Cross-Modal Center Loss for 3D Cross-Modal Retrieval

Longlong Jing* Elahe Vahdani* Jiaxing Tan Yingli Tian
The City University of New York

Abstract

Cross-modal retrieval aims to learn discriminative and modal-invariant features for data from different modalities. Unlike the existing methods which usually learn from the features extracted by offline networks, in this paper, we propose an approach to jointly train the components of cross-modal retrieval framework with metadata, and enable the network to find optimal features. The proposed end-to-end framework is updated with three loss functions: 1) a novel cross-modal center loss to eliminate cross-modal discrepancy, 2) cross-entropy loss to maximize inter-class variations, and 3) mean-square-error loss to reduce modality variations. In particular, our proposed cross-modal center loss minimizes the distances of features from objects belonging to the same class across all modalities. Extensive experiments have been conducted on the retrieval tasks across multi-modalities including 2D image, 3D point cloud and mesh data. The proposed framework significantly outperforms the state-of-the-art methods for both cross-modal and in-domain retrieval for 3D objects on the ModelNet10 and ModelNet40 datasets.

1. Introduction

With the stream of multimedia data flourishing on the Internet in the format of videos, images, text, etc, cross-modal retrieval task has attracted more and more attention from the multimedia communities. Cross-modal retrieval is the task of retrieving data from one modality given a query from a different modality. Inspired by the representation power of deep learning, a series of deep learning-based methods have been proposed for cross-modal retrieval [27, 52, 51]. These methods operate by learning modal-invariant representations in a common space.

The features from different modalities generally have different distributions. Therefore, a fundamental requirement for cross-modal retrieval task is to bridge the gap among different modalities which is commonly done by representation learning. The existing methods mainly extract the features of each modality by offline pre-trained models, and apply a projection function to transfer the features into a common representation space. By this transformation, the similarity of features from different modalities can be directly measured. Hence, the main challenge during this process is to learn discriminative and modal-invariant features.

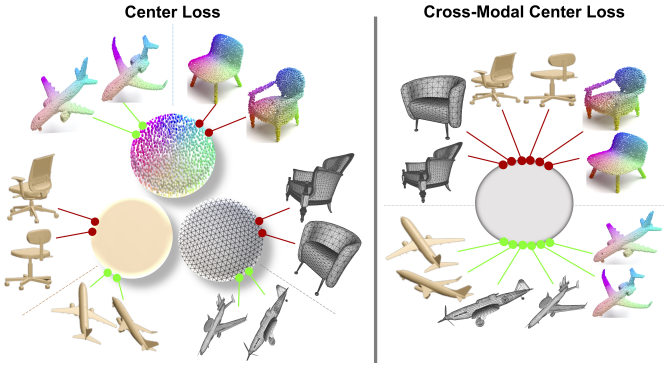

Figure 1. Traditional center loss vs. the proposed cross-modal center loss. Our proposed cross-modal center loss (right) finds a unique center for each class across all modalities. Traditional center loss (left) finds a center for each modality and each class and ignores the relation among centers of different modalities. Our proposed cross-modal center loss specifically eliminates the discrepancy across multiple modalities and thus is very effective for learning modal-invariant features.

By learning discriminative features, we ensure that data from the same class are mapped closely to each other in the feature space while different classes are separated as far as possible. In many studies, cross entropy or mean square error loss in the label space are used to maximize the inter-class variations. In order to compare the features extracted from different modalities, the features need to be modal-invariant. Various methods are proposed to reduce the cross-domain discrepancy by using adversarial loss, sharing a projection network, using triplet loss with pairs/triplets of different modalities, maximizing cross-modal pairwise item correlation [29, 42, 34, 20, 10].

Even though the existing methods [42, 51] achieved promising results in the cross-modal retrieval tasks, they

* Equal contribution.

suffer from the following limitations: 1) Their core idea is to minimize the cross-modal discrepancy over the features from multiple modalities extracted by pre-trained neural networks. For example, in the task of image-text retrieval, image and text features are extracted by pre-trained models(VGG [37] and SentenceCNN [21]), and then learning is performed on these extracted features instead of the metadata. Because these feature extractors (VGG, SentenceCNN) are not trained or finetuned for cross-modal retrieval task, they are not optimally representative. Instead, the network should be jointly trained with multimodal data to fully address the retrieval task. 2) The existing loss functions are mainly designed for two types of modalities, mainly image and text, and may not generalize well for cases when more than two modalities are available. It is essential to develop a simple yet effective loss function that can be easily extended for multiple modalities.

In this paper, we propose a new loss function, called Cross-modal Center Loss, specifically designed to minimize the intra-class variation across multiple modalities. Our loss function is directly inspired by the traditional unimodal center loss which learns a center for each class and minimizes the distance between objects and their corresponding centers in the feature space. Fig. 1 shows the comparison between the traditional center loss and the newly proposed cross-modal center loss. Having multi-modal data, the traditional center loss minimizes the distance of objects and their centers in separate features spaces defined for each modality. Instead, our proposed cross-modal center loss learns a unique center C for each class in the common space of all modalities. Specifically, it minimizes the distance of multi-modal objects and their centers in the same common feature space for all modalities. When more multimodal data is available, the cross-modal center loss will be able to learn more reliable centers for each class in the common space.

