Self-Supervised Modality-Invariant and Modality-Specific Feature Learning for 3D Objects

Anonymous authors
Paper under double-blind review

Abstract

While most existing self-supervised 3D feature learning methods mainly focus on point cloud data, this paper explores the inherent multimodal attributes of 3D objects. We propose to jointly learn effective features from different modalities including image, point cloud, and mesh with heterogeneous networks from unlabeled 3D data. Our proposed novel self-supervised model learns two types of distinct features: modality-invariant features and modality-specific features. The modality-invariant features capture high-level semantic information across different modalities with minimum modality discrepancy, while the modality-specific features capture specific characteristics preserved in each modality. These two types of features provide a more comprehensive representation for 3D data. The quality of the learned features are evaluated on different downstream tasks including 3D object recognition, 3D within-modal retrieval, and 3D cross-modal retrieval tasks with three data modalities including image, point cloud, and mesh. Our proposed method significantly outperforms the state-of-the-art self-supervised methods for all the three tasks and even achieves comparable performance with the state-of-the-art supervised methods on the ModelNet10 and ModelNet40 datasets.

1 Introduction

Self-supervised learning methods learn visual features from large-scale datasets without requiring any manual annotations. The core of self-supervised learning is to define a pretext task and learn visual features through the processing of accomplishing the pretext task. Since it can be easily scaled up to large-scale datasets, recently some self-supervised methods achieved comparable or even better performance on some downstream tasks than supervised methods (Asano et al., 2019; Chen et al., 2020; He et al., 2019; Jing & Tian, 2019; Misra & van der Maaten, 2019).

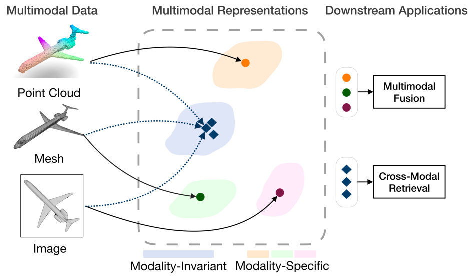

Figure 1: The proposed self-supervised model jointly learns two types of distinct features: modality-invariant features and modality-specific features. The modality-invariant features capture features for multiple modalities in the same metric space making the cross-modal retrieval task possible, while the modality-specific features encode complementary information among different modalities and the fusion of these features can be used for downstream tasks such as recognition.

Most of the existing self-supervised learning methods focus on learning features for only one modality. As a rising trend to model 3D visual features, various methods were proposed to learn point cloud features from point cloud either by reconstructing point cloud (Achlioptas et al., 2017; Gadelha et al., 2018; Yang et al., 2018; Zhao et al., 2019b), by generating point cloud with Generative Adversarial Networks (Li et al., 2018a; Sun et al., 2018; Wu et al., 2016), or by accomplishing pre-defined pretext tasks (Hassani & Haley, 2019; Zhang & Zhu, 2019). Recently Jing *et al.* started to explore multimodal correspondence of 3D data as a supervision signal for 3D self-supervised feature learning (Jing et al., 2020).

Generally, 3D data are inherently multimodal such as mesh, point cloud, multi-view images, etc. The correspondence among multiple modalities is a rich source of supervision signals for self-supervised learning. However, only a few work (Jing et al., 2020) attempted to utilize the multimodal correspondence for self-supervised learning. To fully explore the potential of using it for self-supervised learning, as shown in Fig. 1, we propose a novel framework to jointly learn modality-invariant and modality-specific features for 3D objects.

The modality-invariant features aim to reduce modality gaps. For each object, no matter its modality, the features from different modalities are firstly extracted by different encoders and then mapped into the same universal space to reduce the modality discrepancy. Although these features are extracted from different modalities, the features for each object share the same underline high-level semantic information such as the context and structure of the objects. Mapping these features to the same space helps the network to capture the shared correlated features that are invariant to different modalities. The modality-invariant features can be directly compared making the 3D cross-modal retrieval task feasible.

Different from the modality-invariant features, our proposed model also learns modality-specific features that preserve specific characteristics of each modality. For each object, each modality has distinctive characteristics such as images explicitly encode texture information while point clouds explicitly encode the 3D local structure of the objects. The representations from different modalities encode features from different perspectives and might be complementary to each other. Therefore, the features from different modalities can be fused to form a more robust and comprehensive representation for the data samples which can potentially benefit downstream tasks such as 3D object recognition.

These modality-specific features along with the modality-invariant features in a common space jointly provide a comprehensive multimodal representation of 3D objects. To learn both modality-invariant and modality-specific features, we propose three different constraints: cross-modal invariant constraint enforces the network to maximize the similarity of features from different modalities for the same object, cross-view invariant constraint enforces the network to maximum similarity of features from different views of data for objects in the same modality, and soft orthogonal constraint avoids the redundancy between the modality-invariant and modality-specific features. Our proposed framework is evaluated on different downstream tasks including 3D object recognition, 3D within-domain retrieval, and 3D cross-modal retrieval tasks using two popular 3D object benchmark datasets (i.e. ModelNet40 and ModelNet10) with three different modalities (i.e.image, point cloud, and mesh). In both datasets with all the downstream tasks, our proposed framework significantly outperforms the state-of-the-art self-supervised models. The main contributions of this paper are summarized as follows:

- We propose a novel self-supervised learning framework to jointly learn modality-invariant and modality-specific features for 3D objects without using any manual labels.
- To the best of our knowledge, we are the first to extensively explore the self-supervised 3D cross-modal retrieval for 3D objects with three modalities including image, point cloud, and mesh.
- Our proposed method significantly outperforms the state-of-the-art self-supervised methods on multiple downstream tasks and even achieves comparable performance with the state-of-the-art supervised methods on the ModelNet10 and ModelNet40 datasets.

2 Related Work

Self-supervised 2D Feature Learning: Many methods have been proposed to learn visual features from unlabeled 2D data including videos and images. Based on the source of supervision signal, there are four types of self-supervised learning methods: generation-based method, context-based method, free semantic label-based method, and cross-modal-based method. The generation-based methods learn features by reconstructing the data such as generating images or videos with GAN

(Goodfellow et al., 2014; Ledig et al., 2017; Zhang et al., 2016; Srivastava et al., 2015). The context-based methods learn features by using spatial context or temporal context including Jigsaw puzzle (Noroozi & Favaro, 2016), geometric transformation (Gidaris et al., 2018; Jing & Tian, 2018), clustering (Caron et al., 2018), frame order reasoning (Misra et al., 2016). The free semantic label-based methods learn features either by data generated by game engines or to distil features from other unsupervised learning features (Pathak et al., 2017). The cross-modal-based methods learn features by the correspondence between a pair of channels of data including video-audio (Korbar et al., 2018) or video-text. Recently, more researchers explore to apply these self-supervised learning methods to 3D point cloud data (Hassani & Haley, 2019; Jing et al., 2020; Zhang & Zhu, 2019; Sauder & Sievers, 2019).

