while corresponding rendered images are resized to $(224, 224)$. To enhance the robustness of our model, we apply a series of data augmentation techniques during the pre-training stage. Specifically, we utilize scaling, translation, and rotation transformations for the point cloud data, and apply random crop, color jittering, and random horizontal flips to the image data.

Implementation Details. Our proposed PointCMC is implemented by PyTorch[33] framework and a single GTX3090 GPU. To ensure a fair comparison with existing methods, we select DGCNN[50] and RSCNN[28] as the point cloud encoders and Swin Transformer[31] as the image encoder. The intermediate representations of the point cloud are shaped as $\{B, C, N_1\}$, where B denotes the batch size, C represents the feature dimension, and N_1 represents the number of point groups. Similarly, the intermediate image representations are structured as $\{B, C, N_2\}$, where B, C have the same meaning as above, and N_2 represents the number of image patches. These local representations are then fed into the L2L module, which includes 6-layer intra-attention blocks, 6-layer co-attention blocks, 3-layer merge-attention blocks, and a simple MLP-based FAB module. We utilize the Adam optimizer with a decay rate of 1×10^{-4} and a learning rate of 1×10^{-3}. In addition, we also apply the Cosine annealing[32] as the learning rate scheduler and train the model for 100 epochs. After pre-training, all downstream tasks are performed based on the point cloud encoder E^p.

Table 1: **Linear classification results on the ModelNet40 dataset.** After training the model, we evaluate its performance by fitting a linear SVM classifier onto the ModelNet40 test dataset and reporting the overall accuracy(%). Notably, our proposed PointCMC method outperforms previous self-supervised approaches on both DGCNN and RSCNN backbones.

Method	ModelNet40
3D-GAN [51]	83.3
Latent-GAN [1]	85.7
SO-Net [23]	87.3
FoldingNet [61]	88.4
MRTNet [9]	86.4
3D-PointCapsNet [67]	88.9
DepthContrast [66]	85.4
ClusterNet [64]	86.8
VIP-GAN [12]	90.2
DGCNN + Multi-Task [14]	89.1
DGCNN + Self-Contrast [7]	89.6
DGCNN + Jigsaw [41]	90.6
DGCNN + STRL [16]	90.9
DGCNN + Rotation [34]	90.8
DGCNN + OcCo [48]	89.2
DGCNN + CrossPoint [2]	91.2
DGCNN + PointCMC	**92.2**
RSCNN + GLR [39]	89.5
RSCNN + CrossPoint [2]	91.5
RSCNN + PointCMC	**92.4**

4.2. Downstream Tasks

Our method is evaluated extensively in three downstream tasks, utilizing pre-trained point cloud encoders. These tasks include 3D object classification, few-shot object classification, and 3D object part segmentation. Our approach's performance is evaluated using two key metrics, namely

Table 2: **Linear classification results on the ScanObjectNN dataset.** PointCMC also surpasses existing works in both DGCNN and RSCNN backbones. This shows that our method is still efficient in the real-world dataset.

Method	ScanObjectNN
DGCNN + Jigsaw [41]	59.5
DGCNN + OcCo [48]	78.3
DGCNN + STRL [16]	77.9
DGCNN + CrossPoint [2]	81.7
DGCNN + PointCMC	**84.4**
RSCNN + GLR [39]	80.3
RSCNN + CrossPoint [2]	84.6
RSCNN + PointCMC	**85.1**

overall classification accuracy (OA) and mean intersection over union (mIoU).

Transfer to 3D object classification. We compare the OA of our method with state-of-the-art methods on the ModelNet40 (synthetic) and ScanObjectNN (real-world) benchmarks. The ModelNet40 dataset includes 40 classes with 12,331 3D CAD models, of which 9,843 are in the training data splits and 2,468 are in the test data splits. The ScanObjectNN dataset contains 2,880 objects in 15 classes, extracted from real-world indoor scene scans, with 2,304 objects in the training data splits and 576 in the test data splits. We sample 1,024 points per object and use DGCNN and RSCNN backbones for classification. Following the same setup as in previous work[2][48], we freeze the point cloud encoder and train a simple SVM classifier on the downstream task training data splits.