With the proposed cross-modal center loss, the cross-modal discrepancy between different modalities of the data can be eliminated. The proposed cross-modal center loss can be employed in conjunction with other loss functions to jointly learn features for cross-modal retrieval task. To verify the effectiveness of the proposed loss function, we further propose an end-to-end framework for cross-modal retrieval task to learn discriminative and modal-invariant features. The proposed framework is optimized with three loss functions including the cross-entropy in the label space to learn discriminative features, the cross-modal center loss to specifically eliminate the cross-modal discrepancy in a universal space, and the mean square error loss to minimize the cross-modal distance per object. Furthermore, a weight sharing strategy is applied to learn modal invariant features in the common space.

Different from the previous cross-modal retrieval methods which extract the features of image or text by offline networks, we propose to jointly train the entire framework from the metadata without being limited by pre-trained models from other datasets. The effectiveness of the proposed framework is evaluated on a novel 3D cross-modal retrieval task which has not been explored by existing supervised methods. Our method significantly outperforms the recent state-of-the-art methods on 3D cross-modal retrieval task and in-domain retrieval task. The main contributions of this paper are summarized as follows:

- We propose a novel cross-modal center loss to map the representations of different modalities into a common feature space.
- We propose an end-to-end framework for cross-modal retrieval task by jointly training multiple modalities using the proposed cross-modal center loss. The proposed framework can be extended to various cross-modal retrieval tasks.
- The proposed framework significantly outperforms the state-of-the-art methods on cross-modal and in-domain retrieval tasks across images, point cloud, and mesh for 3D shapes. To the best of our knowledge, this is the first supervised learning method for object retrieval across 2D image and 3D point cloud and mesh data.

2. Related Work

Feature Learning for 3D Objects: 3D data are inherently multi-modal and can be represented in various ways such as point cloud, multi-view images, mesh, volumetric data, etc. Various deep learning-based methods have been proposed for 3D feature learning including unordered point cloud-based methods [23, 24, 30, 32, 41, 44, 47], multi-view images-based methods [38, 39], and volumetric voxelized data-based methods [5, 22, 31, 26, 40]. Qi *et al.* proposed the first deep learning-based model (i.e. PointNet) to directly learn the features from unordered point cloud data. To specifically model the local information for each point [30], Wang *et al.* proposed a dynamic graph convolution neural network (DGCNN) with EdgeConv using k nearest neighbor (KNN) points [44]. Su *et al.* proposed to learn the features for 3D objects with multi-view CNN operating on 2D images that are rendered from different views of 3D data [38]. MeshNet [8] and MeshCNN [12] were proposed to learn features directly from the mesh data by modelling the geometric relations of mesh faces of the object. Recently, few studies attempted to learn modal-invariant features with self-supervised learning. Jing *et al.* proposed MVI for modal and view-invariant feature learning by contrasting where the learned features can be used for cross-modal retrieval [17].

Cross-modal retrieval: Several methods have been proposed for cross-modal retrieval task, mainly targeting image-text retrieval. One straightforward solution for this task is to formulate the problem as a linear projection [16, 18, 33, 50]. Most recently, deep learning-based methods have been proposed for representation learning due to the powerful feature learning capability. As a deep version of CCA, Andrew *et al.* proposed deep canonical correlation analysis (DCCA) to adapt deep neural network to model the complex nonlinear transformations by projecting two highly linear correlated views into the same common space [2]. As a further step, Wang *et al.* proposed deep canonically correlated autoencoders (DCCAE) which is a two-autoencoder design and is jointly optimized by the combination of the canonical correlation between the learned representations and the reconstruction errors of the autoencoders [43]. Peng *et al.* proposed a two-stage framework called Cross-Media Deep Networks (CMDM) which acquires inter- and intra-modality features and then hierarchically combines the representations to further learn the rich cross-media correlations [28]. However, these deep learning-based methods did not concentrate on inter- and intra-modality relations in their designs. The CMDN later is extended by Peng *et al.* to cross-modal correlation learning (CCL) by adding inter-modal interactions in the first stage while adding intra-modal semantic constraints in the second stage [29].

To learn modal-invariant features, Wang *et al.* proposed adversarial cross-modal retrieval (ACMR) which adapted adversarial learning to minimize the domain gap by using a discriminator to predict the corresponding modality of the representations [42]. With the adversarial loss function, this method significantly outperformed the previous state-of-the-art methods on popular benchmarks with a large margin. Zhen *et al.* proposed deep supervised cross-modal retrieval (DSCMR) to learn the representations in the common space in regard to both inter-class and intra-class relations [51]. The DSCMR increases the inter-class variations via the discrimination loss in both the label space and the common representation space. Although the DSCMR achieved state-of-the-art performance on image-text retrieval task, our analysis show that this method have poor generalization ability to settings with diverse data samples.

