3D Cross-Modal Retrieval Using Noisy Center Loss and SimSiam for Small Batch Training

Yeon-Seung Choo[1,*], Boeun Kim[2], Hyun-Sik Kim[1], and Yong-Suk Park[1]
[1] Contents Convergence Research Center, Korea Electronics Technology Institute, Seoul, 03924 KOR
[e-mail: {piksal, hskim, yspark}@keti.re.kr]
[2] Artificial Intelligence Research Center, Korea Electronics Technology Institute, Seongnam-si, 13488 KOR
[e-mail: kbe36@keti.re.kr]
[*]Corresponding author: Yeon-Seung Choo

Received November 5, 2023; revised January 3, 2024; accepted February 27, 2024; published March 31, 2024

Abstract

3D Cross-Modal Retrieval (3DCMR) is a task that retrieves 3D objects regardless of modalities, such as images, meshes, and point clouds. One of the most prominent methods used for 3DCMR is the Cross-Modal Center Loss Function (CLF) which applies the conventional center loss strategy for 3D cross-modal search and retrieval. Since CLF is based on center loss, the center features in CLF are also susceptible to subtle changes in hyperparameters and external inferences. For instance, performance degradation is observed when the batch size is too small. Furthermore, the Mean Squared Error (MSE) used in CLF is unable to adapt to changes in batch size and is vulnerable to data variations that occur during actual inference due to the use of simple Euclidean distance between multi-modal features. To address the problems that arise from small batch training, we propose a Noisy Center Loss (NCL) method to estimate the optimal center features. In addition, we apply the simple Siamese representation learning method (SimSiam) during optimal center feature estimation to compare projected features, making the proposed method robust to changes in batch size and variations in data. As a result, the proposed approach demonstrates improved performance in ModelNet40 dataset compared to the conventional methods.

Keywords: Center Loss, Cross-Modal, Object Retrieval, Representation Learning, Self-Supervised Learning, Supervised Learning

This research was supported by the Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (Project Name: Open Metaverse Asset Platform for Digital Copyrights Management, Project Number: RS-2022-KC000812 (R2022020034), Contribution Rate: 100%).

http://doi.org/10.3837/tiis.2024.03.008 **ISSN : 1976-7277**

1. Introduction

Recent advances in sensor and hardware technology have brought forth improvements in deep learning methods. Data acquisition from 3D objects is no longer limited to image-based methods. Various modalities, such as images, meshes, and point clouds, are used to obtain 3D information. For example, Simultaneous Localization and Mapping (SLAM) and Light Detection and Ranging (LiDAR) use information from point clouds and meshes for 3D space and object reconstruction [1-5]. Neural Radiance Field (NeRF) is a technique that has become popular for 3D object reconstruction from a partial set of 2D images [6]. With the emergence of technologies that use different types of modalities, it has become necessary to understand the 3D object attributes at the raw data level and identify their relationship or relevance in different modalities.

Object retrieval is a task that retrieves objects with similar characteristics using feature comparison. A similarity-based search is performed on the extracted features of the objects. To enable efficient object retrieval in a supervised environment, it is necessary to extract the features that are distinctive for each object instance. Specifically, the inter-class variance should be high to provide a clear distinction among the different classes. To perform cross-modal search tasks, it is also necessary to maintain low inter-modal variance to cluster together features that show similar characteristics among the different modalities. The same class of objects in different modalities should exhibit similar feature distribution, making them rank higher when searching based on similarity criteria. Therefore, a different approach is needed to address cross-modal retrieval tasks.

One of the most prominent methods used for 3D Cross-Modal Retrieval (3DCMR) is the Cross-Modal Center Loss function (CLF) [7]. CLF extends the center loss approach, which was originally proposed for classification tasks, to the cross-modal domain [8]. Specifically, CLF uses the idea of converging to the center feature, which acts as an anchor in a 3D multi-modal environment. It demonstrated notable performance in 3DCMR with its unique and intuitive approach. While CLF benefits from the strengths of center loss, it also exhibits drawbacks inherent in center loss. Due to CLF's high reliance on central features, it is susceptible to external configuration variations, such as hyperparameters. For instance, there can be a significant performance difference depending on the selection of the batch size.

Furthermore, the Mean Square Error (MSE) employed in CLF calculates the Euclidean distance between modalities to reduce the inter-modal variance. However, the MSE, which simply focuses on representing the similarity between modalities, is unable to adapt to the changes when testing inference using different datasets. It also lacks robustness against changes in batch size for the same reason. To extract features that can reduce inter-modal variance in actual evaluation environments, a method is required that can respond to dataset changes during inference. Instead of using simple Euclidean distance, the overall context needs to be compared using projected features.

The method of performing feature comparisons based on projection features has been studied for a long time using various approaches. Among these, Simple Siamese Representation Learning (SimSiam) stands out as one of the most prominent methods [9]. SimSiam avoids the issue of model collapse by using only positive samples. Another distinctive feature of SimSiam is its robustness regardless of the batch size, as it performs training using only positive samples. Therefore, SimSiam could be applied to a multi-modality consisting of positive multi-modality samples and used for 3DCMR tasks.