Our proposed method achieves state-of-the-art results on both point cloud benchmark datasets, as demonstrated in Table 1 for ModelNet40 and Table 2 for ScanObjectNN. However, other methods may not be directly comparable to ours due to differences in pre-training methods and the selection of backbone networks. Despite this limitation, our approach shows great performance in self-supervised contrastive learning.

Transfer to Few-shot object classification. To evaluate the effectiveness of our method under the constraint of limited fine-tuning data, we conduct few-shot object classification experiments on two benchmark datasets: ModelNet40 and ScanObjectNN. In few-shot classification, N represents the number of classes, while K represents the number of samples in each class. Following the settings in[2][48], we compare our method with four configurations. The experimental results in Table 3 for ModelNet40 and Table 4 for ScanObjectNN demonstrate that our proposed approach outperforms the previous state-of-the-art method in three out of four settings for both DGCNN and RSCNN backbones. While our method does not achieve the highest rank in some settings, the margin with the state-of-the-art method is relatively small. This observation highlights that our approach can learn more generalized representations.

Table 3: **Few-shot classification results on ModelNet40 dataset.** We report the mean accuracy(%) and standard deviation(%) over ten runs on independent experiments.

Method	5 way		10 way	
	10 shot	20 shot	10 shot	20 shot
3D-GAN [51]	55.8±3.4	65.8±3.1	40.3±2.1	48.4±1.8
FoldingNet [61]	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 [67]	42.3±5.5	53.0±5.9	38.0±4.5	27.2±4.7
PointNet++ [37]	65.4±2.8	68.6±2.2	46.6±1.5	50.0±2.3
PointCNN [26]	65.4±2.8	68.6±2.2	46.6±1.5	50.0±2.3
RSCNN [28]	65.4±8.9	68.6±7.0	46.6±4.8	50.0±7.2
DGCNN + Rand	31.6±2.8	40.8±4.6	19.9±2.1	16.9±1.5
DGCNN + Jigsaw [41]	34.3±1.3	42.2±3.5	26.0±2.4	29.9±2.6
DGCNN + cTree [43]	60.0±2.8	65.7±2.6	48.5±1.8	53.0±1.3
DGCNN + OcCo [48]	90.6±2.8	92.5±1.9	82.9±1.3	86.5±2.2
DGCNN + CrossPoint [2]	**92.5±3.0**	94.9±2.1	83.6±5.3	87.9±4.2
DGCNN + PointCMC	92.2±5.0	**95.5±3.3**	**87.5±5.1**	**91.4±3.0**
RSCNN + Rand	40.2±2.9	49.8±3.2	29.6±2.4	28.3±3.4
RSCNN + GLR [39]	91.5±4.9	94.7±3.6	86.6±4.0	88.9±4.5
RSCNN + CrossPoint [2]	**93.9±4.2**	95.6±4.0	89.8±4.1	92.5±3.6
RSCNN + PointCMC	93.5±5.1	**95.8±3.4**	**90.2±4.3**	**92.9±3.3**

Table 4: **Few-shot classification results in ScanObjectNN dataset.** We report the mean accuracy(%) and standard deviation(%) over ten runs on independent experiments.

Method	5 way		10 way	
	10 shot	20 shot	10 shot	20 shot
DGCNN + Rand	62.0±5.6	67.8±5.1	37.8±4.3	41.8±2.4
DGCNN + Jigsaw [41]	65.2±3.8	72.2±2.7	45.6±3.1	48.2±2.8
DGCNN + cTree [43]	68.4±3.4	71.6±2.9	42.4±2.7	43.0±3.0
DGCNN + OcCo [48]	72.4±1.4	77.2±1.4	57.0±1.3	61.6±1.2
DGCNN + CrossPoint [2]	74.8±1.5	79.0±1.2	62.9±1.7	73.9±2.2
DGCNN + PointCMC	**78.3±6.8**	**84.4±5.9**	**68.6±4.2**	**76.3±3.8**
RSCNN + Rand	69.2±5.6	73.3±6.3	43.7±5.1	48.8±4.6
RSCNN + GLR [39]	77.2±7.2	83.4±5.7	65.2±4.9	72.0±4.4
RSCNN + CrossPoint [2]	83.5±6.7	88.3±4.3	**78.8±4.3**	79.6±3.8
RSCNN + PointCMC	**84.0±5.7**	**88.4±4.8**	78.4±4.1	**80.2±3.6**