Most of the existing work use the image and text features extracted by offline networks and directly minimize the cross-modal gap in the common space using these features. In this paper, we propose an end-to-end jointly trained framework and a novel cross-modal center loss to learn discriminative and modal-invariant features directly from metadata.

3. Methods

We propose an end-to-end framework with joint training of multiple modalities for cross-modal retrieval task based on the proposed cross-modal center loss. The overview of our proposed framework for 3D cross-modal retrieval task is shown in 2. As shown in the figure, The features for different modalities including Mesh, point cloud, and image are extracted by different networks, then these features are projected to a common space via two shared fully connected layers. The cross-modal discrepancy is eliminate in the universal space with our proposed loss functions. The formulation of the proposed cross-modal center loss is introduced in the following sections.

3.1. Problem Formulation

Dataset S contains N instances where the i-th instance t_i is a set of M modalities with a semantic label y_i. The set of modalities of t_i is denoted by s_i. Formally:

$$S = \{t_i\}_{i=1}^{N}, \quad t_i = (s_i, y_i), \quad s_i = \{x_i^m\}_{m=1}^{M}$$

Generally, the modality samples $\{x_i^1, x_i^2, \cdots, x_i^M\}$ are in M different representation spaces and their similarities cannot be directly measured. The goal of the cross-modal retrieval task is to learn M projection functions f_m for each modality $m \in [1, M]$, where $v_i^m = f_m(x_i^m, \theta_m)$ and θ_m is a learnable parameter. As a result, v_i^m is a projected feature in the common representation space. Distance between the projected features is a measure of similarity between the samples across all modalities. Therefore, samples from the same class should be mapped closely to each other independent of their modalities: $d(v_i^m, v_i^{m^*}) \sim$ low. On the other hand, samples from different classes should be projected as far as possible: $d(v_i^m, v_j^{m^*}) \sim$ high (where $i \neq j$)

3.2. Loss Function

The core of the cross-modal retrieval is to obtain discriminative and modal-invariant features for data of different modalities with heterogeneous networks. To learn discriminative features, we use the cross entropy loss over the sharing head of our network, while our proposed cross-modal center loss and mean square error help with learning modal-invariant features.

Cross-modal center Loss: Given the extracted features $\{v_i^m\}_{i=1}^{N}$ $(m \in [1, M])$ for N instances and M modalities, our proposed cross-modal center loss is formulated in Eq. 1:

$$L_c = \frac{1}{2} \sum_{i=1}^{N} \sum_{m=1}^{M} \|v_i^m - C_{y_i}\|_2^2, \tag{1}$$

where $C_{y_i} \in \mathbb{R}^k$ denotes the center of class y_i in the common space and k is the dimension of features. Com-

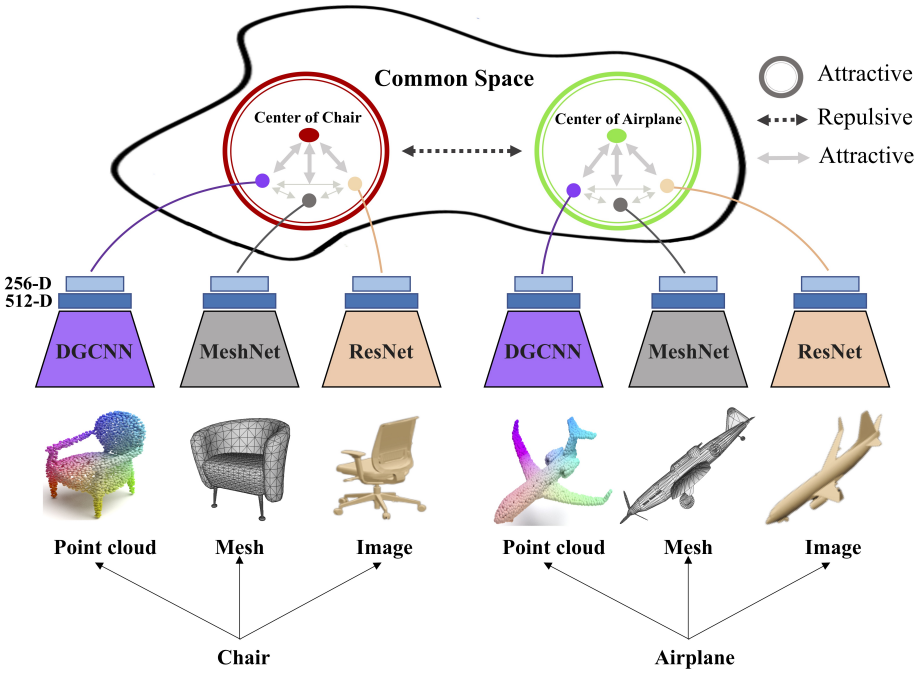

Figure 2. An overview of the proposed framework for 3D cross-modal retrieval task. Mesh, point cloud, and multi-view 2D image features are extracted by MeshNet, DGCNN, and ResNet, respectively, then projected to a common space via two shared fully connected layers. With the cross-modal center loss in conjunction with the cross-entropy loss and mean square error loss, the proposed framework can learn discriminative and modal invariant features.