In this paper, we propose a novel Noisy Cross-Modal Center Loss Function (NCF), an extension of CLF with added noise, to overcome degradation due to small batch sizes and to

establish anchor features that are robust in actual inference evaluation. The method also incorporates the use of projected features between modalities to enhance performance. The contributions offered by the proposed method are as follows:

- The proposed method utilizes noisy center features in addition to the robust center loss for training, aiming to enhance robustness compared to conventional approaches. Through this additional adjustment, the proposed method demonstrates improved performance in response to variations in actual experiments.
- The proposed method discards MSE and adapts SimSiam for 3DCMR tasks to reduce inter-modal variance. SimSiam trains similar attributions using only positive samples. Performance improvements have been observed by using projection features from positive samples and applying SimSiam when using small batch sizes.
- A quantitative comparison with the baseline CLF is conducted using ModelNet40 multi-modal 3D object dataset [10]. Both qualitative and quantitative results are discussed and analyzed.

2. Related Works

3DCMR is a methodology for searching objects across different modalities and faces the additional challenge of comparing inter-modal similarities. To address this issue, various methods have been proposed over time.

2.1 3D Object Feature Learning in Each Modality

Currently, 3D object models can be represented in different modalities, such as 2D images, meshes, and point clouds. With the development of deep learning, it has become essential to apply a suitable network for each modality to represent optimal features in constrained dimensions. Each modality uses different methods to capture distinctive features.

Various feature aggregation-based approaches from multi-view images have been predominantly proposed in the image modality. The representative approach is Multi-view Convolutional Neural Networks (MVCNN), which obtains multi-view images from 3D objects and extracts features for aggregation using CNN [11]. Following MVCNN, approaches such as Triplet-Center Loss (TCL) and Multi-View Transformation Network (MVTN) have been proposed [12, 13]. TCL aims to maximize margins through triplet-center loss. MVTN uses a differentiable renderer to acquire optimal viewpoint images. In addition, various other Graph Convolutional Network (GCN) based approaches have been suggested [14-16].

In the mesh modality, the representation process accompanies various components, such as vertices, faces, and neighboring indices. The emphasis is placed on the continuity among these components. MeshNet introduced a framework employing spatial and structural descriptors for triangular mesh data [17]. The framework attempted to represent optimal features employing both spatial discontinuity specificities, like vertices and neighboring indices. In addition to MeshNet, mesh representation methods have been proposed by several other approaches [18-20].

Unlike meshes, point clouds are composed of an unordered collection of individual points with no particular order and continuity. PointNet utilized Multi-Layer Perceptron (MLP) and max pooling to aggregate features for identifying structural features within these unordered point sets [21]. PointNet also employed a symmetric network structure to leverage permutation-invariant feature representation. Subsequently, various other convolutional networks have been proposed [22-25]. A graph-based K-Nearest Neighbor (KNN) approach

was introduced in Dynamic Graph Convolutional Neural Network (DGCNN) to capture geometric relationships between points, which led to significant performance improvements [26].

Likewise, the representation methods have been developed to suit the attributions of each modality. In cross-modal tasks, the objective is to integrate these networks and find representations of similar features among the different modalities.

2.2 Self-Supervised Learning of Visual Representations

Self-supervised learning, where the model trains itself without labels, is currently one of the most active research areas in artificial intelligence. Its research primarily focuses on using contrastive learning that leverages relative feature differences in specific samples to extract features that are as similar as possible or as different as possible. To achieve this, contrastive learning typically applies pairs of positive and negative samples while training.

The Simple Framework for Contrastive Learning of Visual Representations (SimCLR) method received much attention by studying data augmentation-based contrastive learning techniques and exhibiting a wide range of experiments [27]. Subsequently, the Momentum Contrast (MoCo) method addressed the computational complexity issue by comparing all negative samples [28]. MoCo suggested a new learning approach that stores and utilizes negative representations like a memory bank, to overcome the computational complexity caused by the excessive number of negative samples. However, these methods gradually displayed limitations, leading to the introduction of non-contrastive learning techniques that aim to perform learning using only positive samples.

In previous studies, the sole use of positive samples led to model collapse, where the network outputs almost identical values resulting in similar embedding vectors, causing ineffective learning. To prevent model collapse, Bootstrap Your Own Latent (BYOL) proposed the use of a momentum encoder and an asymmetric network structure with a stop gradient strategy [29]. As a result, the network was only updated once, enabling smooth network training. Following BYOL, SimSiam offered a solution for preventing model collapse using a simpler structure without using a momentum encoder while simultaneously improving performance. SimSiam also demonstrated modest to good efficiency across different batch sizes.

The model collapse phenomenon creates the illusion of loss convergence, where the network exhibits the same output for any training input values. If augmented data derived from the same image are used during the feature learning process, the extracted features will propagate consistently to the same network. As a result, the network will try to decrease loss easily by outputting constant values that are almost similar with no relevance to the input. To address this issue, methods that employ asymmetric structures and stop-gradient strategies (e.g., BYOL and SimSiam) as well as methods that use negative sample placement (e.g., SimCLR) have been proposed. The cross-modal SimSiam used in the proposed method prevents model collapse by using a different encoder network per modality. The results from backpropagation end up in different backbone networks. Consequently, networks learn distinctively for each modality, preventing the occurrence of model collapse phenomenon.

2.3 Feature Representation Learning in Multi-modality

Cross-modal retrieval has been previously researched extensively in image-point and image-text modalities. These studies have focused on representation learning which integrates various attributes across modalities to represent them at the same feature level. The results from the studies are currently being applied to perform various tasks.