Transfer to 3D Object part segmentation. We evaluate our method on the ShapeNetPart[62] benchmark dataset for 3D object part segmentation. The dataset consists of 16 classes, comprising 16,881 3D objects, each annotated with 50 parts. Part segmentation requires classifying each point, making it more challenging than classifying the entire point cloud object as it involves capturing local patterns. We use the DGCNN branch for pre-training on ShapeNet and evaluate the effectiveness of our method using a simple segmentation head. The features of the three abstract layers are labeled with the encoder and then concatenated onto each point feature after forward propagating through the DGCNN to predict the class of each point using a linear projection layer. In Table 5, we compare supervised and unsupervised methods, showing that the unsupervised DGCNN pre-trained by PointCMC performs 0.8% better than the supervised DGCNN with random weight initialization, indicates that PointCMC provides better weight initialization to the backbone. PointCMC outperforms the pre-

vious best method by 0.4%, indicating that its multi-scale learning strategy enables the encoder to capture more fine-grained features than other self-supervised methods. As shown in Figure 5, PointCMC demonstrates superior performance in segmenting certain intricate areas.

Table 5: **Part Segmentation result on ShapeNetPart dataset.** 'mIoU' denotes the mean intersection over union across all object classes in the dataset. Compared with the current supervised and self-supervised methods, PointCMC can achieve the best performance.

Category	Method	mIoU(%)
Supervised	PointNet [35]	83.7
	PointNet++ [37]	85.1
	DGCNN [50]	85.1
Unsupervised	Self-Contrast [7]	82.3
	Jigsaw [41]	85.3
	OcCo [48]	85.0
	PointContrast [56]	85.1
	Liu et al. [27]	85.3
	CrossPoint [2]	85.5
	PointCMC	**85.9**

4.3. Ablation study

We conduct ablation studies on DGCNN and RSCNN to demonstrate the effectiveness of our training strategy. Our evaluation is performed from two perspectives: (1) the impact of modules, and (2) the impact of the number of corresponding 2D images.

Impact of modules. We establish multi-scale cross-modal correspondences by mapping them to the same feature space, equal to creating positive samples in contrastive learning. We hypothesize that enhancing multi-scale cross-modal correspondences can facilitate cross-modal transfer more effectively than single-scale correspondences. To verify our hypothesis, we train networks with single-scale correspondences and evaluate their classification performance using a linear SVM classifier on the ModelNet40 and ScanObjectNN datasets. Our multi-scale network increases the overall accuracy by 0.9% and 1% over the second-best approach on ModelNet40 using DGCNN and RSCNN feature extractors, respectively (Table 6). The L2L-only module network exhibits the best classification results, possibly because embedding local features of the images near the point cloud features promotes fine-grained semantic transformation. We visualize the T-SNE[46] plots of the ablation experiments on the ModelNet10 test splits in Figure 4, which reveal that multi-scale networks better distinguish categories with indistinct boundaries, such as tables and chairs, compared to single-scale networks.