Self-supervised 3D Feature Learning: Several self-supervised learning methods have been proposed to learn features for 3D point cloud objects by reconstructing point cloud data (Achlioptas et al., 2017; Gadelha et al., 2018; Yang et al., 2018; Zhao et al., 2019b), by generating point cloud with GANs (Li et al., 2018a; Sun et al., 2018; Thabet et al., 2019; Wu et al., 2016), or by training networks to solve pre-defined pretext tasks (Hassani & Haley, 2019; Jing et al., 2020; Sauder & Sievers, 2019; Zhang & Zhu, 2019). Sauder *et al.* proposed to learn point cloud features by training networks to recognize the relative position of two segments of point cloud (Sauder & Sievers, 2019). Zhang *et al.* designed clustering and contrastive as pretext task to train networks to learn point cloud features (Zhang & Zhu, 2019). Hassani *et al.* proposed to train networks with multiple pre-defined pretext tasks including clustering, prediction, and reconstruction for point cloud data (Hassani & Haley, 2019). Jing *et al.* proposed to utilize cross-modal relations of point clouds and multi-view images as the supervision signal to jointly learn image and point cloud features for 3D objects (Jing et al., 2020). However, the point cloud and image features learned by the network in (Jing et al., 2020) are not modality-invariant. To thoroughly utilize the cross-modal coherent attributes of 3D data, here we propose to learn modality-invariant and modality-specific features for 3D objects with three different modalities including image, point cloud, and mesh.

Multimodal Feature Learning: The multimodal feature learning has been widely studied in other research fields including video action recognition (Feichtenhofer et al., 2016; Simonyan & Zisserman, 2014; Wang et al., 2015), video captioning (Venugopalan et al., 2015), cross-modal retrieval (Ging et al., 2020; Lee et al., 2018; Li et al., 2019b), etc. The features from different modalities usually capture features from different perspectives, therefore, these features might be complementary to each other. However, the multimodal feature learning has not been widely explored in 3D object recognition task which is a fundamental task for 3D applications. Our model can learn modality-specific features and the fusion of the modality-specific features from multiple modalities can provide a more comprehensive representation for 3D objects.

Cross-Modal Retrieval Task: The cross-modal retrieval aims to retrieval data from one modality by using the query from another modality (e.g. retrieval image using text) (Ging et al., 2020; Lee et al., 2018; Li et al., 2019b; Wang et al., 2017; Zhen et al., 2019). The challenge for this task is to learn features with minimum modality discrepancy for data from multiple modalities. Many deep learning methods have been proposed for retrieval task such as adversarial cross-modal retrieval (ACMR) (Wang et al., 2017) and deep supervised cross-modal retrieval (DSCMR) (Zhen et al., 2019). Normally, all these methods require large-scale labelled datasets for training. In this paper, we explore a less studied task, 3D cross-modal retrieval, in a self-supervised learning way. Our model can learn modality-invariant features without using any manual labels while achieves comparable performance to the state-of-the-art supervised methods.

3 Method

An overview of the proposed framework is shown in Fig. 2. The core of our method is to optimize heterogeneous networks to jointly learn both modality-invariant and modality-specific features. The framework contains three heterogeneous feature encoders to extract hidden features for three different data modalities. Then these hidden features for each modality are mapped into two types of spaces: one is the universal feature space for the modality-invariant features and the other is the modality-specific space of each data modality for the modality-specific features. The entire framework is jointly trained end-to-end with a combination of our proposed constraints. The general formulation of our proposed method is described in the following subsections.

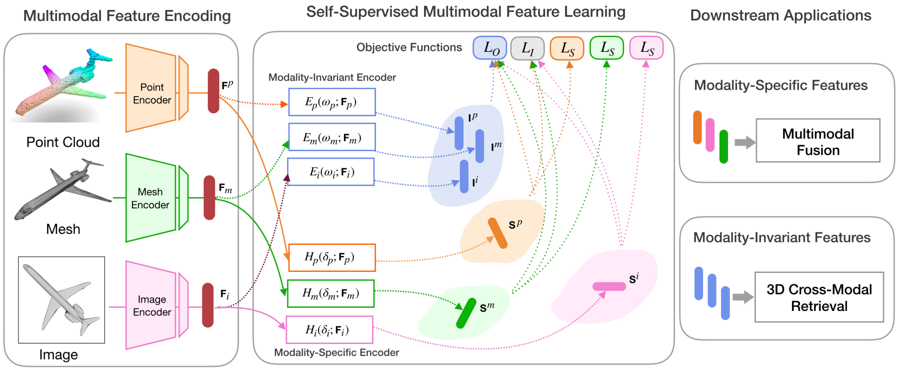
Multimodal Feature Encoding
Self-Supervised Multimodal Feature Learning
Downstream Applications
Point Cloud
Mesh
Image
Point Encoder
Mesh Encoder
Image Encoder
$\mathbf{F}^p$
$\mathbf{F}_m$
$\mathbf{F}_i$
Objective Functions
L_O
L_I
L_S
L_S
L_S
Modality-Invariant Encoder
$E_p(\omega_p; \mathbf{F}_p)$
$E_m(\omega_m; \mathbf{F}_m)$
$E_i(\omega_i; \mathbf{F}_i)$
$\mathbf{I}^p$
$\mathbf{I}^m$
$\mathbf{I}^i$
$\mathbf{S}^p$
$\mathbf{S}^m$
$\mathbf{S}^i$
$H_p(\delta_p; \mathbf{F}_p)$
$H_m(\delta_m; \mathbf{F}_m)$
$H_i(\delta_i; \mathbf{F}_i)$
Modality-Specific Encoder
Modality-Specific Features
Multimodal Fusion
Modality-Invariant Features
3D Cross-Modal Retrieval

Figure 2: An overview of the proposed self-supervised modality-invariant and modality-specific feature learning for 3D objects. The hidden features for mesh, point cloud, and image are extracted by corresponding encoders, then these hidden features for each modality are mapped into two spaces including a universal space for capturing modality-invariant features and a private feature space for each modality for capturing modality-specific features. The self-supervised learned features can be further used for various downstream tasks such as 3D cross-modal retrieval and multimodal fusion for 3D recognition.