CrossPoint is a cross-modal study conducted in the image-point modality [30]. CrossPoint estimates the similarity between image and point cloud data by applying self-supervised contrastive learning using inter-class and inter-modal comparison with data augmentation. In the proposed two-stage method, the inter-modal representations are learned by comparing augmented cloud point object data, and inter-modal representations are estimated by comparing 2D multi-view images. Following CrossPoint, similar methods have been proposed to extract features through cross-modal comparisons based on the self-supervised approach [31-34].

Methods that leverage multi-modality between images and text are one of the most widely researched areas. Deep Supervised Cross-modal Retrieval (DSCMR) tries to find a common representation space to compare images and text directly [35]. DSCMR demonstrates good performance by using three fundamental losses. First, it compares features in the common space after the backbone network. Next, the labels from the extracted label space are compared using the projection head. Lastly, cross-modal comparisons are performed to find consistent feature representations between images and text. DSCMR compares features at the semantic level using label predictions. Unlike DSCMR, Multimodal Contrastive Training (MCT) considers inter- and intra-modality relationships between images and text using contrastive learning [36]. Inspired by Supervised Contrastive Learning (SCL), MCT employs data augmentation to extract common features between relevant images and text [37].

In the field of 3DCMR, CLF serves as the reference baseline work. CLF extends center loss to encapsulate multi-modality. CLF extracts and compares features from all modalities and ensures that they converge to the center feature for the corresponding class regardless of the modality. This method yielded excellent results in the early stages of the research.

However, center loss-based methods are inherently vulnerable to external parameters, such as batch size. Significant performance variations can be observed depending on the batch size used. Moreover, it has been proven that the use of MSE does not adapt to changes robustly during the inference stage in actual evaluation. Therefore, we propose a method that improves upon CLF to overcome these drawbacks and to offer an efficient approach for 3DCMR.

3. Proposed Method

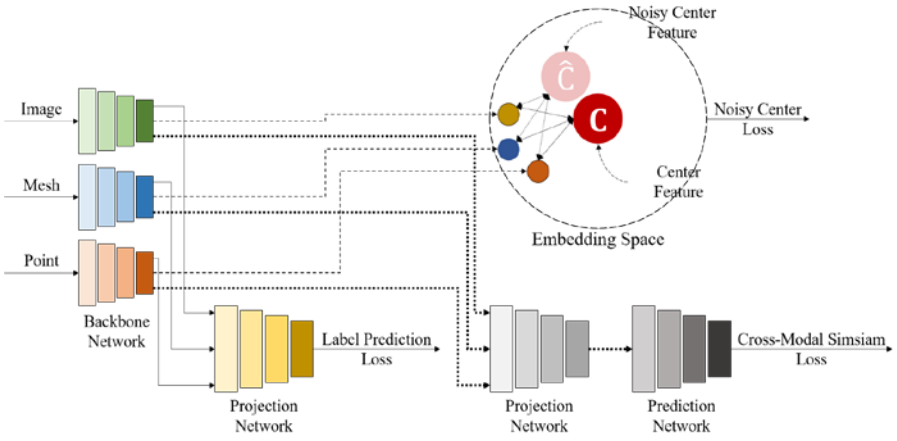

Fig. 1. The framework of the proposed method.

As previously pointed out, existing methods based on center features are sensitive to external factors. It has been observed that high-performance achievement is unattainable when small batch sizes are used. The proposed method consists of several processes that enable the extraction of robust features. **Fig. 1** shows the framework of the proposed method.

3.1 Preliminary Information

For the embedding features f that are extracted through the backbone network for each modality, C is the set of center features $c \in \mathbb{R}^{512}$, representing each class. In addition, f_b and c_{y_b} represent the features of the respective batch data b and the center feature c corresponding to its ground truth label, y_b. The conventional center loss can be expressed as in (1), where B represents the batch size.

$$\frac{1}{2}\sum_{b=1}^{B}\|f_b - c_{y_b}\|_2^2 \tag{1}$$

The cross-modal center loss used in CLF, which is derived from the conventional center loss, is shown in (2). In the equation, M stands for multi-modality. Center features are computed from batch data regardless of modality in CLF.

$$\frac{1}{2}\sum_{b=1}^{BM}\|f_b - c_{y_b}\|_2^2 \tag{2}$$

The initial equation for SimSiam loss is shown in (3). For each training image, SimSiam acquires two augmented data. Features z_1 and z_2 are projected through the encoder. Features p_1 and p_2 are extracted through the predictor using z_1 and z_2.

$$\frac{1}{2}\cos(p_1, z_2) + \frac{1}{2}\cos(p_2, z_1) \tag{3}$$

Cosine similarity, cos, is used for comparing feature similarities in (3) and can be expressed as shown in (4).

$$\cos(x_1, x_2) = -\frac{x_1}{\|x_1\|_2} \cdot \frac{x_2}{\|x_2\|_2} \tag{4}$$

In (3), model collapse occurs if training is performed using only positive samples. Therefore, the ‘stop gradient’ is needed to suppress the training of Encoder features z_1 and z_2. The results after applying the ‘stop gradient’ can be expressed as $\hat{z}_1$ and $\hat{z}_2$. The final SimSiam loss is as shown in (5).