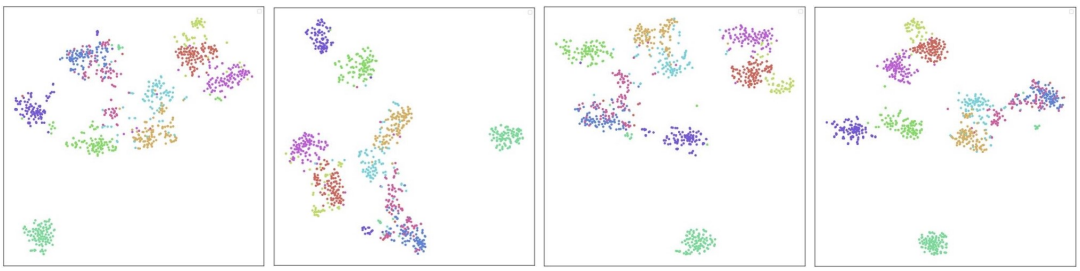
Table 6: **Linear classification results on ModelNet40 with different modules,** where A denotes the DGCNN feature extractor and B denotes the RSCNN feature extractor. Our results indicate that the multi-scale network outperforms the single-scale network in both backbones. Note that all other experimental settings remain consistent across different ablation studies.

Backbones	L2L	L_{CLGC}	L_{CGC}	accuracy(%)
A	✓			91.3
A		✓		89.5
A			✓	90.6
A	✓	✓	✓	**92.2**
B	✓			91.4
B		✓		89.9
B			✓	90.8
B	✓	✓	✓	**92.4**

Table 7: **Linear classification results on ModelNet40 with different attention blocks,** where A represents the DGCNN feature extractor and B represents the RSCNN feature extractor. In both backbones, the network within intra- and inter-attention blocks outperforms the single-block network.

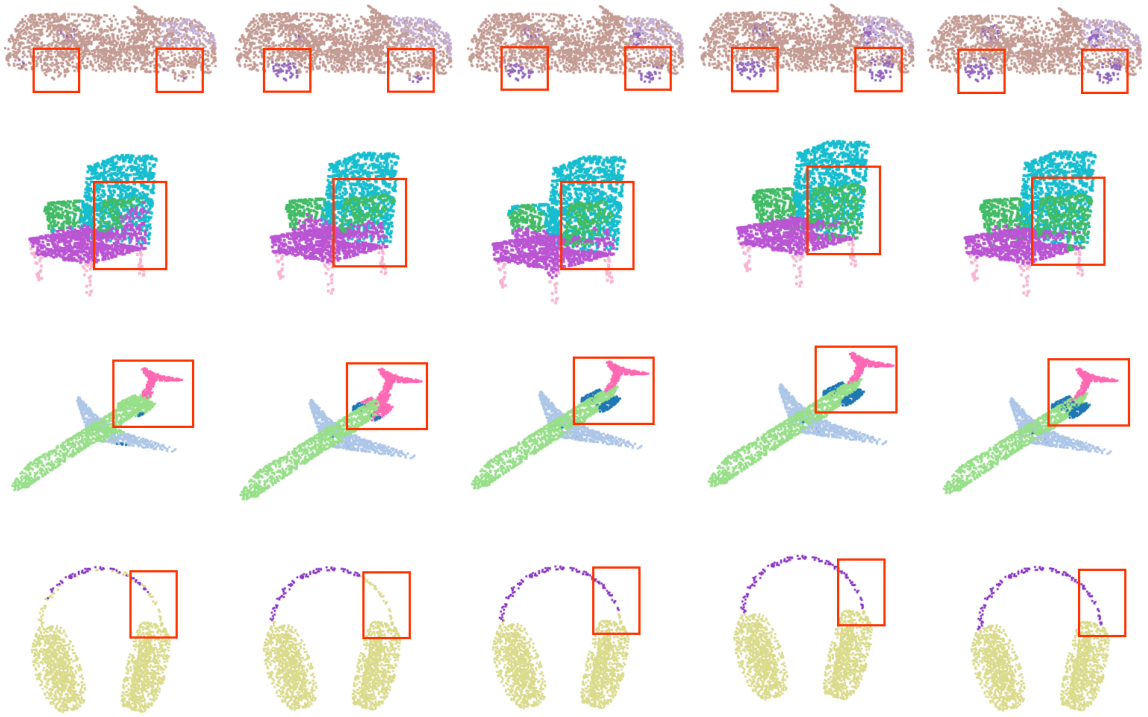
Backbones	intra	inter	accuracy(%)
A	✓		91.5
A		✓	91.9
A	✓	✓	**92.2**
B	✓		91.5
B		✓	92.0
B	✓	✓	**92.4**