3.1 Problem Setup

For a dataset D contains N unlabeled instances where the i-th instance d_i is a set of M modalities, it can be formulated as:

$$D = \{d_i\}_{i=1}^N, \quad d_i = \{x_i^m\}_{m=1}^M. \tag{1}$$

Here, each data instance d_i consists of $\{x_i^1, x_i^2, \cdots, x_i^M\}$ in M different modalities. Normally the learned representations for these M modalities are in different feature spaces and their similarities cannot be directly measured. Our proposed model learns two types of distinct features for each modality x_i^m: modality-invariant features $\mathbf{I}_i^m$ which are invariant to the modality, and modality-specific features $\mathbf{S}_i^m$ which model the specific characteristics preserved in each modality. The modality-invariant features $\mathbf{I}_i^m$ and the modality-specific features $\mathbf{S}_i^1, \mathbf{S}_i^2, ..., \mathbf{S}_i^m$ provide a more comprehensive representations for the object d_i. All these features are jointly learned with our proposed framework for each modality from unlabeled data.

3.2 Multimodal Feature Encoding

For each data instance d_i, each modality sample x_i^m is firstly mapped into a hidden vector $\mathbf{F}_i^m$ by a feature encoder G_m specifically designed for the modality m:

$$\mathbf{F}_i^m = G_m(\theta_m, x_i^m), \tag{2}$$

while θ_m is the learnable parameters of G_m. Normally, this hidden representation F_m is in a separate modality specific space. Given a dataset with M different modalities, there are M different feature encoders as $G_1(\theta_1, x_i^1), G_2(\theta_2, x_i^2), ..., G_m(\theta_M, x_i^M)$. Therefore, for each instance d_i, the hidden vectors $\{\mathbf{F}_i^1, \mathbf{F}_i^2, ..., \mathbf{F}_i^m\}$ are obtained by M encoders.

To jointly learn both modality-invariant and modality-specific features, our model maps the hidden representation $\mathbf{F}_i^m$ for each modality x_i^m into two feature spaces (i.e. the universal feature and the modality-specific spaces.) To learn these two distinct feature spaces, two parallel heads with neural networks are added on each feature encoder $G_m(\theta_m, x_i^m)$ to map this hidden vector into both modality-invariant and modality-specific features. The mapping process for hidden vector $\mathbf{F}_i^m$ of data x_i^m can be formulated as:

$$\mathbf{I}_i^m = E_m(\omega_m, \mathbf{F}_i^m), \tag{3}$$

$$\mathbf{S}_i^m = H_m(\delta_m, \mathbf{F}_i^m), \tag{4}$$

while H_m maps the hidden representation $\mathbf{F}_i^m$ to the universal feature space and E_m maps $\mathbf{F}_i^m$ into the modality-specific feature space. During the learning, $\mathbf{I}_i^m$ and $\mathbf{S}_i^m$ can be obtained by training the entire network with a combination of our proposed three different constraints.

3.3 Self-supervised Modality-Specific Feature Learning

The modality-specific features aim to model the specific characteristics preserved in each modality. We propose the cross-view invariant constraint to learn the modality-specific features for each modality. The cross-view constraint maximizes the similarity of features from different views of the same object while minimizes the similarity of features from the data of different objects. Inspired by the recent remarkable progress achieved by contrastive learning (Chen et al., 2020), we employ contrastive loss to capture the modality-specific features.

Given each data sample x_i^m from modality m, the data of two views $\{x_i^{m1}, x_i^{m2}\}$ can be obtained by performing a set of data augmentation techniques over the data x_i^m. The general hidden vectors $\{\mathbf{F}_i^{m1}, \mathbf{F}_i^{m2}\}$ are firstly extracted by feature encoder $G_m(\theta_m)$, then the modality-specific features are extracted by the network $H_m(\delta_m)$. Suppose the extracted modality-specific features for data $\{x_i^{m1}, x_i^{m2}\}$ are $\mathbf{S}_i^{m1}$ and $\mathbf{S}_i^{m2}$, the cross-view invariant constraint is optimized by the contrastive loss over the extracted modality-specific features among a batch as:

$$\mathcal{L}_S = \sum_{1 \leq m \leq M} \mathcal{L}_s(\mathbf{S}_i^{m1}, \mathbf{S}_i^{m2}), \tag{5}$$

$$\mathcal{L}_s(\mathbf{S}_i^{m1}, \mathbf{S}_i^{m2}) = -\log \frac{h\big(\mathbf{S}_i^{m1}, \mathbf{S}_i^{m2}\big)}{h\big(\mathbf{S}_i^{m1}, \mathbf{S}_i^{m2}\big) + \sum_{k=1}^{B} \mathbb{1}_{\{k \neq i\}} h\big(\mathbf{S}_i^{m1}, \mathbf{S}_k^{m2}\big)}, \tag{6}$$

where $h(\mathbf{u}, \mathbf{v}) = \exp\big(\frac{\mathbf{u}^\top \mathbf{v}}{||\mathbf{u}||_2 ||\mathbf{v}||_\mathbf{2}} / \tau\big)$ is the exponential of cosine similarity measure, B is the batch size, and τ is the temperature hyper-parameter.

This objective function enforces the networks to capture mutual information across different views of the data from the same modality. After the training finished, the encoder $G_m(\theta_m)$ and the modality-specific heads $H_m(\delta_m)$ can capture modality-specific features for each modality.

3.4 Self-supervised Modality-Invariant Feature Learning

For modality-invariant feature learning from data of multiple modalities, the cross-modal invariant constraint is proposed to enforce the network to capture the high-level semantic information that exists across all the modalities and learns the features that invariant to the modality. For each object, we have a collection of m modalities $\{M_1, \ldots, M_m\}$. To capture the modality-invariant features across different modalities, we train the network with pair-wise multimodal pairs and to maximize the high-level semantic information that co-exists between two different pairs. For any two different modalities of M_a and M_b, the high-level semantic information across the modalities M_a and M_b can be captured by maximizing the feature similarity between the features from M_a and M_b. Given any two modalities of data from the same object, the network is optimized over the features extracted by learning the contrastive loss over the features $\mathbf{I}_i^{M_a}$ and $\mathbf{I}_i^{M_b}$ extracted by the modality invariant feature encoder $\mathbf{E}_m(\omega_m)$.

