MM-Point: Multi-View Information-Enhanced Multi-Modal Self-Supervised 3D Point Cloud Understanding

Hai-Tao Yu[1,3], **Mofei Song**[2,3] *

[1] School of Cyber Science and Engineering, Southeast University, Nanjing, China
[2] School of Computer Science and Engineering, Southeast University, Nanjing, China
[3] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
{yuht, songmf}@seu.edu.cn

Abstract

In perception, multiple sensory information is integrated to map visual information from 2D views onto 3D objects, which is beneficial for understanding in 3D environments. But in terms of a single 2D view rendered from different angles, only limited partial information can be provided. The richness and value of Multi-view 2D information can provide superior self-supervised signals for 3D objects. In this paper, we propose a novel self-supervised point cloud representation learning method, **MM-Point**, which is driven by *intra-modal* and *inter-modal* similarity objectives. The core of MM-Point lies in the Multi-modal interaction and transmission between 3D objects and multiple 2D views at the same time. In order to more effectively simultaneously perform the consistent cross-modal objective of 2D multi-view information based on contrastive learning, we further propose ***Multi-MLP*** and ***Multi-level Augmentation*** strategies. Through carefully designed transformation strategies, we further learn Multi-level invariance in 2D Multi-views. MM-Point demonstrates state-of-the-art (**SOTA**) performance in various downstream tasks. For instance, it achieves a peak accuracy of 92.4% on the synthetic dataset ModelNet40, and a top accuracy of 87.8% on the real-world dataset ScanObjectNN, comparable to fully supervised methods. Additionally, we demonstrate its effectiveness in tasks such as few-shot classification, 3D part segmentation and 3D semantic segmentation.

Introduction

In recent years, the demand for 3D perception technologies has surged in the real world. As a fundamental 3D representation, point cloud learning plays a crucial role in numerous tasks, including 3D object *classification*, *detection*, and *segmentation*. However, the cost of point cloud annotation is high, and 3D scans with labels are usually scarce in reality. To address these challenges, many studies have turned their attention to self-supervised methods.

Interestingly, while 3D point clouds can be obtained by sampling, 2D multi-view images can be generated by rendering. Some recently proposed methods (Yang et al. 2020; Huang et al. 2020) have focused on generating high-quality point cloud representations using multi-modal data.

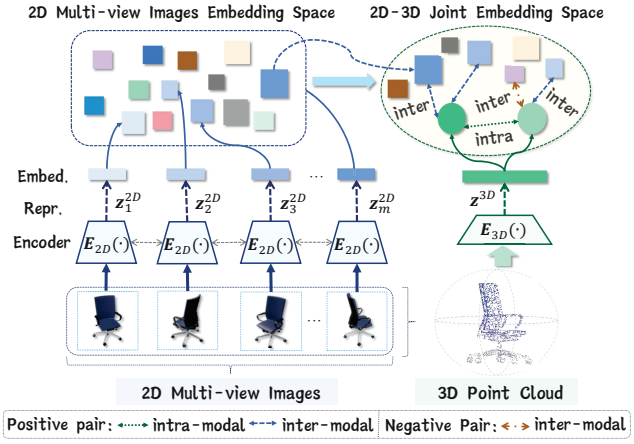

Figure 1: Schematic of MM-Point multi-modal contrastive learning. Given 2D multi-views rendered from a 3D object, the model distinguishes the views of the same object from those of different objects in the embedding space, enabling self-supervised learning of deep representation.

This naturally raises a question: Can we better facilitate the understanding of 3D point cloud representations by leveraging the abundant information hidden in 2D Multi-views? We reconsider the contrast paradigm in 2D-3D, hypothesizing that 3D objects and rendered 2D Multi-views share mutual information. Our primary motivation is to view each 2D view as a unique pattern that is useful for guiding 3D objects and improving their representations, as each 2D view observes different aspects of 3D objects and representations from different angles are distinctive. From another perspective, we encourage the 3D-2D relationship to be consistent across different-angle views, *i.e.* the similarity correspondence between the 2D images and 3D point cloud in one view should also exist in another view.

Due to the distinct nature of 3D representations from 2D visual information, a simple alignment between these two types of representations may lead to limited gain or even negative transfer in multi-modal learning. Therefore, we propose a novel 3D point cloud representation learning framework and a multi-modal pre-training strategy between 3D and 2D, namely **MM-Point** (as shown in Fig.1), which

*Corresponding Author

transfers the features extracted from 2D multi-views into 3D representations. Specifically, *intra-modal* training is used to capture the intrinsic patterns of 3D augmented data samples. The *inter-modal* training scheme aims to learn point cloud representations by accepting 2D-3D interactions. Meanwhile, considering the differences in multi-modal data properties, We further ponder: how to effectively simultaneously transfer multiple 2D view information from different angles to 3D point cloud objects in a self-supervised manner?

Extending this, we propose a **Multi-MLP** strategy to construct multi-level feature learning between each 2D view and 3D object, thus the consistency goal is set to contrast between 2D multi-views and 3D objects across different feature spaces. Such architectural design not only extracts shared information in 2D-3D multi-modal contrast, but also preserve specific information in multiple 2D views at the same time, thereby the overall consistency goal could better extract semantic information between 3D point clouds and as many different 2D views as possible, resulting in improved 3D representations.

Additionally, treating each angle of 2D view equally during training with the same type of augmentation transformation may lead to a suboptimal representation for downstream tasks in multimodal contrast. We suggest a **Multi-level Augmentation** strategy based on multi-view, integrating rendered 2D multi-view information with augmentation information. Moreover, we control the strength of all augmentation modules, ensuring the mutual information between 3D point cloud and 2D image augmentation pairs remains low and within a certain range, thus enabling the model to gradually accumulate more complex information during learning.