Furthermore, we investigate the impact of each block within the L2L module on our approach. Specifically, we explore the effects of the intra-attention and inter-attention blocks. Our results, presented in Table 7, demonstrate that employing only the intra-attention block results in lower linear classification accuracy than using only the inter-attention block. However, combining both blocks achieves the highest classification accuracy. Additionally, we evaluate the effects of the co-attention and merge-attention mechanisms (Table 8). The results show that employing only the merge-attention mechanism leads to lower linear classifica-



(a) L2L Only (b) CLGC Only (c) CGC Only (d) Multi-scale

Figure 4: **T-SNE [46] visualization on the ModelNet10 test dataset.** We show the feature distribution extracted by different module networks: (a) L2L-only network; (b) CLGC-only network; (c) CGC-only network; and (d) Multi-scale network. Our proposed Multi-scale network can better distinguish between different classes than single-scale networks.



(a) Self-Contrast (b) OcCo (c) CrossPoint (d) PointCMC (e) GT

Figure 5: **Visualization of part segmentation results.** We visualize the part segmentation results from (a) Self-Contrast; (b) OcCo; (c) CrossPoint; (d) PointCMC; and (e) Ground-Truth(GT).

tion accuracy than using only the co-attention mechanism. Nevertheless, the highest classification accuracy is achieved when both mechanisms are jointly employed. These findings indicate that appropriately suppressing unimodal interaction and enhancing intermodal interaction can facilitate knowledge transfer.

Impact of images. We investigate the effect of the number of rendered images on our network by randomly selecting n images from different viewpoints and computing the average of their features for loss calculation. Table X shows the network's results with different numbers of images on the linear SVM classifier. Our findings suggest that the backbone achieves the best result when using only one image as input. We attribute the decrease in classification accuracy to the redundant information captured from multiple rendered images of the same object in the image modality.

Table 8: **Linear classification results on ModelNet40 with different attention mechanisms,** where A represents the DGCNN feature extractor and B represents the RSCNN feature extractor.

Backbones	co-att	merge-att	accuracy(%)
A	✓		92.0
A		✓	91.7
A	✓	✓	**92.2**
B	✓		92.1
B		✓	91.9
B	✓	✓	**92.4**

Table 9: **Linear classification results on ModelNet40 with different numbers of rendered images.** PointCMC performs better with one rendered image than with multiple rendered images. We choose one image for all experiences.

Numbers of rendered images	1	2	3	4	5
Linear accuracy(%)	**92.2**	92.1	91.8	91.8	91.5

5. Conclusion

In this paper, we propose PointCMC, a novel self-supervised training strategy for point cloud representation learning. Our downstream experiments demonstrate that enforcing cross-modal correspondences improves point cloud representations. We also validate our hypothesis that multi-scale correspondences enable the model to achieve the best performance through our ablation experiments. In future works, it would be interesting to investigate transferring the learned image knowledge to more efficient models and applying our method to point cloud-based tasks such as segmentation and detection. Moreover, leveraging point cloud geometric knowledge as auxiliary inputs for image learning is a promising direction for future research.

References

[1] P. Achlioptas, O. Diamanti, I. Mitliagkas, and L. Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 1, 3, 6, 7

[2] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9902–9912, 2022. 1, 2, 3, 5, 6, 7, 8

[3] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. 3

[4] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6

[5] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton. A simple framework for contrastive learning of visual representations. *international conference on machine learning*, 2020. 3

[6] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 3

[7] B. Du, X. Gao, W. Hu, and X. Li. Self-contrastive learning with hard negative sampling for self-supervised point cloud learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3133–3142, 2021. 3, 6, 8

[8] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017. 3

[9] M. Gadelha, R. Wang, and S. Maji. Multiresolution tree networks for 3d point cloud processing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 6