To fully utilize the multimodal correspondence, we train the network with all the pair combinations (a, b) from M modalities. In this way, the high-level semantic information across all the modalities can be captured by learning the relations among the modality pairs. By considering all the pairs of different modalities, the entire modality-invariant objective function that we optimize is:

$$\mathcal{L}_I = -\sum_{1 \leq a < b \leq M} \log \frac{h\big(\mathbf{I}_i^{M_a}, \mathbf{I}_i^{M_b}\big)}{\sum_{k=1}^{B} h\big(\mathbf{I}_i^{M_a}, \mathbf{I}_k^{M_b}\big)}. \tag{7}$$

By optimizing with all the pairs, the high-level semantic information across all the modalities is maximized through the learning process. Ideally, more modalities of data provide more supervision signal from the correspondences and potentially can lead to better performance. With this objective function, the features from different modalities are directly optimized in the same universal space, therefore, are modality-invariant.

3.5 Soft Orthogonal Feature Learning

Ideally, our model jointly learns the modality-invariant and modality-specific features. However, without other constraints, the model may learn redundant features between the two types of features. To further ensure the model to learn different aspects of the data for each modality, we further constrain the relation between the modality-invariant features $\mathbf{I}_i^m$ and modality-specific features $\mathbf{S}_i^m$ by enforcing a soft orthogonality constraint between a pair of features. For each batch of data, both the modality-invariant and modality-specific features are firstly normalized to zero mean and unit l_2 norm. Let $\mathbf{H}^I_m$ and $\mathbf{H}^S_m$ be the matrices whose rows denote the hidden vectors $\mathbf{I}^m_i$ and $\mathbf{S}^m_i$ for modality m of each object. Then the orthogonality constraint between the invariant and specific feature vectors for modality m is calculated as:

$$\left\|\mathbf{H}^I_m \mathbf{H}^{S^\top}_m\right\|^2_F. \tag{8}$$

Here, $\|\cdot\|^2_F$ is the squared Frobenius norm. In addition to the constraints between the invariant and specific vectors, we also add orthogonality constraints between the modality-specific vectors. The overall difference loss is then computed as:

$$\mathcal{L}_{\mathrm{O}} = \sum_{1 \le i \le M} \left\|\mathbf{H}^I_{M_i} \mathbf{H}^{S^\top}_{M_i}\right\|^2_F + \sum_{1 \le i < j \le M} \left\|\mathbf{H}^S_{M_i} \mathbf{H}^{S^\top}_{M_j}\right\|^2_F. \tag{9}$$

3.6 Jointly Learning

When jointly trained with the above three objective functions, a linear weighted combination of all the loss functions is employed to optimize the entire framework. The final loss to optimize the framework is as:

$$\mathcal{L} = \alpha L_I + \beta L_S + \lambda L_O. \tag{10}$$

After the jointly training finished, the network encoder for each modality is obtained as the pre-trained model and can be used for other downstream tasks. The joint training enables the feature encoders of different modalities to learn comprehensive and robust features.

3.7 Framework Architecture

The effectiveness and generalizability of our proposed model are evaluated on 3D datasets with three different modalities including image, point cloud, and mesh. As shown in Fig. 2, our framework consists of three heterogeneous backbone networks including an image feature encoder, a point cloud feature encoder, and a mesh feature encoder. Two distinct parallel MLP layers are employed over the output of each feature encoder to produce the modality-invariant and modality-specific features for each modality. The details of the framework architecture and implementation details can be found in the appendix.

4 Experimental Results

4.1 Experimental Setup

Self-supervised learning: The proposed framework is jointly trained using SGD optimizer with an initial learning rate of 0.001, the moment of 0.9, and weight decay of 0.0005. The network is optimized with a mini-batch size of 96 for $90,000$ iterations and the learning rate decreases by 90% every $30,000$ iterations. Data augmentation used for point cloud network includes randomly rotated between [0, 2π] degrees along the up-axis, randomly jittered the position of each point by Gaussian noise with zero mean and 0.02 standard deviation. Data augmentation for images includes randomly cropped and randomly flipped with 50% probability. Data augmentation for mesh includes random rotation with a degree between [0, 2π].

Datasets: Two 3D object benchmarks including ModelNet40 (Wu et al., 2015) and ModelNet10 (Wu et al., 2015) are used to evaluate the proposed method. The ModelNet40 contains about $12.3k$ objects covering 40 object classes, while about $9.8k$ are used for training and about $2.5k$ for testing. The ModelNet10 consists of $4,900$ objects belong to 10 categories with $3,991$ for training and 909 for testing.

4.2 Benchmarking Self-Supervised 3D Object Recognition

Following the prior state-of-the-art self-supervised learning methods (Achlioptas et al., 2017; Chen et al., 2003; Gadelha et al., 2018; Girdhar et al., 2016; Hassani & Haley, 2019; Jing et al., 2020; Kazhdan et al., 2003; Sharma et al., 2016; Wu et al., 2016; Yang et al., 2018; Zhao et al., 2019b), we compare the performance with them on 3D object recognition task reporting the TOP-1 classification accuracy of a Support Vector Machine (SVM) over the self-supervised learned features. Compared to the existing methods which mainly learn features for one modality, our method jointly learns features from multiple modalities which makes it possible to fuse the multimodal features for more robust representations. To thoroughly evaluate the performance, we compare the performance by using the single modality and by using the features fused from multiple modalities on the ModelNet40 dataset. For the results of using multiple modalities, the features from these modalities are

Method	Modality	Acc (%)
SPH (Kazhdan et al., 2003)	Mesh	68.2
T-L Network (Girdhar et al., 2016)	Point	74.4
LFD (Chen et al., 2003)	Image	75.5
VConv-DAE (Sharma et al., 2016)	Point	75.5
3D-GAN (Wu et al., 2016)	Point	83.3
FV (Sánchez et al., 2013)	Image	84.8
Latent-GAN (Achlioptas et al., 2017)	Point	85.7
MRTNet-VAE (Gadelha et al., 2018)	Point	86.4
Contrast (Zhang & Zhu, 2019)	Point	86.8
FoldingNet (Yang et al., 2018)	Point	88.4
PointCapsNet (Zhao et al., 2019b)	Point	88.9
MultiTask (Hassani & Haley, 2019)	Point	89.1
XMV (Jing et al., 2020)	Point	89.8
ContextPred (Sauder & Sievers, 2019)	Point	90.6
Orientation (Poursaeed et al., 2020)	Point	90.7
Ours	Image	87.0
Ours	Point	89.7
Ours	Mesh	90.4
Ours	Image & Point	90.6
Ours	Image & Mesh	91.5
Ours	Mesh & Point	92.3
Ours	**3 Modalities**	**92.9**

Table 1: The comparison with the state-of-the-art self-supervised methods for 3D object recognition on the ModelNet40 dataset.