To evaluate our proposed 2D-3D multi-modal self-supervised learning framework, we assess the performance of MM-Point across several downstream tasks. The learned 3D point cloud representation can be directly transferred to these tasks. For 3D object shape classification, we performed on both synthetic dataset ModelNet40 and real-world object dataset ScanObjectNN, achieving state-of-the-art performance, surpassing all existing self-supervised methods. Furthermore, part segmentation and semantic segmentation experiments validate MM-Point's capability to capture fine-grained features of 3D point clouds.

In summary, our research contributions are as follows:

- We introduce **MM-Point**, a novel 3D representation learning scheme based on 2D-3D multi-modal training.
- Our research applies multi-modal contrastive learning to multi-view settings, maximizing shared mutual information between different 2D view and the same 3D object.
- We propose **Multi-MLP** and **Multi-level Augmentation** strategies, thereby ensuring more effective learning of 3D representation in multi-modal contrast under a self-supervised setting, achieving effective pre-training from 2D multi-views to 3D objects.
- **MM-Point** demonstrates remarkable transferability. The pre-trained 3D representations can be directly transferred to numerous downstream tasks, achieving state-of-the-art performance in extensive experiments.

Related Work

Self-supervised Point Cloud Learning

Unsupervised learning for point cloud understanding can be broadly divided into generative or discriminative tasks based on the proxy tasks. Generative models learn features by self-reconstruction, as exemplified by methods like JigSaw (Qi et al. 2018) . Furthermore, Point-BERT (Yu et al. 2022) predicts discrete labels, while Point-MAE (Pang et al. 2022) randomly masks patches of the input point cloud and reconstructs the missing parts. However, these methods are computationally expensive. On the other hand, discriminative methods predict or discriminate the enhanced versions of the inputs. Our work learns point cloud representations based on this approach. Recent works (You et al. 2021; Sun et al. 2021) have explored contrastive learning for point clouds following the success of image contrastive learning. PointContrast (Xie et al. 2020) is the first unified framework investigating 3D representation learning with a contrastive paradigm. Compared to these works, we investigate contrastive pre-training of point clouds from a new perspective, leveraging the semantic information hidden in 2D multi-views to design a multi-modal learning network, and thus enhancing the representational capacity of 3D point clouds.

Multi-modal Representation Learning

Recent studies on self-supervised learning have leveraged the multi-modal attributes of data (Caron et al. 2020a; Wang et al. 2020) , with a common strategy being the exploration of natural correspondences across differing modals, emphasizing the extraction of cross-modal shared information. For instance, in the field of visual-textual multi-modalities, large-scale image-text pairs (Li et al. 2020) have been pre-trained, enabling these models to be applicable for numerous downstream tasks. A few works have integrated point cloud representations with other modalities, such as voxels (Shi, Wang, and Li 2020; Shi, Zhou, and Li 2020) or multi-view images (Shi, Wang, and Li 2019; Qian et al. 2020) . Prior3D (Liu and Li 2020) proposed a geometric prior contrastive loss to enhance the representation learning of RGB-D data. Tran *et al.* (Tran et al. 2022) employ self-supervision through local correspondence losses and global losses based on knowledge distillation. Of note, CrossPoint (Afham 2022), most related to our work, learns 2D-3D cross-modal representations via contrastive learning, concentrating on shared features across different modes. In comparison, our approach diverges significantly. MM-Point is able to utilize information from multiple views in 2D space simultaneously for self-supervision of 3D point cloud features. By contrasting with Multi-views simultaneously, it better facilitates the maximization of mutual information shared across multi-modalities, leading to more robust and enhanced performance in various downstream tasks.

The Proposed Method

The proposed MM-Point multi-modal pre-training architecture is depicted in Fig. 2. It comprises two contrasting training strategies: *intra-modal* point cloud contrastive learning and *inter-modal* 2D-3D contrast, each accompanied by