paring to the original center loss [46], our proposed cross-modal center loss learns by eliminating the cross-modality gap and reducing the intra-class variation. To learn modal-invariant features, the cross-modal center loss optimizes the network to learn a center C_{y_i} for class y_i and minimize the distance between the features and their corresponding centers within each training batch. After each training iteration, the center of each class, C_j, is updated by $\triangle C_j$ with data from all modalities belonging to class j:

$$\triangle C_j = \frac{\sum_{i=1}^{N}\sum_{m=1}^{M}\delta(y_i = j)(C_j - v_i^m)}{1 + \sum_{i=1}^{N}\delta(y_i = j)}, \quad (2)$$

where

$$\delta(condition) = \begin{cases} 1 & condition = True \\ 0 & otherwise \end{cases} \quad (3)$$

Given a large batch size, the model can learn a robust center for each class, leading to produce features with small intra-class variation across all modalities. One advantage of the proposed cross-modal center loss is that it can be easily extended to more modalities. When data with more modalities are available, it provides more robust centers and may lead to better optimized features.

Discriminative Loss: To learn discriminative features, cross entropy loss in the label space is employed to optimize the network. Given N samples from M modalities, the discriminative loss is calculated by the cross-entropy loss between the MLP prediction $\hat{y_i^m}$ from each extracted feature v_i^m, and its label y_i.

$$L_d = -\frac{1}{N}(\sum_{i=1}^{N}\sum_{m=1}^{M} y_i^m \cdot log(\hat{y_i^m})), \quad (4)$$

where $\hat{y_i^m}$ is predicted by the two shared layers as:

$$\hat{y_i^m} = MLP(v_i^m). \quad (5)$$

Trained with cross-entropy loss, samples from the same category have higher similarities, while samples from different categories have lower similarities. Jointly trained with cross-modal center loss and cross-entropy loss, the network is able to learn both modal-invariant and discriminative features.

To further reduce the cross-modal discrepancy for each instance, we propose a loss function based on mean square error to minimize the distances between the features of all cross-modal sample pairs. The loss function across M modalities for each instance i is defined as the following where $\{v_i^1, v_i^2, \cdots, v_i^M\}$ are the extracted features:

$$L_m = \sum_{\alpha,\beta\in[1,M]\alpha\neq\beta} \left\| v_i^\alpha - v_i^\beta \right\|_2^2. \quad (6)$$

The three proposed loss functions are used to jointly train the network to learn discriminative and modal-invariant features:

$$Loss = \alpha_c L_c + \alpha_d L_d + \alpha_m L_m, \tag{7}$$

where α_c, α_d, and α_m are the weights for each loss term. Our proposed joint loss function in Eq. 7 can be optimised by stochastic gradient descent. The details of the optimization procedure is summarized in *Algorithm 1*.

Algorithm 1 Optimization procedure of the proposed framework

Require: The training data set $S = \{(t_i, y_i)\}_{i=1}^{n}$, the dimensionality of the common representation space k, the mini-batch size n_b, the learning rate τ , the maximal number of epochs $\mathcal{N}$.

Ensure: The optimized parameters in the M sub-networks $\theta_m, m \in [1, M]$.

Initialization : Randomly initialize the parameters of M subnetworks $\theta_m, m \in [1, M]$ and the parameters of the shared MLP classifier θ_P.

for $j = 1$ to $\mathcal{N}$ **do**
  for $b = 1$ to $\left\lfloor \frac{n}{n_b} \right\rfloor$ **do**
    Construct a training mini-batch by randomly selecting n_b samples from S.
    Extract the representations v_i^m for each sample x_i^m in the mini-batch by forward propagation, where $m \in [1, M]$, and $i \in [1, n_b]$.
    For each v_i^m, acquire the class prediction y_i^m by: $y_i^m = MLP(v_i^m)$
    Calculate the mini-batch training loss L by Eq. 7.
    Update the parameters of the entire network, where each part is updated by:
    a) Parameters of linear classifier P is updated by minimizing J in Eq. 7 with:
    $$\theta_P = \theta_P - \tau \frac{\partial J}{\partial \theta_P}$$
    b) Parameters of the sub-networks, θ_m, by minimizing J with descending their stochastic gradient:
    $$\theta_m = \theta_m - \tau \frac{\partial J}{\partial \theta_m}, \quad m \in [1, M]$$
    c) Center of each class is updated by Eq. 2.
  end for
end for

3.3. Framework Architecture

The proposed loss function can be applied to various cross-modal retrieval tasks. To verify the effectiveness of the proposed loss function, we designed an end-to-end framework for 3D cross-modal retrieval task to jointly train multiple modalities including image, mesh, and point cloud. The overview of the proposed framework for 3D cross-modal retrieval is shown in Fig. 2. As shown in the figure, there are three networks: $F(\theta)$ for image feature extraction, $G(\beta)$ for point cloud feature extraction, and $H(\gamma)$ for mesh feature extraction. Our framework can be easily extended to cases with more modalities or to different cross-modal retrieval tasks.