Method	Modality	Acc (%)
VoxNet (Maturana & Scherer, 2015)	Voxel	85.9
Subvolume (Qi et al., 2016)	Voxel	89.2
PointNet (Qi et al., 2017a)	Point	89.2
MVCNN (Su et al., 2015)	Image	90.1
Pairwise (Johns et al., 2016)	Image	90.7
MeshNet (Feng et al., 2019)	Mesh	91.9
PointNet++ (Qi et al., 2017b)	Point	91.9
SpecGCN (Wang et al., 2018)	Point	92.1
PointCNN (Li et al., 2018b)	Point	92.2
DGCNN (Li et al., 2019a)	Point	92.2
PointWeb (Zhao et al., 2019a)	Point	92.3
SpiderCNN (Xu et al., 2018)	Point	92.4
KPConv (Thomas et al., 2019)	Point	92.9
InterpCNN (Mao et al., 2019)	Point	93.0
PointTransformer (Zhao et al., 2020)	Point	93.7
Ours	Image	87.0
Ours	Point	89.7
Ours	Mesh	90.4
Ours	Image & Point	90.6
Ours	Image & Mesh	91.5
Ours	Mesh & Point	92.3
Ours	**3 Modalities**	**92.9**

Table 2: Comparison of our **self-supervised** method over the state-of-the-art **supervised** methods ModelNet40.

extracted and then concatenated together to represent the object, and the TOP-1 classification accuracy over the concatenated features are reported for comparison.

The performance comparison against other state-of-the-art self-supervised methods is shown in Table 1. The overall performance of our method are much better. When only one modality is used, our performance based on mesh modality or point cloud modality is comparable to the-state-of-the-art methods. Compared to the other two modalities, the performance based on image modality is lower and the performance can be improved if more multi-view images are used to represent each object. When fusing any two modalities of features, the performance is consistently improved, while the highest performance is achieved when all the modalities are used. These results demonstrate that the features from multiple modalities are indeed complementary to each other while validating the hypothesis of our method.

To demonstrate the strength and potential of our method, we further compare the performance of our self-supervised method with the state-of-the-art supervised methods on 3D object recognition on the ModelNet40 dataset in Table 2. With the advantage of utilizing multimodal features, our proposed self-supervised method even outperforms most of the supervised learning methods and the performance is only 0.8% lower than the most recent Transformer-based method PointTransformer (Zhao et al., 2020). This demonstrates the potential of utilizing the multimodal features for the fundamental 3D understanding tasks.

4.3 Benchmarking Self-Supervised 3D Cross-Modal and Within-Modal Retrieval

Another advantage of our proposed method is that the learned modality-invariant features from different modalities can be directly compared in the universal space. To thoroughly evaluate the performance of the learned modality-invariant features, we verify the effectiveness with the self-supervised 3D cross-modal and within-modal retrieval tasks among the modalities including image, point cloud, and mesh. The Euclidean distance over the normalized modality-invariant features is used to measure the similarity of data from different modalities. Following the convention, the Mean Average Precision (mAP) score is used to indicate the performance.

The learned modality-invariant features in the universal feature space for three different data modalities make the cross-modal retrieval for 3D objects possible, which is, as far as we know, not explored by any other self-supervised methods. To demonstrate the ability of our proposed method, we compare with two types of methods: (1) other self-supervised 3D feature learning methods including XMV (Jing et al., 2020) and Contrast (Zhang & Zhu, 2019); (2) the supervised cross-modal retrieval model DSCMR (Zhen et al., 2019) which achieved the state-of-the-art performance on four image-text retrieval benchmarks including Wikipedia (Pereira et al., 2013), Pascal (Rashtchian et al., 2010), NUS-WIDE-10k (Chua et al., 2009), and XMediaNet (Peng et al., 2017; 2018) datasets. We conduct 6 pairs of cross-modal retrieval tasks (Mesh2Point, Mesh2Image, Point2Mesh, Point2Image, Image2Point, and Image2Mesh) and 3 pairs of within-modal retrieval tasks (Mesh2Mesh, Point2Point, and Image2Image) on the ModelNet40 dataset.

The performance comparison is shown in Table 3. The existing self-supervised learning methods normally only learn modality-specific features for one or two modalities like Contrast (Zhang & Zhu, 2019) and XMV (Jing et al., 2020), the learned features of these methods cannot be applied to cross-modal retrieval tasks, and their

Task	Self-Supervised			Supervised
Method	XMV (Jing et al., 2020)	Contrast (Zhang & Zhu, 2019)	**Ours**	DSCMR (Zhen et al., 2019)
		3D Cross-Modal Retrieval		
Image2Mesh	—	—	**69.3**	76.9
Image2Point	—	—	**69.8**	73.8
Point2Image	—	—	**70.5**	72.7
Mesh2Point	—	—	**71.0**	70.2
Point2Mesh	—	—	**71.0**	71.6
Mesh2Image	—	—	**71.2**	75.2
		3D Within-Modal Retrieval		
Image2Image	36.3	—	**71.2**	81.1
Point2Point	48.4	45.3	**71.4**	70.8
Mesh2Mesh	—	—	**71.6**	74.8

Table 3: The comparison with the self-supervised learning methods and a state-of-the-art **supervised** method for the 3D cross-modal and within-modal retrieval tasks on the ModelNet40 dataset.

performance for the within-modal retrieval tasks are much lower due to lacking carefully designed constraints. Our model learns modality-invariant features for multiple modalities and the performance for all the retrieval tasks are much higher than these self-supervised learning methods (Jing et al., 2020; Zhang & Zhu, 2019) and even outperform the supervised method DSCMR (Zhen et al., 2019) on some tasks such as Mesh2Point and Point2Point. The performance comparison with these models demonstrates the effectiveness of the modality invariance ability of the learned features.

4.4 Ablation Study

To thoroughly evaluate the impact of each component of the propose model, we conducted two sets of ablation studies to evaluate the impact of (1) each loss function and (2) different modalities. The results are shown in Table 4 and Table 5.