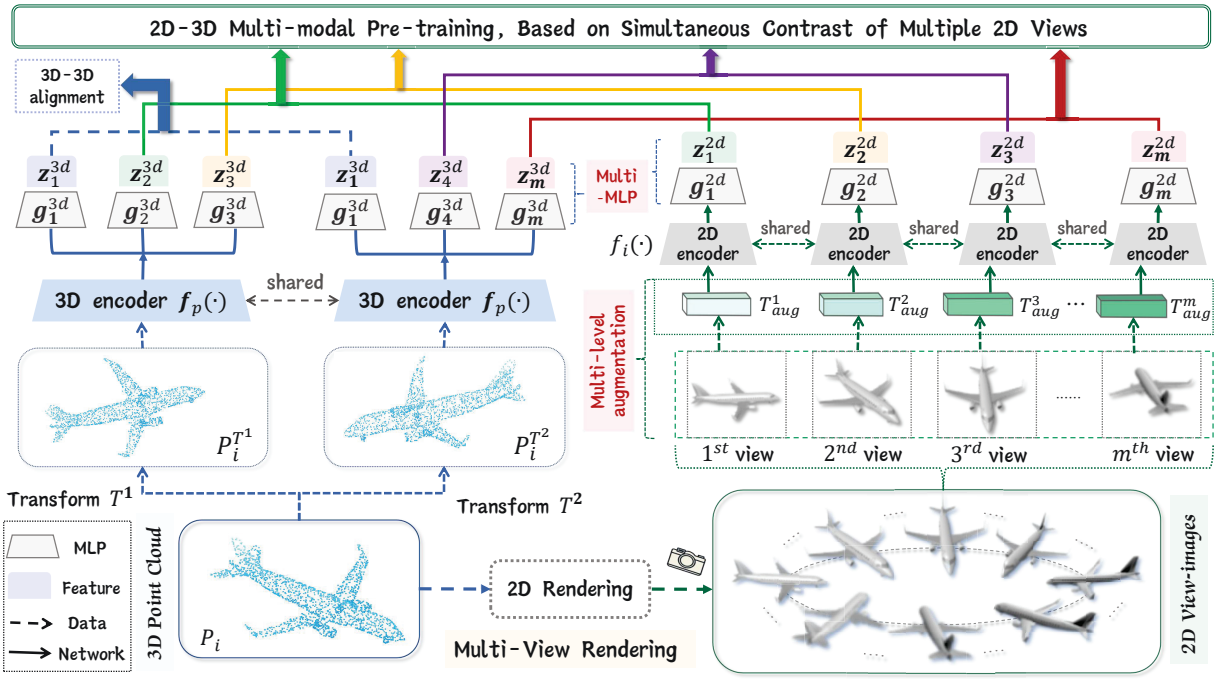

Figure 2: Schematic of the MM-Point architecture. MM-Point carries out intra-modal self-supervised learning in the 3D point cloud (blue path) and cross-modal learning between 2D and 3D (other color paths), aligning 2D multi-view features with 3D point cloud features. To better utilize the information from the 2D multi-views, MM-Point introduces two strategies: 1) Multi-MLP: constructing a multi-level feature space; 2) Multi-level augmentation: establishing multi-level invariance.

different types of loss functions. The *inter-modal* training scheme enhances 3D point cloud features by interacting with rendered 2D multi-views, while the point cloud representation focuses on the shared mutual information among multiple 2D images at the same time, inherently boosting the diversity of multi-modal contrastive learning. Furthermore, we employ a Multi-MLP strategy to project multi-view and multi-modal features. Finally, we put forward a Multi-level augmentation invariance strategy based on 2D multi-views information rendered from the same 3D object.

Intra-modal and Inter-modal Alignment

We propose to learn a encoder for aligning 3D point cloud features with visual features of 2D views. The pretraining process jointly handles the alignment of multiple modalities.

For each point cloud P_i, we obtain two variants, P_i^1 and P_i^2, through augmentation operations. We then encode the augmented point clouds separately into the feature space $F_i^1, F_i^2 \in \mathbb{R}^{n\times d}$. The projected features are subsequently mapped into the latent space, producing the representations z_i^1 and z_i^2. By performing contrastive loss, we enforce that the distance between feature representations of the same object is smaller than the distance between different objects.

MM-Point aims to learn two objectives through cross-modal alignment: $f_P(.)$ and $f_I(.)$. Cross-modal contrast seeks to minimize the distance between point clouds and the corresponding rendered 2D images while maximizing the distance from other images.

Given a sample pair (P_i, I_i), where P_i and I_i represent the embeddings of the 3D point cloud and its rendered 2D image description, respectively. For each sample P_i in the mini-batch $\mathcal{M}$, the negative sample set N_i is defined as $\mathrm{N}_i = \{I_j \mid \forall I_j \in \mathcal{M}, j \neq i\}$. The corresponding cross-modal contrastive loss on $\mathcal{M}$ is as follows:

$$Loss_{inter} = \mathbb{E}_{i\in\mathcal{M}}\left[-\log \frac{\exp\left(f_P(P_i)^T f_I(I_i)\right)}{\exp\left(f_P(P_i)^T f_I(I_i)\right)+\sum_{I_j\in \mathrm{N}_i}\exp\left(f_P(P_i)^T f_I(I_j)\right)}\right] \quad (1)$$

In addition, we map cross-modal features to different spaces and decouple training within and across modalities. We design projection heads with larger dimensions for cross-modal contrastive learning.

Multi-Modal Learning Based on Multi-View

In this section, we describe the contrast between 3D point clouds and 2D multi-view images. By pairing multiple 2D images from different angles with a 3D object, this collaboration scheme benefits the model.

Rethinking the Contrast between 3D Object and 2D Multi-view For 3D point clouds and their corresponding rendered 2D multi-view images, we simply fix the 3D point clouds and enumerate positive and negative samples from the rendered 2D view images. Suppose the loss $\mathcal{L}_{\text{contrast}}^{P_i V_i}$ treats the 3D point cloud P_i as an anchor and enumerates the 2D rendered views V_i. Symmetrically, we can obtain the loss by anchoring on the 2D V_i and enumerating the 3D P_i. Then, we sum them up as the overall contrastive loss: $\mathcal{L}(P_i, V_i) = \mathcal{L}_{\text{contrast}}^{P_i,V_i} + \mathcal{L}_{\text{contrast}}^{V_i,P_1}$.

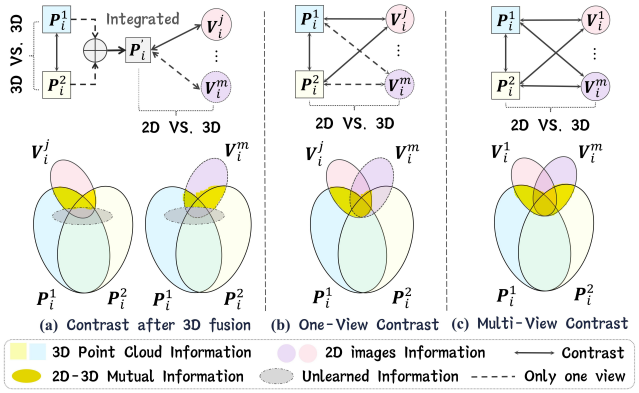

Figure 3: Mutual Information plot for the modal comparison between 3D and 2D. The yellow area signifies mutual information. The figure illustrates the impact of different strategies (a-c) on the contribution to mutual information.