$$\frac{1}{2}\cos(p_1, \hat{z}_2) + \frac{1}{2}\cos(p_2, \hat{z}_1) \tag{5}$$

3.2 Noisy Center Feature

The proposed method employs NCF to extract robust center features and adapts to variations in the data during the inference process. Through the NCF, the proposed method not only uses the center features extracted from the training but also incorporates comparisons with noise-added center features to robustly handle various changes. The noisy center loss L_C, where NCF is applied to multi-modality M, is represented by (6). G, σ, and θ denote the Gaussian noise function, the standard deviation of the noise, and the mean of the noise, respectively. Furthermore, the parameters α and β indicate their respective weights.

$$L_C = \frac{1}{2}\sum_{b=1}^{BM}\left(w_1 \left\| f_b - c_{y_b} \right\| + w_2 \left\| f_b - G\left(c_{y_b}, \sigma, \theta\right) \right\|\right) \tag{6}$$

3.3 SimSiam for Intra-class Variation

After enabling inter-class discrimination with the adaptation of the NCF, it is necessary to reduce inter-modal variance for extracting similar features regardless of the modalities. Therefore, to decrease the feature distribution within the same class, a loss function that utilizes feature comparisons is applied.

As previously mentioned, the conventional approach employed by the CLF, such as using the simple Euclidean distance in MSE, may not be an efficient method since the feature distribution will be different for the inference data. Therefore, it is necessary to apply a robust method that strongly reduces inter-modal variance to take changes in data into account. In addition, there is a need to apply a robust loss to deal with the sensitivity of the center features. At the same time, a method that comprehensively utilizes the samples in each modality is necessary.

The proposed method addresses these issues by employing cross-modal SimSiam, which utilizes positive samples to extract features for cross-modal comparison. For a given batch data, features extracted from different modalities are denoted as $Z = \{\hat{z}_m\}_{m=1}^{M}$ and $P = \{p_m\}_{m=1}^{M}$. Cross-modal SimSiam treats each modality features as positive samples of each other, transforming them to extract similar features. Equation (7) represents the proposed cross-modal SimSiam.

$$L_{CS} = \frac{1}{M \cdot (M-1)}\sum_{i=1}^{M-1}\sum_{j=i}^{M}\left\{\cos\left(p_i, \hat{z}_j\right) + \cos\left(p_j, \hat{z}_i\right)\right\} \tag{7}$$

In (7), the proposed method treats each modality as the augmented data of each other, thereby utilizing projected features to depict the similarity between different modalities in a multi-modal environment. By defining inter-modal similarity in this manner, the proposed method increases inter-class variance and decreases inter-modal variance simultaneously, enabling the creation of meaningful feature clusters. **Fig. 2** illustrates the difference between (a) the original SimSiam and (b) the proposed cross-modal SimSiam.

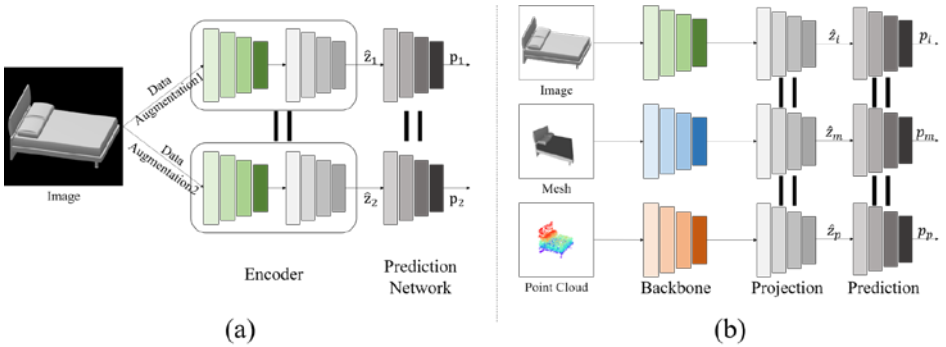

Fig. 2. The architecture of (a) SimSiam and (b) the proposed cross-modal SimSiam.

3.4 Loss Functions

The final formulation of the proposed method is as follows. First, the embedding features extracted through the backbone network are processed to obtain projected features using the MLP-based projection network, $Proj$. Subsequently, the label prediction loss L_P used for feature classification through label prediction is denoted as (8).

$$L_P = -\frac{1}{N}\left(\sum_{n=1}^{N}\sum_{m=1}^{M} y_n^m \cdot \log\left(Proj\left(f_n^m\right)\right)\right) \tag{8}$$

The final loss L, which combines the label prediction loss L_P, the noisy center loss L_C using NCF, and the cross-modal SimSiam loss L_{CS}, is represented in (9), where α, β, and γ represent the weights for each loss.

$$L = \alpha L_P + \beta L_C + \gamma L_{CS} \tag{9}$$

4. Experimental Results

To evaluate the effectiveness of the proposed method in performing 3DCMR tasks, we present the dataset used for evaluation and the details of the experimental procedure. In addition, we provide both quantitative and qualitative evaluation results in comparison to existing methods.