Task	L_I	L_S	L_I, L_S	L_I, L_S, L_O
		3D Cross-Modal Retrieval		
Image2Mesh	66.1	7.2	65.8	**69.3**
Image2Point	66.7	4.5	66.3	**69.8**
Point2Image	68.3	6.5	67.8	**70.5**
Mesh2Point	69.6	5.8	70.6	**71.0**
Point2Mesh	69.6	5.1	70.6	**71.0**
Mesh2Image	68.9	6.9	68.6	**71.2**
		3D Within-Modal Retrieval		
Image2Image	69.2	60.9	68.4	**71.2**
Point2Point	69.8	60.3	70.7	**71.4**
Mesh2Mesh	70.7	25.1	71.8	**71.6**
		3D Multimodal Recognition		
Multimodal	92.3	90.9	92.8	**92.9**

Table 4: Ablation study for evaluating impact of each loss function to 3D retrieval and 3D multimodal recognition tasks on ModelNet40.

Task	Mesh-Image	Image-Point	Point-Mesh	All
		3D Cross-Modal Retrieval		
Image2Mesh	60.2	—	—	**69.3**
Image2Point	—	62.9	—	**69.8**
Point2Image	—	64.9	—	**70.5**
Point2Mesh	—	—	62.4	**71.0**
Mesh2Point	—	—	62.7	**71.0**
Mesh2Image	61.7	—	—	**71.2**
		Within-Modality 3D Retrieval		
Image2Image	63.2	64.8	—	**71.2**
Point2Point	—	70.3	62.0	**71.4**
Mesh2Mesh	67.3	—	64.1	**71.6**
		3D Multimodal Recognition		
Multimodal	90.9	89.9	91.5	**92.9**

Table 5: Ablation study for the number of modalities for 3D object retrieval and recognition tasks on the ModelNet40 dataset. 'All' indicates all the three modalities are used.

Ablation Study of Losses. The proposed framework is trained with a combination of three objective functions. To thoroughly evaluate the impact of each objective function, we perform ablation studies on three downstream tasks including cross-modal retrieval, within-modal retrieval, and recognition on the ModelNet40 dataset and report the performance in Table 4. The L_I is to enforce the network to capture modality-invariant features, and it alone achieves relative high performance for both retrieval and recognition tasks. The L_S is to enforce the network to capture modality-specific features and the learned features are not modality-invariant which leads to low performance for the cross-modal retrieval task. When the two objective functions L_I and L_S are jointly employed, the performance of the recognition and most of the retrieval tasks are improved indicating the two objective functions are complementary with each other. When all three objective functions are used, our model achieves the best performance for all tasks demonstrating the effectiveness of our objective function design.

Ablation Study of Different Modalities. Our model is jointly trained on the data with three modalities including image, point cloud, and mesh. Ideally, more modalities of data can provide the network with more multimodal correspondences and can potentially lead to better performance. To thoroughly evaluate the impact of the number of modalities, we conduct ablation studies by only training our framework with two modalities and report the performance on both the recognition and retrieval tasks in Table 5. When only trained with two

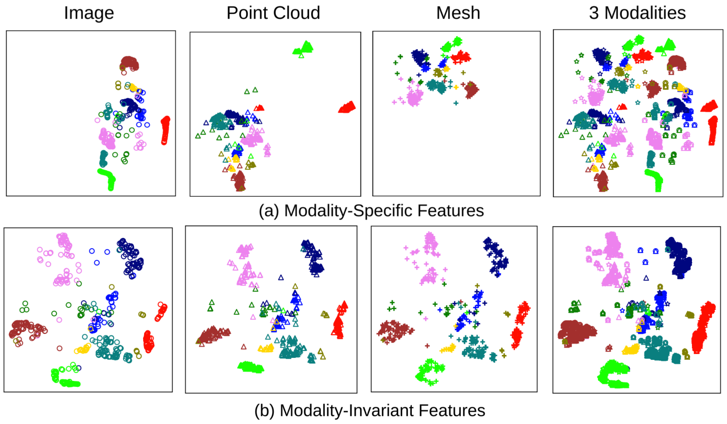

Figure 3: The qualitative visualization of modality-specific and modality-invariant features on the ModelNet40 dataset.

modalities, the performance for both the retrieval and recognition tasks are lower than all the three modalities are used during training. These results are consistent with our hypothesis that more modalities of data can provide more supervision signals which can lead to better performance on downstream tasks.

4.5 Comparing with Multimodal Supervised Methods

Modality	Supervised (%)	Ours (Self-Supervised) (%)
Point Cloud (Li et al., 2019a)	91.85	**92.62**
Mesh (Feng et al., 2019)	83.70	**92.62**
Image (Su et al., 2015)	92.51	91.07
3 Modalities	93.83	**94.05**

Table 6: Performance comparison of our **self-supervised** learning method over the **supervised** counterparts for 3D object recognition on the ModelNet10 dataset.

To demonstrate the advantage of the multimodal self-supervised learning, we compare supervised learning and our proposed self-supervised learning method with the **same** backbone networks for 3D object recognition on the ModelNet10 dataset. For the supervised training, we follow the exact setting as proposed in the papers (Feng et al., 2019; Li et al., 2019a; Su et al., 2015). For the self-supervised learning, the features are extracted by our learned models and then a Support Vector Machine (SVM) is used for performing the recognition.

The performance comparison are shown in Table 6. Even with a linear classifier, our proposed self-supervised learning method achieves comparable performance with the supervised method using the same backbone network. When the features from multiple modalities are fused together, the performance is significantly improved for both methods. These results confirm the potential of self-supervised multimodal feature learning.

4.6 Qualitative Feature Visualization

To visually demonstrate the ability of learning modality-specific and modality-invariant features, we compare qualitative visualization of features from the ModelNet40 dataset by using the t-SNE method (Maaten & Hinton, 2008). As shown in Fig. 3 (a), the distributions of the modality-specific features from three modalities are different which confirms that the learned features are indeed modality-specific. Fig. 3 (b) shows that the modality-invariant features extracted from different modalities have similar distributions and mixed in the common space indicating that the features are indeed modality-invariant.

5 Conclusion

In this paper, we have proposed a novel self-supervised learning method to jointly learn both modality-invariant and modality-specific features from unlabeled 3D datasets. The features learned from different modalities have been extensively evaluated on different tasks. Our method significantly outperforms other self-supervised learning methods on multiple downstream tasks. The multimodal features learned by our model even achieves comparable performance with the most recent state-of-the-art supervised methods on some tasks, indicating that the self-supervised multimodal feature learning for 3D object is a promising research direction.

References

Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. *arXiv preprint arXiv:1707.02392*, 2017.

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. *arXiv preprint arXiv:1911.05371*, 2019.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018.

Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. In *Computer graphics forum*, pp. 223–232. Wiley Online Library, 2003.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pp. 1–9, 2009.

Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1933–1941, 2016.