Combining the theoretical proof in *CMC* (Tian, Krishnan, and Isola 2019), minimizing the loss should be equivalent to maximizing the lower bound of $I(z_i; z_j)$, that is: $I(z_i; z_j) \geq \log(k) - \mathcal{L}_{\text{contrast}}$. In this case, z_i and z_j represent the latent representations of the point cloud and image, respectively, while k denotes the number of negative sample pairs. Besides, research (Hjelm et al. 2019; Chen et al. 2021) has shown that the boundaries of $I(z_i; z_j)$ may not be clear, and finding a better mutual information estimator is more important. We consider the 3D point cloud and rendered multi-angle 2D views to construct all possible relationships between different 2D views and 3D point clouds. By involving all pairs, our optimized objective function is:

$$\mathcal{L}_F = \sum_{1 \leq j \leq M} \mathcal{L}(P_1, V_j) + \mathcal{L}(P_2, V_j) \tag{2}$$

When learning 2D-3D contrast, the mutual information will change proportionally with the number of 2D views. When the view count reaches a certain level, multi-modal mutual information will reach a stable level. In Fig.3, the visualization demonstrates that contrasting 2D Multi-views with 3D point clouds enables the point cloud features to capture more mutual information between different 2D views.

Multi-modal Contrastive based on 2D Multi-view MM-Point aims to maximize the mutual information between the differently augmented 3D point clouds and 2D views from various angles in the same scene. Naturally, too few or too many rendered 2D views exhibit poor mutual information performance, while an optimal position exists in between. We randomly sample m 2D rendered views from different angles. The value of m will be introduced in experiment.

Let $\mathcal{D}$ represent the pretraining dataset $\mathcal{D} = \left\{P_i, \{\mathbf{I}_{ij}, M_{ij}\}_{j=1}^{m}\right\}_{i=1}^{n}$, where n denotes the number of 3D objects and m represents the number of 2D views. Let P_i be the 3D point cloud of the i-th object, and I_{ij} denotes the 2D image of the j-th view of the i-th 3D object. For the multi-view 2D images I_{ij}, we extract features $\mathbf{H}_{ij}^{I} \in \mathbb{R}^{1 \times C}$,

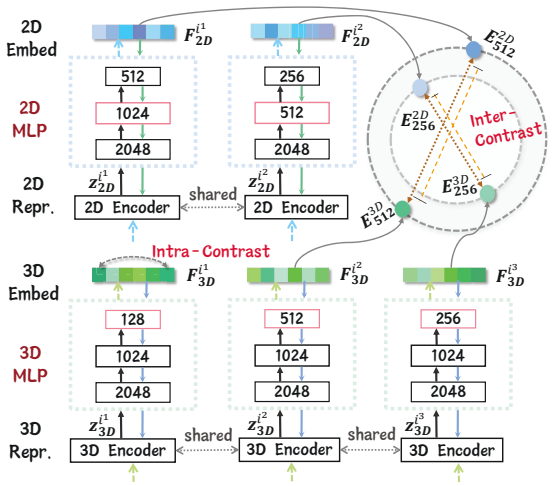

Figure 4: Schematic of the Multi-MLP strategy. Multi-modal learning enhances point cloud representation through 2D-3D interaction. Multiple 2D features and 3D features are contrasted through a multi-level feature space and adjusted via multi-path backpropagation.

where $j \in \{1, m\}$. These 2D image features correspond to the feature $\mathbf{Z}_i^P$ of the i-th 3D object. Building on the 2D-3D cross-modal contrastive scheme, we further extend it and design a cumulative loss $Loss_{inter\text{-}plus}$:

$$\begin{aligned} J_{inter}^{i}(P, I) &= \sum_{k=1}^{n} \exp\left(\text{sim}\left(\mathbf{Z}_i^P, \mathbf{Z}_k^P\right)/\tau\right) + \exp\left(\text{sim}\left(\mathbf{Z}_i^P, \mathbf{H}_k^I\right)/\tau\right), \\ \mathcal{L}_{inter}^{i}(P, I) &= -\log \frac{\exp\left(\text{sim}\left(\mathbf{Z}_i^P, \mathbf{H}_i^I\right)/\tau\right)}{J_{inter}^{i}(P, I) - \exp\left(\text{sim}\left(\mathbf{Z}_i^P, \mathbf{Z}_i^P\right)/\tau\right)}, \\ Loss_{inter\text{-}plus} &= \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\mathcal{L}_{inter}^{i}(P, I_j) + \mathcal{L}_{inter}^{i}(I_j, P)\right) \end{aligned} \tag{3}$$

Multi-modal Contrastive based on Multi-MLP To alleviate the consistency conflicts between different 2D views and 3D object, we construct a multi-level feature learning strategy based on Multi-MLP. The multi-modal objective is then achieved by contrasting 2D multi-view and 3D objects in different dimensional feature spaces, as depicted in Fig.4.

Building upon the modality-specific and cross-modal training paths designed, for multiple 2D views, we further stack different MLPs on the encoder $\left\{\boldsymbol{f}^P, \boldsymbol{f}^I\right\}$, mapping them to distinct feature spaces. The mappings for the 3D encoder and 2D encoder are as follows:

$$\begin{aligned} &F\left(\{\mathbf{G}_H\}_{H=1}^{m}; \boldsymbol{f}^P\right); \\ &F\left(\{\mathbf{G}_H\}_{H=1}^{m}; \boldsymbol{f}^I\right) \end{aligned} \tag{4}$$

Here, $\{\mathbf{G}_H\}_{H=1}^{m}$ denotes the additional projection heads of the MLP layers of the m 2D multi-views, each with different output dimensions. The output feature dimension of *cross-modal* projection heads exceeds that of *intra-modal* outputs. The feature extraction process for the j-th 2D view corresponding to the i-th 3D object is as follows:

$$\left\{\mathbf{F}^H\right\}_{H=j} = \{\mathbf{G}_H\}_{H=j}\left(\mathbf{H}_{ij}^{I}\right) = \{\mathbf{G}_H\}_{H=j}\left(f_I\left(I_{ij}\right)\right) \tag{5}$$

Multi-level Augmentation Invariance

An overview of our method is shown in Fig.5. We propose an improvement to the contrastive framework for 3D point clouds and 2D multi-views through a multi-level augmentation module. Crucially, we employ an incremental strategy to generate multi-level augmentation, which in turn applies distinct transformations to the 2D views.