4.1 Dataset

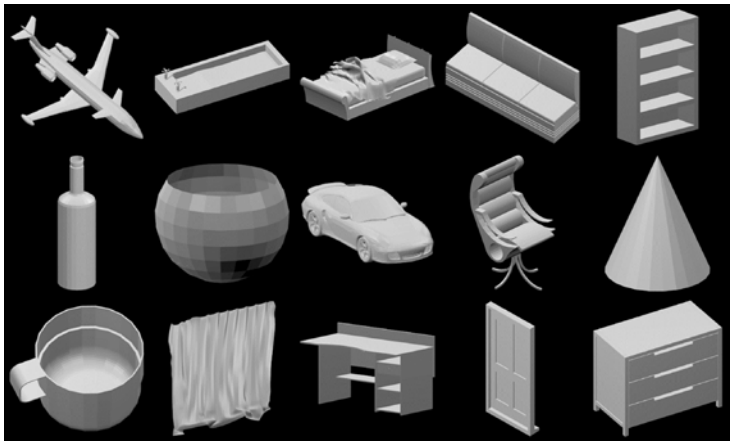

Fig. 3. Sample data from the ModelNet40 dataset.

To verify and validate the effectiveness of the proposed method, we use the ModelNet40 dataset for evaluation. The ModelNet40 dataset is a collection of 3D CAD object data from 40 different class categories. It consists of 9,840 training data and 2,468 test data samples. The use of ModelNet40 facilitates quantitative comparison with the baseline CLF method in performing 3DCMR tasks. We use the same pre-processed dataset for evaluation which is available at *https://github.com/LongLong-Jing/Cross-Modal-Center-Loss*.

4.2 Experimental Details

The experiments were conducted using a Linux Ubuntu server machine equipped with an Intel Xeon E5-2698 2.20 GHz CPU, 256 GB RAM, and four NVIDIA TITAN Xp GPUs.

To verify the performance of the proposed method, training was carried out using various batch sizes, and we applied cosine learning rate decay during the training process. The backbone network used 512-dimensional features. The proposed cross-modal SimSiam method uses a projection network that composes the encoder and a prediction network that extracts feature p from the projected feature $\hat{z}$. Unlike the original SimSiam, networks in the proposed cross-modal SimSiam use a simple multilayer perceptron (MLP) structure with no batch normalization. The MLP consists of a linear layer, a Rectified Linear Unit (ReLU) activation function, and an additional linear layer. The network for label prediction also uses the same MLP structure. We used 1,024 sampling points for point cloud and four multi-view images to extract representative features. The proposed method uses the same networks used in CLF for individual multi-modality feature extraction. Features for images, meshes, and point clouds are extracted using ResNet [38], MeshNet [17], and DGCNN [26], respectively.

In addition, to extract distinctive features, data augmentation was applied for training each modality. In the image modality, random crop and horizontal flip were applied to the 180 randomly captured multi-view images of an object. In the mesh modality, random jittering was applied. Random translation, rotation, jittering, and scaling were applied in the point-cloud modality.

In the experiments, the performance evaluation was based on mean Average Precision (mAP). Average Precision (AP) is a single value that represents the overall performance of the algorithm. In the context of a 3DCMR task, it reflects the accuracy of retrieval results for each class. When a similarity search for a given object is performed, similar objects, i.e., objects

belonging to the same class as the given object, should be placed at the front of the result list. Simply put, the AP quantifies how high in the list of retrieval results are the objects from the same class. As stated in the CLF, given test set T and target class y, if the number of data items belonging to class e is v_y, the AP is computed as shown in (10).

$$AP(y) = \frac{1}{v_y} \sum_{t=1}^{T} Prec(t) \cdot \mathbb{I}_{\{0,1\}}(t) \tag{10}$$

In (10), $Prec$ denotes the precision of the first t-th retrieved data, and $\mathbb{I}$ is an indicator function that outputs a 1 if the current class is the same as the target class y and 0 otherwise. From (10), the mAP is defined as follows:

$$mAP = \frac{1}{Y} \sum_{y=1}^{Y} AP(y) \tag{11}$$

4.3 Quantitative Results

To provide evidence of the effectiveness of the proposed approach, comparative experiments were carried out under various conditions. Initially, the proposed method was tested for learning with different batch sizes, and the results are presented in **Table 1**. CLF shows better performance as the batch size increases. However, the proposed method performs better when the batch size becomes smaller. Given a small batch size, if the relationship between modalities is defined using simple Euclidean distance, there is a limited amount of data that can be compared per learning iteration during training. Therefore, it can be concluded that there is potential for improving performance through context-based learning using projected features, rather than relying on simple comparisons.

Table 1. Comparison of the proposed method with baseline (CLF) using various batch sizes.

Batch Size: 12										
From	Image	Image	Image	Mesh	Mesh	Mesh	Point	Point	Point	Mean
To	Image	Mesh	Point	Image	Mesh	Point	Image	Mesh	Point	mAP
CLF	45.67	13.89	32.32	25.50	06.98	08.29	59.50	27.68	15.87	26.19
ours	51.61	10.57	22.16	10.03	38.98	15.48	21.17	15.04	52.20	**26.36**
Batch Size: 24										
From	Image	Image	Image	Mesh	Mesh	Mesh	Point	Point	Point	Mean
To	Image	Mesh	Point	Image	Mesh	Point	Image	Mesh	Point	mAP
CLF	63.56	73.22	72.08	88.44	68.81	84.60	82.44	67.46	83.56	76.02
ours	86.53	82.39	78.04	78.12	79.34	74.25	63.27	71.9	71.29	**76.13**
Batch Size: 48										
From	Image	Image	Image	Mesh	Mesh	Mesh	Point	Point	Point	Mean
To	Image	Mesh	Point	Image	Mesh	Point	Image	Mesh	Point	mAP
CLF	85.64	86.94	85.59	88.91	86.50	86.67	85.44	84.67	86.62	**86.33**
ours	89.23	88.94	86.18	87.23	87.90	84.93	81.75	83.11	79.40	85.41

To compare the performance of the NCF utilized in the proposed method, experiments were conducted using a fixed batch size. **Table 2** presents the results which show that applying NCF yields better performance. NCF can perform robust learning in noisy environments, resulting in better performance.