Yutong Feng, Yifan Feng, Haoxuan You, Xibin Zhao, and Yue Gao. Meshnet: mesh neural network for 3d shape representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 8279–8286, 2019.

Matheus Gadelha, Rui Wang, and Subhransu Maji. Multiresolution tree networks for 3d point cloud processing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 103–118, 2018.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.

Simon Ging, Mohammadreza Zolfaghari, Hamed Pirsiavash, and Thomas Brox. Coot: Cooperative hierarchical transformer for video-text representation learning. *arXiv preprint arXiv:2011.00597*, 2020.

Rohit Girdhar, David F Fouhey, Mikel Rodriguez, and Abhinav Gupta. Learning a predictable and generative vector representation for objects. In *European Conference on Computer Vision*, pp. 484–499. Springer, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pp. 2672–2680, 2014.

Kaveh Hassani and Mike Haley. Unsupervised multi-task feature learning on point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8160–8171, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Longlong Jing and Yingli Tian. Self-supervised spatiotemporal feature learning by video geometric transformations. *arXiv preprint arXiv:1811.11387*, 2(7):8, 2018.

Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *arXiv preprint arXiv:1902.06162*, 2019.

Longlong Jing, Yucheng Chen, Ling Zhang, Mingyi He, and Yingli Tian. Self-supervised feature learning by cross-modality and cross-view correspondences. *arXiv preprint arXiv:2004.05749*, 2020.

Edward Johns, Stefan Leutenegger, and Andrew J Davison. Pairwise decomposition of image sequences for active multi-view recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3813–3822, 2016.

Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, volume 6, pp. 156–164, 2003.

Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NIPS*, pp. 7773–7784, 2018.

Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, 2017.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 201–216, 2018.

Chun-Liang Li, Manzil Zaheer, Yang Zhang, Barnabas Poczos, and Ruslan Salakhutdinov. Point cloud gan. *arXiv preprint arXiv:1810.05795*, 2018a.

Guohao Li, Matthias Muller, Ali Thabet, and Bernard Ghanem. Deepgcns: Can gcns go as deep as cnns? In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9267–9276, 2019a.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4654–4662, 2019b.

Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on χ-transformed points. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 828–838, 2018b.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Jiageng Mao, Xiaogang Wang, and Hongsheng Li. Interpolated convolutional networks for 3d point cloud understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1578–1587, 2019.

Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *IROS*, pp. 922–928. IEEE, 2015.

Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. *arXiv preprint arXiv:1912.01991*, 2019.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *ECCV*, pp. 527–544. Springer, 2016.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual represions by solving jigsaw puzzles. In *ECCV*, 2016.

Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *CVPR*, volume 2, 2017.

Yuxin Peng, Xin Huang, and Yunzhen Zhao. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on circuits and systems for video technology*, 28(9):2372–2385, 2017.

Yuxin Peng, Jinwei Qi, and Yuxin Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27(11):5585–5599, 2018.

Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):521–535, 2013.

Omid Poursaeed, Tianxing Jiang, Han Qiao, Nayun Xu, and Vladimir G. Kim. Self-supervised learning of point clouds via orientation estimation. In *8th International Conference on 3D Vision, 3DV 2020, Virtual Event, Japan, November 25-28, 2020*, pp. 1018–1028. IEEE, 2020.

Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *CVPR*, pp. 5648–5656, 2016.

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017a.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pp. 5099–5108, 2017b.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pp. 139–147. Association for Computational Linguistics, 2010.

Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, 105(3):222–245, 2013.

Jonathan Sauder and Bjarne Sievers. Self-supervised deep learning on point clouds by reconstructing space. In *Advances in Neural Information Processing Systems*, pp. 12942–12952, 2019.

Abhishek Sharma, Oliver Grau, and Mario Fritz. Vconv-dae: Deep volumetric shape learning without object labels. In *European Conference on Computer Vision*, pp. 236–250. Springer, 2016.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised Learning of Video Representations using LSTMs. In *ICML*, 2015.

Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pp. 945–953, 2015.

Yongbin Sun, Yue Wang, Ziwei Liu, Joshua E Siegel, and Sanjay E Sarma. Pointgrow: Autoregressively learned point cloud generation with self-attention. *arXiv preprint arXiv:1810.05591*, 2018.

Ali Thabet, Humam Alwassel, and Bernard Ghanem. Mortonnet: Self-supervised learning of local features in 3d point clouds. *arXiv preprint arXiv:1904.00230*, 2019.

Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6411–6420, 2019.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pp. 4534–4542, 2015.

Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 154–162, 2017.

Chu Wang, Babak Samari, and Kaleem Siddiqi. Local spectral graph convolution for point set feature learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 52–66, 2018.

Limin Wang, Yuanjun Xiong, Zhe Wang, and Yu Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015.

Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (TOG)*, 38(5):1–12, 2019.

Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in neural information processing systems*, pp. 82–90, 2016.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Yifan Xu, Tianqi Fan, Mingye Xu, Long Zeng, and Yu Qiao. Spidercnn: Deep learning on point sets with parameterized convolutional filters. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 87–102, 2018.

Yaoqing Yang, Chen Feng, Yiru Shen, and Dong Tian. Foldingnet: Point cloud auto-encoder via deep grid deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 206–215, 2018.

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

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, pp. 649–666. Springer, 2016.

Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5565–5573, 2019a.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip Torr, and Vladlen Koltun. Point transformer. *arXiv preprint arXiv:2012.09164*, 2020.

Yongheng Zhao, Tolga Birdal, Haowen Deng, and Federico Tombari. 3d point capsule networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1009–1018, 2019b.

Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10394–10403, 2019.

A Appendix

A.1 Framework Architecture

As shown in the Fig. 2 in the main paper, our framework consists of three heterogeneous backbone networks including an image feature encoder, a point cloud feature encoder, and a mesh feature encoder. The MeshNet (Feng et al., 2019), dynamic graph convolutional neural network (DGCNN) (Wang et al., 2019), and ResNet (He et al., 2016) are employed as backbone networks to extract representation features from mesh, point cloud, and image, respectively. The architecture of backbone networks is described as follows.

Point Cloud Feature Encoder: The 3D point cloud feature learning network (G_p) employs DGCNN as the backbone model due to its capability to model local structures of each point by dynamically constructed graphs. There are four EdgeConv layers and the number of kernels in each layer is 64, 64, 64, and 128, and the EdgeConv layers aim to construct graphs over k nearest neighbors calculated by KNN and the features for each point are calculated by an MLP over all the k closest points. After the four EdgeConv blocks, a 512-dimension fully connected layer is used to extract per-point features for each point and then a max-pooling layer is employed to extract global features for each object.