In 2D multi-view data, there are two types of information contained: the shared semantic information across all multi-views and the private information specific to each individual 2D view. If each view is treated equally with the same type or intensity of enhancement during training, the model will learn a non-optimal representation.

Meanwhile, in alignment with the *InfoMax* (Bell 1995) principle, the goal of contrastive is to capture as much information as possible about stimuli. The *InfoMin* (Noroozi and Favaro 2016) principle suggests that further reduction of mutual information at the *intermediate optimal* can be achieved by utilizing more robust augmentation.

Furthermore, we explore enhancing representational performance by mining hard samples through *RandomCrop*. Given a 2D view image I, we first determine the cropping ratio s and aspect ratio r from a predefined range. This can be described as follows:

$$(x, y, h, w) = \mathbb{R}_{crop}(s, r, I), \tag{6}$$

where $\mathbb{R}_{\text{crop}}(\cdot,\cdot,\cdot)$ is a random sampling function that returns a quaternion (x, y, h, w). where (x, y) represents the coordinates of the cropping center, and (h, w) represents the height and width of the cropping.

Assuming the number of 2D views is m, the complete augmentation pipeline is specified as $T = \mathit{Combine}\{t_0, t_1, t_2, t_3, \cdots, t_m\}$, where t_0 contains the basic augmentation and $t_1 \sim t_m$ represents a specific type of augmentation. Then, the incremental strategy can be represented as:

$$\begin{aligned} T_1 &= \mathit{Combine}\{t_0, t_1\} \\ T_2 &= \mathit{Combine}\{t_0, t_1, t_2\} \\ T_3 &= \mathit{Combine}\{t_0, t_1, t_2, t_3\} \\ T_m &= \mathit{Combine}\{t_0, t_1, t_2, t_3, \cdots, t_m\} \end{aligned} \tag{7}$$

Using these modules, we transform the 2D multi-view samples $\{x_i\}_{1\le i\le m}$ of the same 3D object into m images with different augmentation intensities:

$$v_i = T_i(x_i), \quad i = 1, 2, 3, \cdots, m \tag{8}$$

The augmentation transformations from T_1 to T_m gradually increase in intensity and number, reducing the shared mutual information between the transformed 2D view and the 3D object. Further, we use a projection head g_i based on Multi-MLP to map the features into the loss space:

$$z_i = g_i(f_i(T_i(v_i))), i = 1, 2, 3, \cdots, m \tag{9}$$

Here, f_i represents the encoder and z_i represents the features in the latent space. Note that the number of projection heads, rendered multi-view 2D images and augmentation modules are consistent. Also, the strength of the augmentation module and the variation trend of feature projection dimensions in Multi-MLP are consistent.

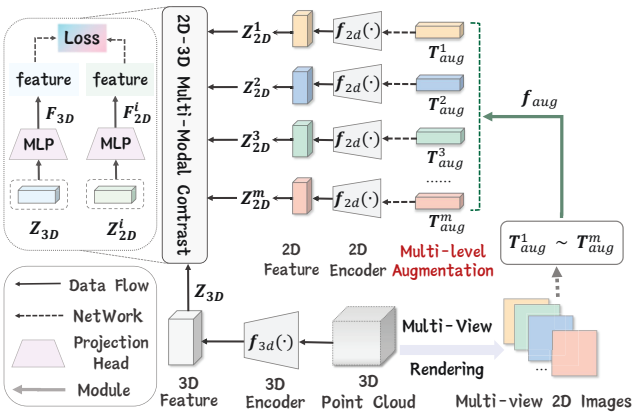

Figure 5: Schematic of Multi-level Augmentation. Multi-level augmentation are generated in an incremental manner. 2D multi-views correspond to a certain level of augmentation indicated by different colors.

Type	Method	Accuracy (%)	
		ModelNet40	ModelNet10
Sup.	PointNet (Qi et al. 2017a)	89.2	-
	GIFT (Dovrat et al. 2019)	89.5	91.5
	MVCNN (Su et al. 2015)	89.7	-
Self.	Point-BERT (Yu et al. 2022)	87.4	-
	Point-MAE (2022)	91.0	-
	Jigsaw3D (Sauder 2019)	90.6	94.5
	Vconv-DAE (Dovrat 2021)	75.5	80.5
	SwAV (Caron et al. 2020b)	90.3	93.5
	OcCo (Wang 2020)	89.2	92.7
	STRL (Liu et al. 2021)	90.9	-
	CrossPoint (Afham 2022)	91.2	-
	MM-Point (ours)	**92.4**	**95.4**
	Improvement	**+1.2**	**+0.9**

Table 1: Linear SVM classification results on ModelNet40 and ModelNet10. *Self.* and *Sup.* represent pre-training with self-supervised and supervised methods.

Therefore, the overall loss function formula is updated as follows, where $L_{contrast}$ refers to the cross-modal loss between the 3D feature z_{j} and a specific 2D view feature z_{i}:

$$L_i = L_{contrast}(z_i, z_j), \quad L_{overall} = \sum_{i=1}^{m} L_i \tag{10}$$

This way, the distribution of the augmentation invariance for t_0 and t_1 is the broadest, and the invariance of t_m is limited to the corresponding features of the largest dimension.

Experiments

In this section, we first introduce the pre-training details of MM-Point. As our focus is on 3D representation learning, we only evaluate the pre-trained 3D point cloud encoder backbones. We sample different downstream tasks and assess the 3D feature representations learned by MM-Point.