3D cross-modal retrieval. For 2D image feature extraction, we utilize ResNet18 [13] as the backbone network with four convolution blocks, all with 3×3 kernels, where the number of kernels are 64, 128, 256, and 512, respectively. Unless specifically mentioned, after the global average pooling, a 512-dimensional final feature vector is acquired in all experiments. Dynamic graph convolutional neural network (DGCNN)[45] is employed as the backbone model to capture point cloud features. DGCNN contains four EdgeConv blocks with the number of kernels set to 64, 64, 64, and 128. After the four EdgeConv block, a fully connected layer with 512 neurons is used to extract point-specific features for each point and then a max-pooling layer is applied to extract global features for each object. MeshNet [8] consists of 2 mesh convolution blocks, which achieved the state-of-the-art results for mesh retrieval, and is selected as the backbone to extract the features from mesh data. Two fully connected layers with size of 256 and 40 are employed to make classification predictions based on the 512-dimensional global features for all three modalities. The entire framework is trained from scratch for 3D cross-modal retrieval task with the proposed loss function.

4. Experiments

Datasets: The ModelNet40 [48] and ModelNet10 [48] datasets are used for evaluation. The ModelNet40 dataset is a 3D object benchmark and contains of $12,311$ CAD models which belong to 40 different categories with $9,843$ used for training and $2,468$ for testing, while ModelNet10 consists of $3,991$ CAD models for training and 909 models for testing belonging to 10 categories. Three modalities are provided in these two datasets including image, point cloud, and mesh.

4.1. Experimental Setup:

Evaluation Metrics: The evaluation results for all experiments are presented in terms of Mean Average Precision (mAP) score which is a classical performance evaluation criterion for cross-modal retrieval task [53, 7, 42]. The mAP for retrieval task is defined to measure whether the retrieved data belong to the same class as the query (relevant) or do not (irrelevant). Given a query and a set of R corresponding retrieved data (R top-ranked data), the Average Precision is defined as:

Source	Target	mAP-v1	mAP-v2	mAP-v4
Image	Image	82.06	86.00	90.23
Image	Mesh	85.58	87.31	89.59
Image	Point Cloud	85.23	86.79	89.04
Mesh	Image	83.58	85.96	88.11
Point Cloud	Image	82.29	85.18	87.11
Mesh	Mesh	88.51	—	—
Mesh	Point Cloud	87.37	—	—
Point Cloud	Mesh	87.58	—	—
Point Cloud	Point Cloud	87.04	—	—

Table 1. Performance of 3D in-domain and cross-modal retrieval task on ModelNet40 dataset in terms of mAP. When the target or source are from image domain, the results are reported for multi-view images: 1 view, 2 views, and 4 views denoted by v1, v2, and v4.

$$AP = \frac{1}{N}\sum_{r=1}^{R} p(r)\cdot\delta(r), \tag{8}$$

where N is the number of relevant data in the retrieved set, $p(r)$ is the precision of first r retrieved data, and $\delta(r)$ is the relevance of the r-th retrieved data (1 if relevant and 0 otherwise).

4.2. 3D Cross-modal Retrieval Task

To evaluate the effectiveness of the proposed end-to-end framework, we conduct experiments on ModelNet40 dataset with three different modalities including multi-view images, point cloud, and mesh. To thoroughly examine the quality of learned features, we conduct two types of retrieval tasks including in-domain retrieval when the source and target objects are from the same domain and cross-domain retrieval when they are from two different domains. When the target or source is from image domain, we evaluate the performance of multi-view images where the number of views is set to 1, 2 and 4. The performance of our method for 3D in-domain and cross-modal retrieval tasks is shown in Table 1.

As shown in Table 1, the proposed framework achieves more than 85% mAP for both in-domain and cross-domain retrieval tasks on ModelNet40 dataset. When the query or target are from the image-domain, the retrieval performance are significantly improved if more image views are used. Even though cross-modal center loss is specifically designed for learning modal invariant features, it is capable of discriminating the features of different classes within the same domain and achieves more than 86% mAP for Image2Image, Point2Point, and Mesh2Mesh in-domain retrieval tasks.

4.3. Impact of Loss Function

The three components of our proposed loss function are denoted as following: cross-entropy loss for each modality in the label space as L_1, cross-modal center loss in the universal representation space as L_2, and mean-square loss between features of different modalities as L_3. To further investigate the impact of each component, we evaluate different combinations for the loss functions including: 1) optimization with L_1, 2) jointly optimization with L_1 and L_3, 3) jointly optimization with L_1 and L_2, and 4) jointly optimization with L_1, L_2, and L_3. These four models are trained with the same setting and hyper-parameters, where the performance is shown in Table 2.