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, pages 139–144, 2020. 3

[11] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[12] Z. Han, M. Shang, Y.-S. Liu, and M. Zwicker. View inter-prediction gan: Unsupervised representation learning for 3d shapes by learning global shape memories to support local view predictions. *national conference on artificial intelligence*, 2019. 6

[13] Z. Han, X. Wang, Y.-S. Liu, and M. 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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10441–10450. IEEE, 2019. 3

[14] K. Hassani and M. Haley. Unsupervised multi-task feature learning on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8160–8171, 2019. 6

[15] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[16] S. Huang, Y. Xie, S.-C. Zhu, and Y. Zhu. Spatio-temporal self-supervised representation learning for 3d point clouds. *international conference on computer vision*, 2021. 3, 6, 7

[17] Z. Huang, Y. Yu, J. Xu, F. Ni, and X. Le. Pf-net: Point fractal network for 3d point cloud completion. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020. 3

[18] L. Jing, L. Zhang, and Y. Tian. Self-supervised feature learning by cross-modality and cross-view correspondences. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[19] L. Jing, L. Zhang, and Y. Tian. Self-supervised feature learning by cross-modality and cross-view correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1591, 2021. 2

[20] M. A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, pages 233–243, 1991. 3

[21] T. Le and Y. Duan. Pointgrid: A deep network for 3d shape understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9204–9214, 2018. 2

[22] C.-L. Li, M. Zaheer, Y. Zhang, B. Poczos, and R. Salakhutdinov. Point cloud gan. *arXiv preprint arXiv:1810.05795*, 2018. 3

[23] J. Li, B. M. Chen, and G. H. Lee. So-net: Self-organizing network for point cloud analysis. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 1, 3, 6

[24] R. Li, X. Li, C.-W. Fu, D. Cohen-Or, and P.-A. Heng. Pu-gan: A point cloud upsampling adversarial network. *international conference on computer vision*, 2019. 3

[25] R. Li, X. Li, P.-A. Heng, and C.-W. Fu. Point cloud upsampling via disentangled refinement. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 3

[26] Y. Li, R. Bu, M. Sun, W. Wu, X. Di, and B. Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018. 1, 7

[27] F. Liu, G. Lin, and C.-S. Foo. Point discriminative learning for unsupervised representation learning on 3d point clouds. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 8

[28] Y. Liu, B. Fan, S. Xiang, and C. Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8895–8904, 2019. 1, 3, 6, 7

[29] Y. Liu, L. Yi, S. Zhang, Q. Fan, T. Funkhouser, and H. Dong. P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding. *arXiv: Computer Vision and Pattern Recognition*, 2020. 2

[30] Y.-C. Liu, Y.-K. Huang, H.-Y. Chiang, H.-T. Su, Z.-Y. Liu, C.-T. Chen, C.-Y. Tseng, and W. H. Hsu. Learning from 2d: Contrastive pixel-to-point knowledge transfer for 3d pretraining. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[32] I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. *neural information processing systems*, 2019. 6

[34] O. Poursaeed, T. Jiang, H. Qiao, N. Xu, and V. G. Kim. Self-supervised learning of point clouds via orientation estimation. *international conference on 3d vision*, 2020. 6

[35] C. R. Qi, H. Su, K. Mo, and L. 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*, pages 652–660, 2017. 1, 2, 8

[36] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5648–5656, 2016. 2

[37] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 1, 2, 7, 8

[38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *international conference on machine learning*, 2021. 3

[39] Y. Rao, J. Lu, and J. Zhou. Global-local bidirectional reasoning for unsupervised representation learning of 3d point clouds. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2020. 2, 3, 5, 6, 7

[40] A. Sanghi. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 626–642. Springer, 2020. 3

[41] J. Sauder and B. Sievers. Self-supervised deep learning on point clouds by reconstructing space. *neural information processing systems*, 2019. 6, 7, 8