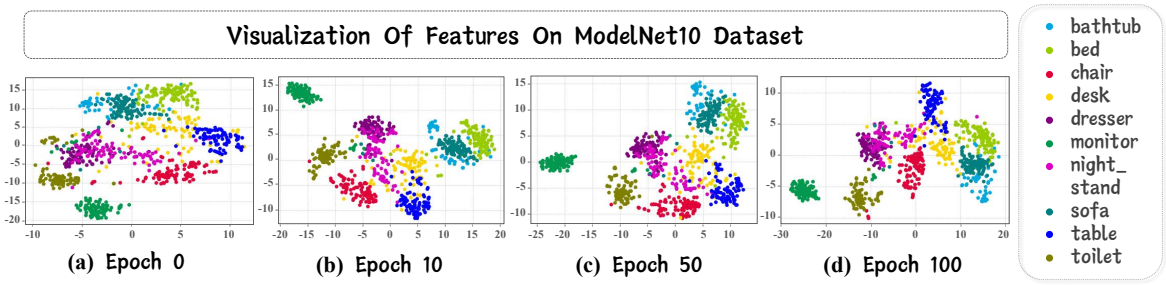

Figure 6: t-SNE visualization of point-level features extracted by MM-Point on ModelNet10 (Qi et al. 2016), with each feature point colored according to its class label. Points with the same color represent semantic similarity.

Method	Accuracy (%)
GBNe (Touvron et al. 2020)	80.5
PRANet (Zhang et al. 2020)	81.0
PointMLP (Ma et al. 2022)	85.2
Point-BERT (Yu et al. 2022)	83.1
MaskPoint (Liu, Cai, and Lee 2022)	84.3
Point-MAE (Pang et al. 2022)	85.2
OcCo (Wang 2020)	78.3
STRL (Liu et al. 2021)	77.9
CrossPoint (Afham 2022)	81.7
MM-Point (ours)	**87.8**
Improvement	**+6.1**

Table 2: Evaluation of 3D point cloud linear SVM classification on ScanObjectNN.

Method	5-way		10-way	
	10-shot	20-shot	10-shot	20-shot
	Results on ModelNet40			
3D-GAN (Wu 2016)	55.8±3.4	65.8±3.1	40.3±2.1	48.4±1.8
PointCNN (Li 2018)	65.4±2.8	68.6±2.2	46.6±1.5	50.0±2.3
RSCNN (Li et al. 2018)	65.4±8.9	68.6±7.0	46.6±4.8	50.0±7.2
Jigsaw (Sauder 2019)	34.3±1.3	42.2±3.5	26.0±2.4	29.9±2.6
OcCo (Wang 2020)	90.6±2.8	92.5±1.9	82.9±1.3	86.5±2.2
CrossPoint (2022)	92.5±3.0	94.9±2.1	83.6±5.3	87.9±4.2
MM-Point (ours)	**96.5±2.8**	**97.2±1.4**	**90.3±2.1**	**94.1±1.9**
	Results on ScanObjectNN			
Jigsaw (Sauder 2019)	65.2±3.8	72.2±2.7	45.6±3.1	48.2±2.8
OcCo (Wang 2020)	72.4±1.4	77.2±1.4	57.0±1.3	61.6±1.2
CrossPoint (2022)	74.8±1.5	79.0±1.2	62.9±1.7	73.9±2.2
MM-Point (ours)	**88.0±6.5**	**90.7±5.1**	**76.7±1.9**	**83.9±4.2**

Table 3: Few-shot classification: SVM classification accuracy comparison on ModelNet40 and ScanObjectNN. We report the average accuracy (%) and standard deviation (%).

Pre-training Setup

Datasets **ShapeNet** (Chang et al. 2015) is a large-scale 3D shape dataset containing 51162 synthetic 3D point cloud objects. For each object in the dataset, we render the 3D object into 2D multi-views, obtaining 24 images per object.

Implementation Details For the point cloud modality, we employ DGCNN (Wang et al. 2019) as the 3D backbone. For the image modality, we use ResNet-50 as the 2D backbone. For all encoders, we append a 2-layer non-linear MLP projection head to generate the final representation. Note that we add different projection heads to obtain features. Pre-training employs AdamW as the optimizer.

3D Object Classification

The point cloud classification experiments were conducted on three datasets. Notably, we utilized a pre-trained encoder with frozen weights for evaluation using a linear SVM.

ModelNet40 (Wu et al. 2015) is a synthetic point cloud dataset obtained by sampling 3D CAD models, consisting of 12311 3D objects. **ModelNet10** (Qi et al. 2016) includes 4899 CAD models with orientations from 10 categories. **ScanObjectNN** (Uy et al. 2019) is a real-world 3D object dataset comprising 2880 unique point cloud objects. This dataset offers a more realistic and challenging setting.

To evaluate the effectiveness of the point cloud representation learned by MM-Point, we first performed random sampling of 1024 points for each object.

The classification accuracy results for ModelNet40 and ModelNet10 are shown in Tab.1. As illustrated, MM-Point outperforms all existing self-supervised methods, achieving classification accuracies of 92.4% and 95.4%, respectively.

To validate the effectiveness in the real world, we conducted experiments on ScanObjectNN, and the evaluation results are presented in Tab.2. In comparison with the state-of-the-art methods, the accuracy has been significantly improved by 6.1%, highlighting the advantages of MM-Point in challenging scenarios in real-world environments.

To visualize the learned 3D representations, we employ t-SNE to reduce the dimensionality of the latent representations and map them onto a 2D plane. Fig.6 presents the visualization of 3D point cloud features learned by MM-Point.

3D Object Few-shot Classification

The conventional setting for FSL is *N-way K-shot*. We conduct experiments using the ModelNet40 and ScanObjectNN datasets. Specifically, we report the results of 10 runs and calculate their mean and standard deviation.