Table 2. Comparison between w/ and w/o NCF.

Batch Size: 48										
From	Image	Image	Image	Mesh	Mesh	Mesh	Point	Point	Point	Mean
To	Image	Mesh	Point	Image	Mesh	Point	Image	Mesh	Point	mAP
w/ NCF	89.23	88.94	86.18	87.23	87.90	84.93	81.75	83.11	79.40	**85.41**
w/o NCF	89.60	88.81	84.92	88.00	88.27	83.89	79.89	81.05	77.44	84.65

4.4 Qualitative Results

To verify whether our proposed method performs well in 3DCMR tasks, we provide qualitative experimental results. The similarity search results are shown in **Fig. 4**. As shown in the figure, our proposed method also demonstrates the ability to conduct similarity-based searches smoothly.

4.5 Discussion

In this paper, we propose a method to improve the conventional CLF to extract robust center features. Center features are highly sensitive, displaying variations in performance depending on the batch size, as seen from the CLF results in **Table 1**. Moreover, using simple Euclidean distance makes it difficult to adapt to data changes when using small batches for training. Therefore, there is a need for robust center feature extraction and stronger inter-class discriminative capabilities. As a result, it was confirmed that when training with small batch sizes, learning based on context through the comparison of positive samples at the projection level improves accuracy compared to the conventional method.

5. Conclusion

This paper proposes a new method that can extract robust features under small batch sizes for 3DCMR tasks. The conventional CLF, which serves as the baseline method, uses cross-modal center loss to extract robust features for all modalities with no limitations. The adoption of center loss resulted in inheriting both its benefits and drawbacks. Fluctuations in performance can be observed in CLF depending on the batch size. When the batch size is small, significant performance degradation occurs. The proposed method treats multi-modal data as augmented positive samples and applies the SimSiam loss which is commonly used in self-supervised learning. In addition, noisy center features are applied to extract robust center features. Experimental results using the 3D CAD data in the ModelNet40 dataset show that the proposed method shows performance improvements when using small batch sizes compared to the conventional method. Consequently, the proposed method enables deep learning with 3D data even in constrained computational environments.