Mesh Feature Encoder: The backbone architecture for mesh data is MeshNet, denoted as G_m. MeshNet contains three main blocks: spatial descriptor, structural descriptor, and mesh convolution block. The spatial descriptor applies fully-connected layers (64, 64) to extract spatial features from face's center. The structural descriptor contains a face rotate convolution within fully-connected layers (32, 32) and (64, 64), and a face kernel correlation with 64 kernels. Two mesh convolution blocks are used to aggregate features with neighboring information which the input/output channels of spatial and structural features are configured as (64, 131, 256, 256) and (256, 256, 512, 512), respectively. After the two mesh convolution blocks, a fully-connected layer (1024) further fuses the neighboring features and a max-pooling layer is employed to extract 512-dimension global features from the aggregated features.

Image Feature Encoder: ResNet18 is employed as the image feature capture network (G_{img}) for 2D images. It contains four convolution blocks with a number of $\{64, 128, 256, \text{and } 512\}$ kernels. Each convolution block includes two convolution layers followed by a batch-normalization layer and a ReLU layer, except the first convolution block which consists of one convolution layer, one batch-normalization layer, and one max-pooling layer. A global average pooling layer, after the fourth convolution blocks, is used to obtain the global features for each image. Unless specifically pointed out, a 512-dimensional vector after the global average pooling layer is used for all our experiments.

A.2 Ablation Study for Number of Views for Image Features

For each object, the features from multiple views of images are extracted and then averaged to obtain the object-level features. Ideally, higher performance should be obtained when more views of images are available since the features of images from different perspectives are complementary with each other. To evaluate the impact of the number of views for image features, we conduct experiments to evaluate the recognition performance with different number of views on the ModelNet40 dataset.

As shown in Table 7, when only one view of images is available, the performance of recognition is only 81.65%, and the performance is significantly boosted when more views of images are available. These results are consistent with our hypothesis that more views of images lead to better performance.

# Views	Recognition Accuracy(%)
1	81.65
2	84.31
4	85.56
8	86.83
16	87.47
32	88.04

Table 7: Ablation study for evaluating impact of numbers of images for the 3D recognition task. ”# Views” indicates how many views of images are used to obtain the image features.

Task	Results	
Cross-Modal Retrieval	Image2Mesh	65.67
	Image2Point	65.19
	Point2Image	65.97
	Mesh2Point	64.95
	Point2Mesh	65.03
	Mesh2Image	66.35
Within-Modal Retrieval	Image2Image	66.44
	Point2Point	65.30
	Mesh2Mesh	66.16
Recognition	Multimodal	94.05

Table 8: More results for the 3D cross-modal retrieval, 3D within-modal retrieval, and 3D multimodal recognition on the ModelNet10 dataset.

A.3 More Results on the ModelNet10 Dataset

Due to the space limitation, we only reported the performance for the recognition task on the ModelNet10 dataset in the main paper. Table 8 shows a complete results of our model for the 3D cross-modal retrieval, 3D within-modal retrieval, and 3D multimodal recognition on the ModelNet10 dataset. As shown in both Table 8 and Table 6 in the main paper, our proposed self-supervised learning method achieves comparable performance with the supervised method using the same backbone network for the recognition task. These results demonstrate the effectiveness of our design.

B Qualitative Visualization of Cross-Modal Retrieval

Fig. 4 shows the top-10 retrieval results for four different queries from the ModelNet40 dataset. The similarity between two objects are measured by Euclidean distance over the $L1$ normalized modality-invariant features. The results show that the objects with similar appearance are closer in the feature space even though they are from different modalities which confirm that the network indeed can learn modality-invariant features from unlabeled data.

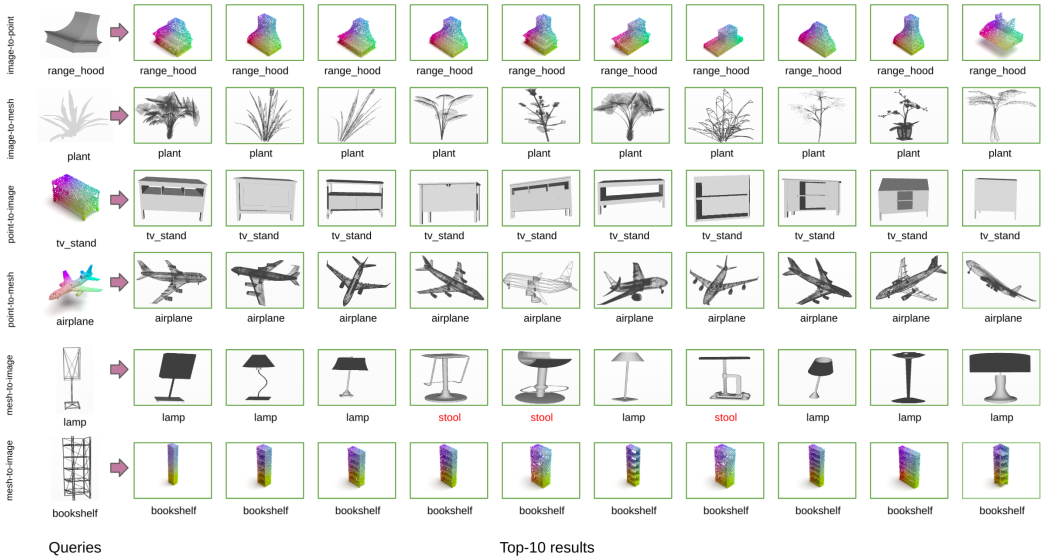
image-to-point
range_hood
range_hood range_hood range_hood range_hood range_hood range_hood range_hood range_hood range_hood range_hood
image-to-mesh
plant
plant plant plant plant plant plant plant plant plant plant
point-to-image
tv_stand
tv_stand tv_stand tv_stand tv_stand tv_stand tv_stand tv_stand tv_stand tv_stand tv_stand
point-to-mesh
airplane
airplane airplane airplane airplane airplane airplane airplane airplane airplane airplane
mesh-to-image
lamp
lamp lamp lamp stool stool lamp stool lamp lamp lamp
mesh-to-image
bookshelf
bookshelf bookshelf bookshelf bookshelf bookshelf bookshelf bookshelf bookshelf bookshelf bookshelf
Queries
Top-10 results

Figure 4: The Top-10 ranking for six query samples on cross-modal retrieval on the ModelNet40 dataset by our models. All the top-10 selected samples have very similar appearance as the query data.