We report the results in Tab.3. On both ModelNet40 and ScanObjectNN, MM-Point achieves ***SOTA*** accuracy, outperforming other few-shot classification methods, and demonstrating substantial improvements across all settings.

Category	Method	Metrics	
		OA (%)	mIoU (%)
Sup.	PointNet (Qi et al. 2017a)	-	83.7
	PointNet++ (Qi et al. 2017b)	-	85.1
	DGCNN (Wang et al. 2019)	-	85.1
Self-sup.	OcCo (Wang 2020)	94.4	85.0
	CrossPoint (Afham 2022)	94.4	85.3
	MM-Point (ours)	**94.5**	**85.7**

Table 4: Overall accuracy and mean IoU results for 3D part segmentation. The metrics are OA(%) and mIoU(%).

Method	Metrics	
	OA(%)	mIoU(%)
Jigsaw3D (Sauder 2019)	84.1	55.6
OcCo (Wang 2020)	84.6	58
CrossPoint (Afham 2022)	86.7	58.4
MM-Point (ours)	**88.7**	**59.1**

Table 5: Semantic segmentation results on S3DIS (Armeni et al. 2016). We report the *mIoU* and *OA* for all 13 classes.

3D Object Part Segmentation

We also extend MM-Point to the task of 3D object part segmentation, a challenging fine-grained 3D recognition task.

The **ShapeNetPart** (Yi et al. 2016) dataset contains 16881 point clouds, with 3D objects divided into 16 categories and 50 annotated parts. We sample 2048 points from each input instance.

For a fair comparison, we follow previous works and add a simple part segmentation head on top of the DGCNN encoder. We evaluate the performance using Overall Accuracy (OA) and mean Intersection over Union (mIoU) metrics. Tab.4 summarizes the evaluation results.

3D Object Semantic Segmentation

The Stanford Large-Scale 3D Indoor Spaces Dataset (**S3DIS**) (Armeni et al. 2016) consists of 3D scan data from 271 rooms across six different indoor spaces. For evaluation, we train our model from scratch on Areas $1 \sim 4$ and Area 6, using Area 5 for validation.

Tab.5 demonstrates the performance of MM-Point. Compared to the randomly-initialized baseline without pre-training, our proposed MM-Point pre-training method yields a significant improvement ($+4.2\%$ mIoU).

Ablation Study

In order to investigate the contributions of each main component in MM-Point, we conduct an extensive ablation study.

Multi-view Contrastive: Number of Views We evaluate the performance of MM-Point by performing multi-modal contrastive with different numbers of 2D views. As shown in Tab.6, we observe that the performance of MM-Point is the lowest when using only one view image. As more 2D image views are added, the classification performance steadily improves. The performance is best when the number of multi-view 2D images M is 4.

Number of 2D images		1	3	**4**	5	6
Accuracy (%)	ModelNet40	91.3	92.2	**92.4**	92.4	92.1
	ScanObjectNN	83.3	86.7	**87.8**	87.6	86.9

Table 6: Ablation test using different numbers of 2D views.

Multi-MLP		Accuracy (%)	
Intra-modal	Inter-modal	ModelNet40	ScanObjectNN
✗	✗	91.4	83.4
✗	✔	91.9	86.7
✔	✗	91.6	84.9
✔	✔	**92.4**	**87.8**

Table 7: A comparison using different Multi-MLP strategies.

Multi-MLP Strategy Employing different MLPs for *intra-modal* and *inter-modal* feature embedding, as well as for different dimensional embedding of multi-views, is a distinctive design in MM-Point. We evaluate our method by: (1) using a unified MLP, (2) employing different dimensions for *intra-modal* and *inter-modal* features, ensuring that *inter-modal* are larger than *intra-modal* dimensions, and (3) using unified output dimensions or multiple different dimensions for 2D multi-views. The results are reported in Tab.7 . These results indicate that our unified framework benefits from using multiple different spaces for multi-modal modeling.

Multi-level Augmentation Strategy To validate the effectiveness of the multi-level augmentation strategy, we experiment with: (1) all 2D views based solely on unified augmentations, (2) all 2D views based on different augmentations, but without increasing the difficulty in a hierarchical manner, and (3) all 2D views based on different multi-level augmentations. The results are reported in Tab.8. The absence of any specific augmentation (indicated by ✗) negatively impacts the performance. This suggests that applying multi-level augmentations to 2D multi-views allows the model to benefit from contrast in the augmentation space.

Multi-level Augmentation		Acc. (%)	
Multi Aug.	Multi-level	ModelNet40	ScanObjectNN
✗	✗	91.7	86.1
✔	✗	92.1	86.9
✔	✔	**92.4**	**87.8**

Table 8: The impact of Multi-level augmentation strategy.

Conclusion

In this paper, we explore a novel pre-training method for 3D representation learning. We introduce MM-Point, a framework that encourages 3D point clouds to learn excellent features from 2D multi-views. Concurrently, MM-Point employs Multi-MLP and Multi-level augmentation strategies to effectively learn more robust 3D modality knowledge from 2D multi-views. MM-Point consistently exhibits state-of-the-art performance on various downstream tasks. Codes are available at https://github.com/HaydenYu/MM-Point.

Acknowledgments

This work was supported by National Natural Science Foundation of China 61906036 and the Fundamental Research Funds for the Central Universities $(2242023k30051)$. This research work was also supported by the Big Data Computing Center of Southeast University

References

Afham, M. 2022. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9902–9912.

Armeni, I.; Sax, A.; Zamir, A.; and Savarese, S. 2016. 3D semantic parsing of large-scale indoor spaces. *arXiv preprint arXiv:1604.04545*.

Bell, T. 1995. Information theory and the central limit theorem. *IEEE Transactions on Information Theory*, 41(3): 726–735.