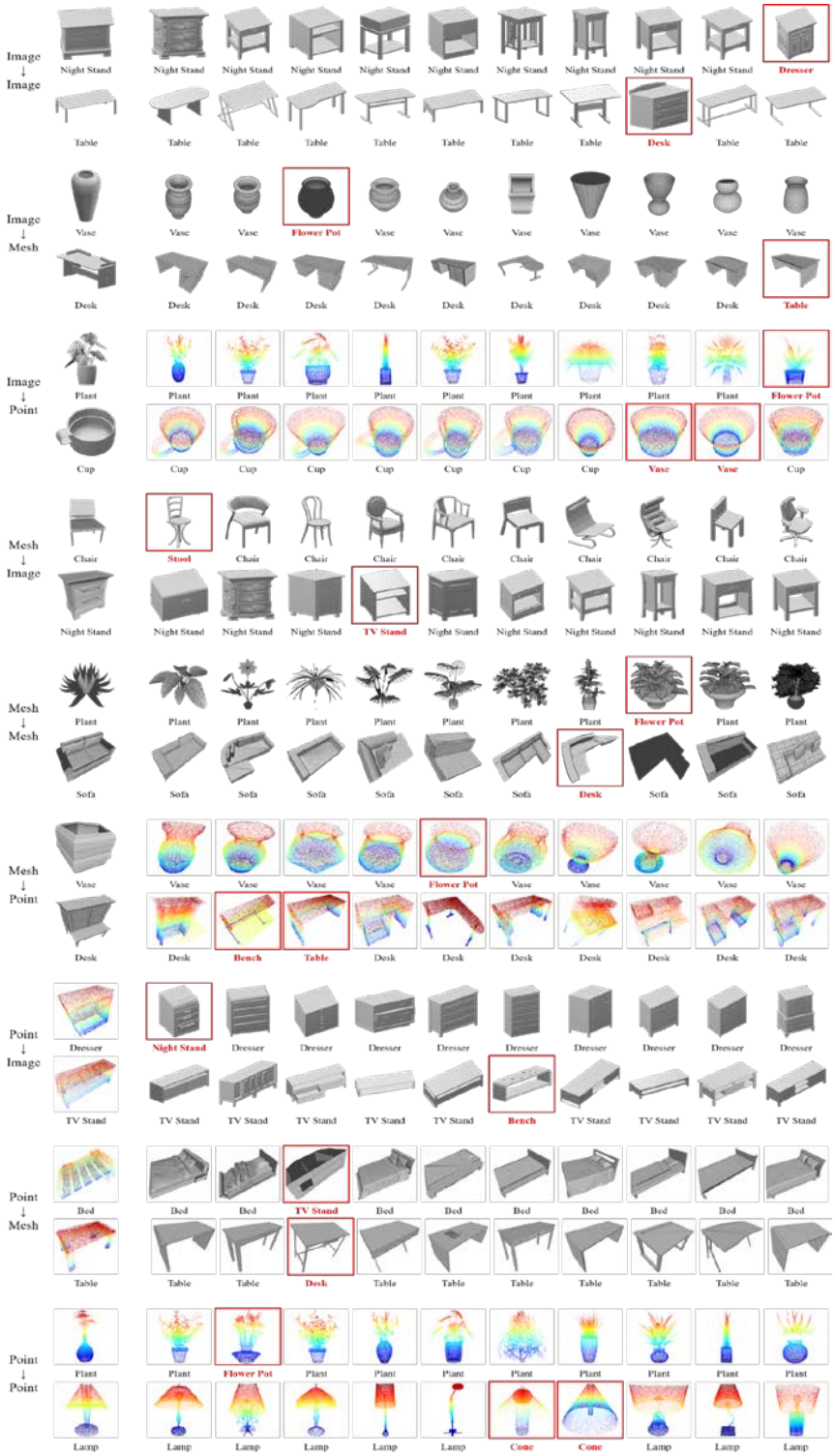

Fig. 4. The qualitative results of 3DCMR using the proposed method on ModelNet40 dataset. The objects in the first column represent the target objects, while the others denote the top 10 similar objects based on similarity criteria.

References

[1] H. Durrant-Whyte, and T. Bailey, "Simultaneous localization and mapping: part I," *IEEE Robotics & Automation Magazine (RAM)*, vol. 13, no. 1, pp. 99-110, June. 2006. Article (CrossRef Link).

[2] T. Bailey, and H. Durrant-Whyte, "Simultaneous localization and mapping (SLAM): part II," *IEEE Robotics & Automation Magazine (RAM)*, vol. 13, no. 3, pp. 108-117, Sept. 2006. Article (CrossRef Link).

[3] Y. Fu, L. Shen, and T. Chen, "3D-Distortion Based Rate Distortion Optimization for Video-Based Point Cloud Compression," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 17, no. 2, pp. 435-449, 2023. Article (CrossRef Link).

[4] Y. Hong, and J. Kim, "3D Mesh Model Exterior Salient Part Segmentation Using Prominent Feature Points and Marching Plane," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 13, no. 3, pp. 1418-1433, 2019. Article (CrossRef Link).

[5] M. Narendra, D. M. L. Valarmathi, D. L. Anbarasi, "Optimization of 3D Triangular Mesh Watermarking Using ACO-Weber's Law," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 14, no. 10, pp. 4042-4059, 2020. Article (CrossRef Link).

[6] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *Proc. of European Conference on Computer Vision (ECCV)*, pp. 405-421, 2020. Article (CrossRef Link).

[7] L. Jing, E. Vahdani, J. Tan, and Y. Tian, "Cross-Modal Center Loss for 3D Cross-Modal Retrieval," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3142-3151, 2021. Article (CrossRef Link).

[8] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A Discriminative Feature Learning Approach for Deep Face Recognition," in *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 499-515, 2016. Article (CrossRef Link).

[9] X. Chen, and K. He, "Exploring Simple Siamese Representation Learning," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15745-15753, 2021. Article (CrossRef Link).

[10] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D ShapeNets: A Deep Representation for Volumetric Shapes," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1912-1920, 2015. Article (CrossRef Link).

[11] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-View Convolutional Neural Networks for 3D Shape Recognition," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 945-953, 2015. Article (CrossRef Link).

[12] X. He, Y. Zhou, Z. Zhou S. Bai, and X. Bai, "Triplet-Center Loss for Multi-View 3D Object Retrieval," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1945-1954, 2018. Article (CrossRef Link).

[13] A. Hamdi, S. Giancola, and B. Ghanem, "MVTN: Multi-View Transformation Network for 3D Shape Recognition," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1-11, 2021. Article (CrossRef Link).

[14] Z. Li, C. Xu, and B. Leng, "Angular Triplet-Center Loss for Multi-View 3D Shape Retrieval," in *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, no. 1, pp. 8682-8689, 2019. Article (CrossRef Link).

[15] X. Wei, R. Yu, and J. Sun, "View-GCN: View-Based Graph Convolutional Network for 3D Shape Analysis," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1847-1856, 2020. Article (CrossRef Link).

[16] W. -Z. Nei, M. -J. Ren, A. -A. Liu, Z. Mao, and J. Nie, "M-GCN: Multi-Branch Graph Convolution Network for 2D Image-based on 3D Model Retrieval," *IEEE Transactions on Multimedia (TMM)*, vol. 23, pp. 1962-1976, 2021. Article (CrossRef Link).

[17] Y. Feng, Y. Feng, H. You, X. Zhao, and Y. Gao, "MeshNet: Mesh Neural Network for 3D Shape Representation," in *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 33, no. 1, pp. 8279-8286, 2019. Article (CrossRef Link).