[42] A. Sharma, O. Grau, and M. Fritz. Vconv-dae: Deep volumetric shape learning without object labels. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3

[43] C. Sharma and M. Kaul. Self-supervised few-shot learning on point clouds. *neural information processing systems*, 2020. 7

[44] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 3

[45] D. Valsesia, G. Fracastoro, and E. Magli. Learning localized generative models for 3d point clouds via graph convolution.

international conference on learning representations, 2018. 3

[46] L. van der Maaten and G. E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 2008. 8, 9

[47] B. Wang, C. Chen, Z. Cui, J. Qin, C. X. Lu, Z. Yu, P. Zhao, Z. Dong, F. Zhu, N. Trigoni, et al. P2-net: Joint description and detection of local features for pixel and point matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16004–16013, 2021. 2

[48] H. Wang, Q. Liu, X. Yue, J. Lasenby, and M. J. Kusner. Unsupervised point cloud pre-training via occlusion completion. *international conference on computer vision*, 2021. 3, 6, 7, 8

[49] P.-S. Wang, Y.-Q. Yang, Q.-F. Zou, Z. Wu, Y. Liu, and X. Tong. Unsupervised 3d learning for shape analysis via multiresolution instance discrimination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2773–2781, 2021. 1

[50] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 1, 2, 6, 8

[51] J. Wu, C. Zhang, T. Xue, W. T. Freeman, and J. B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *neural information processing systems*, 2016. 6, 7

[52] W. Wu, Z. Qi, and L. Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 9621–9630, 2019. 1

[53] Z. Wu, Y. Zhang, M. Zeng, F. Qin, and Y. Wang. Joint analysis of shapes and images via deep domain adaptation. *Computers & Graphics*, 70:140–147, 2018. 3

[54] A. Xiao, J. Huang, D. Guan, and S. Lu. Unsupervised representation learning for point clouds: A survey. *arXiv preprint arXiv:2202.13589*, 2022. 3

[55] J. Xie, Z. Zheng, R. Gao, W. Wang, S.-C. Zhu, and Y. N. Wu. Learning descriptor networks for 3d shape synthesis and analysis. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3

[56] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020. 1, 3, 8

[57] C. Xu, S. Yang, B. Zhai, B. Wu, X. Yue, W. Zhan, P. Vajda, K. Keutzer, and M. Tomizuka. Image2point: 3d point-cloud understanding with pretrained 2d convnets. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[58] Q. Xu, W. Wang, D. Ceylan, R. Mech, and U. Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. *neural information processing systems*, 2019. 6

[59] Y. Xu, T. Fan, M. Xu, L. Zeng, and Y. Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European conference on computer vision (ECCV)*, pages 87–102, 2018. 1

[60] X. Yan, H. Zhan, C. Zheng, J. Gao, R. Zhang, S. Cui, and Z. Li. Let images give you more: Point cloud cross-modal training for shape analysis. *arXiv preprint arXiv:2210.04208*, 2022. 3, 4

[61] Y. Yang, C. Feng, Y. Shen, and D. Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 206–215, 2018. 3, 6, 7

[62] L. Yi, V. G. Kim, D. Ceylan, I.-C. Shen, M. Yan, H. Su, C. Lu, Q. Huang, A. Sheffer, and L. J. Guibas. A scalable active framework for region annotation in 3d shape collections. *international conference on computer graphics and interactive techniques*, 2016. 7

[63] L. Yu, X. Li, C.-W. Fu, D. Cohen-Or, and P.-A. Heng. Pu-net: Point cloud upsampling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2790–2799, 2018. 3

[64] L. Zhang and Z. Zhu. Unsupervised feature learning for point cloud understanding by contrasting and clustering using graph convolutional neural networks. *international conference on 3d vision*, 2019. 1, 3, 6

[65] R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1058–1067, 2017. 3

[66] Z. Zhang, R. Girdhar, A. Joulin, and I. 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. 5, 6

[67] Y. Zhao, T. Birdal, H. Deng, and F. Tombari. 3d point capsule networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3, 6, 7

[68] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3