Loss	L_1	L_1+L_3	L_1+L_2	$L_1+L_2+L_3$
Image2Image	75.09	74.21	84.87	86.0
Image2Mesh	75.38	75.86	86.7	87.31
Image2Point	69.76	70.52	86.11	86.79
Mesh2Mesh	75.53	76.36	88.83	88.59
Mesh2Image	75.2	74.76	85.66	85.96
Mesh2Point	69.64	70.34	87.58	87.37
Point2Point	66.63	68.18	86.89	87.04
Point2Image	69.54	70.34	84.76	85.18
Point2Mesh	69.23	71.88	87.69	87.58

Table 2. The ablation studies for loss functions. L_1 is cross entropy loss, L_2 is cross-modal center loss, and L_3 is mean squared error loss. The number of views for images is fixed to 2.

As illustrated in Table 2, we have the following observations:

- The combination of L_1, L_2 and L_3 achieves the best performance for all cross-modal and in-domain retrieval tasks.
- As the baseline, cross-entropy loss alone achieves relatively high mAP due to the sharing head of the three modalities forcing the network to learn similar representations in the common space for different modalities of the same class.
- By adding cross-modal center loss to cross entropy loss, a constant and significantly improvement in mAP, between 7% to 20%, could be achieved for different retrieval tasks, proving that the proposed cross-modal center loss could significantly reduces the cross-modal discrepancy.
- Particularly, performance of Point2Mesh, Point2Point, and Mesh2Point retrieval tasks are improved by nearly 20% which further validates the effectiveness of the proposed cross-modal center loss.
- Adding the MSE loss to cross entropy and cross-modal center loss also slightly improves the performance.

4.4. Impact of Batch Size

The core idea of the proposed cross-modal center loss is to learn a unique center for each class and to minimize the

distance of data from different modalities in that class to its center. However, calculation based on the whole dataset in each update is inefficient even practical [46]. As a result, the center for each class is defined as the average of features for that class in a mini-batch and updated with optimizer. Therefore, the reliability of the features for each class is highly correlated with the batch size. Using a large enough batch size provides sufficient samples for each class to find a reliable center, while having a small batch size leads to unreliable centers. To analyze the impact of batch sizes to the performance, we conduct experiments on ModelNet40 dataset with different batch sizes $(12, 24, 48, 96)$. The results are shown in Table 3. All models are trained with the same number of epochs and same hyper-parameters.

Batch Size	12	24	48	96
Image2Image	45.67	63.56	85.64	90.23
Image2Mesh	13.89	73.22	86.94	89.59
Image2Point	32.32	72.08	85.59	89.04
Mesh2Mesh	25.5	88.44	88.91	88.51
Mesh2Image	6.98	68.81	86.5	88.11
Mesh2Point	8.29	84.6	86.67	87.37
Point2Point	59.5	82.44	85.44	87.04
Point2Image	27.68	67.46	84.67	87.11
Point2Mesh	15.87	83.56	86.62	87.58

Table 3. The ablation studies for the batch size on the ModelNet40 dataset. The number of views for images is fixed to 4. Same number of epochs are used for all the experiments.

As shown in Table 3, changing the batch size from 12 to 96 significantly improves the performance for all modalities. Due to the limitations of the GPU memory, the largest batch size that we tested is 96. This results indicate that a larger batch size should be used for the proposed cross-modal center loss whenever possible.

4.5. Comparison with Existing Methods on 3D Retrieval

In this section, we compare the performance of our method with the state-of-the-art methods on 3D in-domain and cross-modal retrieval tasks in both ModelNet10 and ModelNet40 datasets.

4.5.1 Comparing with the state-of-the-art Cross-Modal Retrieval Methods

Since there is no method specifically designed for 3D cross-modal retrieval task yet, we re-produce the current state-of-the-art method (DSCMR [51]) that designed for image-text retrieval task. Since DSCMR was originally designed only for image-text retrieval, we extend it to three types of modalities (image, point cloud, and mesh) and jointly trained it on 3D datasets. We conduct experiments for 3D cross-modal retrieval on both ModelNet10 and ModelNet40 datasets.

Method	DSCMR [51]	Ours
Image2Image	82.31	**90.23**
Image2Mesh	77.30	**89.59**
Image2Point	74.33	**89.04**
Mesh2Mesh	74.84	**88.51**
Mesh2Image	76.18	**88.11**
Mesh2Point	70.21	**87.37**
Point2Point	70.80	**87.04**
Point2Image	73.74	**87.11**
Point2Mesh	71.59	**87.58**

Table 4. Comparison with the state-of-the art method on ModelNet40 dataset for 3D cross-modal retrieval task. The number of views for images is fixed to 4. The DSCMR has poor generalization ability of extending to diverse datasets. The proposed jointly trained method significantly outperforms the state-of-the-art method on all retrieval tasks.

As shown in Table 4 and Table 5, our proposed method significantly outperforms the state-of-the-art methods for all of the retrieval tasks on the two benchmarks. The ModelNet10 only consists of 10 categories of data and the DSCMR performs well on this small dataset. However, when extending to ModelNet40 which consists of data belonging to 40 classes, the performance of DSCMR is significantly worse than our proposed method showing that the DSCMR has poor generalization ability when extend to more classes and more diverse dataset. Compared to DSCMR, our method obtained significantly better performance on all the retrieval pairs on both datasets showing our proposed method has very strong generalization ability.