[18] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, "MeshCNN: A Network with an Edge," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1-12, 2019. Article (CrossRef Link).
[19] F. Milano, A. Loquercio, A. Rosinol, D. Scaramuzza, and L. Carlone, "Primal-Dual Mesh Convolutional Neural Networks," in *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 952-963, 2020. Article (CrossRef Link).
[20] Y. Liang, S. Zhao, B. Yu, J. Zhang, and F. He, "MeshMAE: Masked Autoencoders for 3D Mesh Data Analysis," in *Proc. of the European Conference on Computer Vision (ECCV)*, pp. 37-54, 2022. Article (CrossRef Link).
[21] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77-85, 2017. Article (CrossRef Link).
[22] C. R. Qi, L. Yi, H. su, and L. J. Guibas, "PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space," in *Proc. of Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5105-5114, 2017. Article (CrossRef Link).
[23] H. Thomas, C. R. Qi, J. -E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, "KPConv: Flexible and Deformable Convolution for Point Clouds," in *Proc. of the IEEE International Conference on Computer Vision (ICCV)*, pp. 6410-6419, 2019. Article (CrossRef Link).
[24] Y. Shen, C. Feng, Y. Yang, and D. Tian, "Mining Point Cloud Local Structures by Kernel Correlation and Graph Pooling," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4548-4557, 2018. Article (CrossRef Link).
[25] S. S. Mohammadi, Y. Wang, and A. D. Bue, "Pointview-GCN: 3D Shape Classification with Multi-View Point Clouds," in *Proc. of the IEEE International Conference on Image Processing (ICIP)*, pp. 3103-3107, 2021. Article (CrossRef Link).
[26] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic Graph CNN for Learning on Point Clouds," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 5, pp. 1-12, 2019. Article (CrossRef Link).
[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proc. of the International Conference on Machine Learning (ICML)*, vol. 119, pp. 1597-1607, 2020. Article (CrossRef Link).
[28] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726-9735, 2020. Article (CrossRef Link).
[29] J. -B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning," in *Proc. of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 21271-21284, 2020. Article (CrossRef Link).
[30] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. Rodrigo, "CrossPoint: Self-Supervised Cross-Modal Contrastive Learning for 3D Point Cloud Understanding," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9892-9902, 2022. Article (CrossRef Link).
[31] R. Zhang, L. Wang, Y. Qiao, P. Gao, and H. Li, "Learning 3D Representations from 2D Pre-Trained Models via Image-to-Point Masked Autoencoders," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21769-21780, 2023. Article (CrossRef Link).
[32] Y. Yao, Y. Zhang, Z. Yin, J. Luo, W. Ouyang, and X. Huang, "3D Point Cloud Pre-training with Knowledge Distillation from 2D Images," in *arXiv preprint arXiv:2212.08974*, 2022. Article (CrossRef Link).
[33] A. Chen, K. Zhang, R. Zhang, Z. Wang, Y. Lu, Y. Guo, and S. Zhang, "PiMAE: Point Cloud and Image Interactive Masked Autoencoders for 3D Object Detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5291-5301, 2023. Article (CrossRef Link).

[34] R. Dong, Z. Qi, L. Zhang, J. Zhang, J. Sun, Z. Ge, L. Yi, and K. Ma, "Autoencoders as Cross-Modal Teachers: Can Pretrained 2D Image Transformers Help 3D Representation Learning?," in *Proc. of International Conference on Learning Representations (ICLR)*, 2023.
[35] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep Supervised Cross-Modal Retrieval," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10386-10395, 2019. Article (CrossRef Link).
[36] X. Yuan, Z. Lin, J. Kuen, J. Zhang, Y. Wang, M. Maire, A. Kale, and B. Faieta, "Multimodal Contrastive Training for Visual Representation Learning," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6991-7000, 2021. Article (CrossRef Link).
[37] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised Contrastive Learning," in *Proc. of 34th International Conference on Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 18661-18673, 2020. Article (CrossRef Link) .
[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778, 2016. Article (CrossRef Link).

Yeon-Seung Choo was born in Suwon, Korea, in 1993. He received his B.S. degree in Information and communication Engineering from Sunmoon University, Korea, in 2018, and he received an M.S. degree in Image Science at Chung-Ang University in 2020. Since 2020, he has been with Korea Electronics Technology Institute (KETI). His research interests include deep learning, video registration and 3D Cross-Modal Retrieval.

Boeun Kim received the B.S. degree in Electronic and Electrical Engineering from Korea Advance Institute of Science and Technology. She received the M.S. and Ph.D. degrees in Electrical and Computer Engineering at Seoul National University. From 2015 to 2017, she was a research engineer in Samsung Electronics. Since 2017, she has been with Korea Electronics Technology Institute (KETI). Her research interests include deep learning and human motion pattern analysis.

Hyun-Sik Kim received B.S. and M.S. degrees in information and communication engineering from Inha University, South Korea, in 2002 and 2004, respectively. He received his Ph.D. degree in electrical and electronic engineering from Yonsei University, South Korea, in 2017. He joined Korea Electronics Technology Institute (KETI), South Korea in 2004, and is currently a Chief Researcher at the Contents Convergence Research Center. His research interests are in blockchain, immersive content, content copyright protection, and image processing.

Yong-Suk Park received his B.S. and M.S. degrees in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, in 1997 and 1998, respectively, and his Ph.D. degree in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2018. He is currently a Principal Researcher at Korea Electronics Technology Institute (KETI), Seoul, Korea. Before joining KETI in 2003, he was with I&C Technology and Samsung S1, where he worked on projects relevant to Wi-Fi networks and security system integration. His current research interests are in AI-based content protection in virtual ecosystems.