Caron, M.; Bojanowski, P.; Joulin, A.; and Douze, M. 2020a. SwAV: Unsupervised Learning of Features by Swapping Across Views. In *European Conference on Computer Vision*, 100–116. Springer.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020b. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*.

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

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2021. Normalized Information Distance for Contrastive Representation Learning. In *ICML*.

Dovrat, K. 2021. Vconv-DAE: 3D shape segmentation via volumetric convolutional autoencoder. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dovrat, K.; Litany, O.; Bronstein, M.; and Averbuch-Elor, H. 2019. GIFT: A Real-Time and Scalable 3D Shape Search Engine. In *ICCV*.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Mutual information neural estimation. *NeurIPS*.

Huang, J.; Hao, Y.; Zhang, W.; and Liu, Y. 2020. S3DIS: Self-Supervised 3D Scene Representation via Inverse Dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2020. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1087–1101.

Li, Y. 2018. PointCNN. In *CVPR*.

Li, Y.; Bu, R.; Sun, M.; Wu, W.; and Di, X. 2018. PointCNN: Convolution On X-Transformed Points. In *Advances in Neural Information Processing Systems*, 820–830.

Liu, H.; Cai, M.; and Lee, Y. J. 2022. Masked discrimination for self-supervised learning on point clouds. In *European Conference on Computer Vision*, 657–675. Springer.

Liu, S.; Wang, Z.; Qi, X.; and so on. 2021. Stochastic Temporal Repulsion Learning for Self-supervised Representation of 3D Point Clouds. In *CVPR*.

Liu, X.; and Li, B. 2020. Prior3D: Point Cloud Prior for 3D Object Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3615–3624.

Ma, X.; Qin, C.; You, H.; Ran, H.; and Fu, Y. 2022. Rethinking network design and local geometry in point cloud: A simple residual MLP framework. *arXiv preprint arXiv:2202.07123*.

Noroozi, M.; and Favaro, P. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In *European Conference on Computer Vision*, 69–84. Springer.

Pang, Y.; Wang, W.; Tay, F. E.; Liu, W.; Tian, Y.; and Yuan, L. 2022. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, 604–621. Springer.

Qi, C. R.; Liu, W.; Wu, C.; Su, H.; and Guibas, L. J. 2018. JigSaw: A large-scale, low-cost, and high-accuracy annotation framework for 3D object instance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *CVPR*.

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

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017b. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *NeurIPS*.

Qian, R.; Fracastoro, G.; Paudel, D. P.; Pinhanez, C. S.; and Favaro, P. 2020. PointGMM: A Neural GMM Network for Point Clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7178–7187.

Sauder, J. 2019. Self-supervised deep learning on point clouds by reconstructing space. *Advances in Neural Information Processing Systems*, 32.

Shi, S.; Wang, X.; and Li, H. 2019. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–779.

Shi, S.; Wang, X.; and Li, H. 2020. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10529–10538.

Shi, S.; Zhou, X.; and Li, H. 2020. PV-RCNN++: Point-Voxel Feature Set Abstraction for 3D Object Detection. In *Advances in Neural Information Processing Systems*, 3505–3516.

Su, H.; Maji, S.; Kalogerakis, E.; and Learned-Miller, E. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *ICCV*.

Sun, J.; Wang, Y.; Zhang, W.; Zhou, Z.; Kong, T.; and Li, L. 2021. SeedCon: Unsupervised Point Cloud Object Segmentation with Data-Efficient Seed Consensus. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tian, Y.; Krishnan, D.; and Isola, P. 2019. Contrastive multiview coding. In *Advances in Neural Information Processing Systems*, 159–169.

Touvron, H.; Caron, M.; Joulin, A.; Alayrac, J.-B.; Bojanowski, P.; Laptev, I.; and Neverova, N. 2020. Training data-efficient image transformers & distillation through attention. In *ICML*.

Tran, B.; Hua, B.-S.; Tran, A. T.; and Hoai, M. 2022. Self-supervised learning with multi-view rendering for 3d point cloud analysis. In *Proceedings of the Asian Conference on Computer Vision*, 3086–3103.

Uy, M. A.; Pham, Q.-H.; Hua, B.-S.; Nguyen, T.; and Yeung, S.-K. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1588–1597.

Wang, X. E.; Wu, Q.; Wang, X.; and Xiao, J. 2020. Vision-Language Navigation with Self-Supervised Auxiliary Reasoning Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10811–10820.

Wang, Y. 2020. OcCo: Occupancy-aware 3D convolutional networks for indoor/outdoor semantic segmentation. In *CVPR*.

Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S. E.; Bronstein, M. M.; and Solomon, J. M. 2019. Dynamic graph cnn for learning on point clouds. In *Proceedings of the ACM SIGGRAPH Asia 2019 Technical Papers*, 9.

Wu, J. 2016. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*, 82–90.

Wu, Z.; Song, S.; Khosla, A.; Yu, F.; Zhang, L.; Tang, X.; and Xiao, J. 2015. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*.

Xie, S.; Gu, J.; Guo, D.; Qi, C. R.; Guibas, L.; and Litany, O. 2020. 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*, 574–591. Springer.

Yang, Y.; Feng, C.; Shen, Y.; and Tian, D. 2020. PointFlowNet: Learning Representations for Rigid Motion Estimation from Point Clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yi, L.; Su, H.; Guo, X.; and Guibas, L. J. 2016. Scalable shape retrieval with a family of path-based convolutional neural networks. In *CVPR*.

You, A.; Chen, X.; Li, S.; Yan, Q.; and Yang, X. 2021. H3DNet: 3D Object Detection Using Hybrid Geometric Primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu, X.; Tang, L.; Rao, Y.; Huang, T.; Zhou, J.; and Lu, J. 2022. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19313–19322.

Zhang, Y.; Liu, Z.; Zhou, Y.; and Qi, H. 2020. Simple view for point cloud recognition. In *NeurIPS*.