4.5.2 Comparing with the State-of-the-art In-Domain Retrieval Methods

Although designed for cross-modal retrieval task, our model and loss function can be easily extended to in-domain retrieval task. Following the prior state-of-the-art methods for 3D in-domain retrieval task [5, 38, 19, 36, 8, 3, 4, 9, 15, 11, 14, 14], We compare the performance of in-domain 3D object retrieval task on ModelNet40 dataset with different modalities. As shown in Table 6, our method outperforms all the state-of-the-art methods on ModelNet40 dataset validating again the strong generalization ability of our proposed method.

4.6. Qualitative Visualization

T-SNE Feature Embedding Visualization: Fig. 3 (a), (b), and (c) show that the features are distributed as sepa-

Method	DSCMR [51]	Ours
Image2Image	84.49	**91.75**
Image2Mesh	84.09	**91.23**
Image2Point	81.73	**91.37**
Mesh2Mesh	83.92	**90.41**
Mesh2Image	82.52	**89.98**
Mesh2Point	80.81	**90.00**
Point2Point	83.08	**90.99**
Point2Image	84.15	**90.73**
Point2Mesh	84.37	**90.92**

Table 5. Comparison with the state-of-the art methods on ModelNet10 dataset for 3D cross-modal retrieval task. The number of views for images is fixed to 4. The proposed method significantly outperforms the state-of-the-art method.

Method	Domain	MAP
SPH [19]	Mesh	33.3
LFD [6]	Image	40.9
3DShapeNet [5]	Volume	49.2
Deeppano [36]	Image	76.8
MVCNN [38]	Image	80.2
MeshNet [8]	Mesh	81.9
GIFT [3]	Image	81.9
SPNet [49]	Image	85.2
RED [4]	Volume	86.3
Panorama-ENN [35]	Image	86.3
DLAN [9]	Point	85.0
TCL [15]	Image	88.0
SequenceView [11]	Image	89.1
VNN [14]	Image	89.3
ADCNN [1]	Image	91.1
Ours	Mesh	**90.41**
Ours	Point	**90.99**
Ours	Image	**91.75**

Table 6. Comparison with the state-of-the-art in-domain retrieval methods for 3D objects on ModelNet40 Dataset. The number of views for images is 4 for our method. Our method outperforms all the other methods that are specifically designed for in-domain retrieval for 3D data.

rated clusters, demonstrating that the proposed loss is able to discriminate the samples from different classes for each modality. From Fig. 3 (d), the features from three different modalities are mixed together showing that the features learned by the proposed framework in the universal space are indeed model-invariant.

Cross-Modal Retrieval Visualization: Fig. 4 shows the cross-modal retrieval samples for six different queries from ModelNet40 dataset. For each query, the euclidean distance over the normalized features is used to measure the similarity of data from different modalities. The Top-10 closest samples for each query data are visualized. The figure shows the objects with similar appearance are closer in the features space even though they are from different modalities, proving that the network indeed learned model-invariant features.

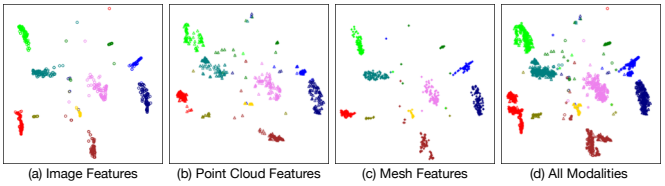

Figure 3. The visualization for the testing data in the ModelNet40 dataset by using t-SNE method [25]. Each point in the figure represents one object. Objects from the same category are rendered with the same color.

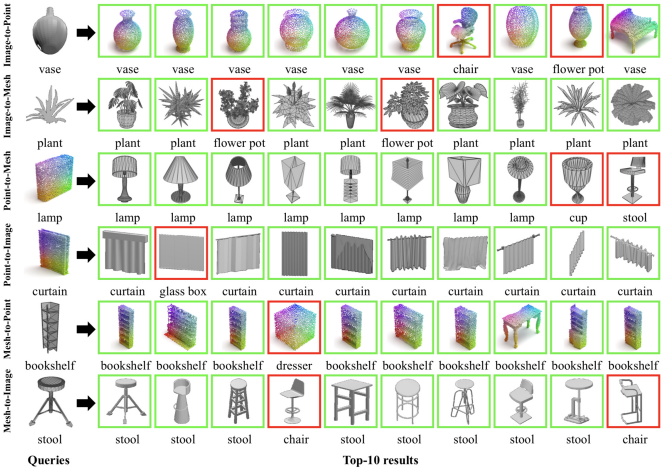

Figure 4. Top-10 retrieval results for six query samples on ModelNet40 dataset with our models. The green bounding boxes indicate that the images belong to the same category as the query, whereas the red bounding boxes indicate wrong matches.

5. Conclusion

In this paper, we have proposed a cross-modal center loss to learn discriminative and modal-invariant features for cross-modal retrieval tasks. The proposed cross-modal center loss significantly reduces the cross-modal discrepancy by minimizing the distances of features belonging to the same class across all modalities, and can be used in conjunction with other loss functions. Extensive experiments have been conducted on retrieval tasks across multi-modalities including image, 3D point cloud and mesh data. The proposed framework significantly outperforms the state-of-the-art methods on the ModelNet40 dataset validating the effectiveness of the proposed cross-modal center loss and the end-to-end framework.

References

[1] Ahmad Alzu'bi, Abdelrahman Abuarqoub, and Ahmed Al-Hmouz. Aggregated deep convolutional neural networks for multi-view 3d object retrieval. In *2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, pages 1–5. IEEE, 2019.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255, 2013.

[3] Song Bai, Xiang Bai, Zhichao Zhou, Zhaoxiang Zhang, and Longin Jan Latecki. Gift: A real-time and scalable 3d shape search engine. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5023–5032, 2016.

[4] Song Bai, Zhichao Zhou, Jingdong Wang, Xiang Bai, Longin Jan Latecki, and Qi Tian. Ensemble diffusion for retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 774–783, 2017.

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[6] Ding-Yun Chen, Xiao-Pei Tian, Yu-Te Shen, and Ming Ouhyoung. On visual similarity based 3d model retrieval. In *Computer graphics forum*, volume 22, pages 223–232. Wiley Online Library, 2003.

[7] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16, 2014.

[8] Yutong Feng, Yifan Feng, Haoxuan You, Xibin Zhao, and Yue Gao. Meshnet: Mesh neural network for 3d shape representation. *AAAI 2019*, 2018.

[9] Takahiko Furuya and Ryutarou Ohbuchi. Deep aggregation of local 3d geometric features for 3d model retrieval. In *BMVC*, volume 7, page 8, 2016.

[10] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[11] Zhizhong Han, Mingyang Shang, Zhenbao Liu, Chi-Man Vong, Yu-Shen Liu, Matthias Zwicker, Junwei Han, and CL Philip Chen. Seqviews2seqlabels: Learning 3d global features via aggregating sequential views by rnn with attention. *IEEE Transactions on Image Processing*, 28(2):658–672, 2018.

[12] Rana Hanocka, Amir Hertz, Noa Fish, Raja Giryes, Shachar Fleishman, and Daniel Cohen-Or. Meshcnn: a network with an edge. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[13] 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*, pages 770–778, 2016.

[14] Xinwei He, Tengteng Huang, Song Bai, and Xiang Bai. View n-gram network for 3d object retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7515–7524, 2019.

[15] Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1945–1954, 2018.

[16] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992.

[17] Longlong Jing, Yucheng Chen, Ling Zhang, Mingyi He, and Yingli Tian. Self-supervised modal and view invariant feature learning. *arXiv preprint arXiv:2005.14169*, 2020.

[18] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):188–194, 2015.

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

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

[21] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[22] Roman Klokov and Victor Lempitsky. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *ICCV*, pages 863–872, 2017.

[23] Huan Lei, Naveed Akhtar, and Ajmal Mian. Spherical kernel for efficient graph convolution on 3d point clouds. *arXiv preprint arXiv:1909.09287*, 2019.

[24] 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*, pages 9267–9276, 2019.

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

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

[27] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8427–8436, 2018.

[28] Yuxin Peng, Xin Huang, and Jinwei Qi. Cross-media shared representation by hierarchical learning with multiple deep networks.

[29] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. Ccl: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Transactions on Multimedia*, 20(2):405–420, 2017.

[30] 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*, pages 652–660, 2017.

[31] 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*, pages 5648–5656, 2016.

[32] 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*, pages 5099–5108, 2017.

[33] Jan Rupnik and John Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, pages 1–4, 2010.

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[35] Konstantinos Sfikas, Ioannis Pratikakis, and Theoharis Theoharis. Ensemble of panorama-based convolutional neural networks for 3d model classification and retrieval. *Computers & Graphics*, 71:208–218, 2018.

[36] Baoguang Shi, Song Bai, Zhichao Zhou, and Xiang Bai. Deeppano: Deep panoramic representation for 3-d shape recognition. *IEEE Signal Processing Letters*, 22(12):2339–2343, 2015.

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] 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*, pages 945–953, 2015.

[39] Jong-Chyi Su, Matheus Gadelha, Rui Wang, and Subhransu Maji. A deeper look at 3d shape classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[40] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *ICCV*, pages 2088–2096, 2017.

[41] 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*, pages 6411–6420, 2019.

[42] 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*, pages 154–162, 2017.

[43] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015.

[44] 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.

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

[46] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pages 499–515. Springer, 2016.

[47] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2019.

[48] 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*, pages 1912–1920, 2015.

[49] Mohsen Yavartanoo, Eu Young Kim, and Kyoung Mu Lee. Spnet: Deep 3d object classification and retrieval using stereographic projection. In *Asian Conference on Computer Vision*, pages 691–706. Springer, 2018.

[50] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6):965–978, 2013.

[51] 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*, pages 10394–10403, 2019.

[52] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3550–3558, 2018.

[53] Yue Ting Zhuang, Yan Fei Wang, Fei Wu, Yin Zhang, and Wei Ming Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.