Supplementary Materials for Multimodal Semi-Supervised Learning for 3D Objects

Zhimin Chen[1]
zhiminc@clemson.edu
Longlong Jing[2]
ljing@gradcenter.cuny.edu
Yang Liang [2]
lyang1@ccny.cuny.edu
YingLi Tian [2]
ytian@ccny.cuny.edu
Bing Li (Corresponding author) [1]
bli4@clemson.edu

[1] Clemson University
Clemson, USA
[2] The City University of New York,
New York, USA

This document contains the supplementary materials for "Multimodal Semi-Supervised Learning for 3D Objects".

1 Comparison with Cross-Modal Center Loss

The Cross-Modal Center [1] (CMC) loss is specifically proposed for cross-modal retrieval based on Center Loss [2] which directly minimizes the intra-class distances of features from multiple modalities in the universal feature space. The CMC loss only minimizes the feature distance of each object to its prototype. Without any constraints about the relation of samples from different classes, the CMC loss may learn features with low inter-class variation. On the contrary, our proposed M2CP loss jointly minimizes the feature distance of each object to its prototype and maximizes the distance to the others. Therefore, M2CP can learn features with both low intra-class variation and high inter-class variation which further produces better classification and retrieval results.

To thoroughly compare the performance of these two losses, we conducted two sets of experiments, one is with M2CP (X_p) loss and the other is with CMC loss (X_c), for the semi-supervised learning on the ModelNet40 dataset. We compared the performance on three different tasks including 3D cross-modal retrieval, 3D in-domain retrieval, and 3D object classification task. As shown in Table 1, the proposed M2CP outperforms CMC loss for all these three tasks with three type percentages of labeled data. For some retrieval tasks such as Image2Point and Image2Mesh, our proposed M2CP outperforms CMC by more than 3% demonstrating its effectiveness.

Amount	2%		5%		10%	
Loss	X_c	X_p	X_c	X_p	X_c	X_p
3D Cross-Modal Retrieval						
Image2Point	76.10	**78.46**	80.98	**83.30**	82.03	**85.93**
Image2Mesh	76.47	**80.05**	82.47	**85.04**	83.57	**87.56**
Point2Image	78.03	**79.28**	80.95	**83.18**	82.69	**85.51**
Point2Mesh	77.65	**78.62**	79.75	**82.58**	81.23	**85.00**
Mesh2Image	77.30	**79.10**	82.03	**84.73**	84.00	**87.10**
Mesh2Point	75.20	**77.42**	80.52	**82.27**	82.22	**85.23**
3D In-Domain Retrieval						
Image	78.84	**81.50**	83.23	**85.71**	84.48	**86.96**
Point	76.08	**78.45**	79.67	**82.05**	83.08	**84.16**
Mesh	77.16	**80.03**	82.43	**84.84**	82.04	**84.29**
3D Object Classification						
Image	80.84	**82.78**	86.23	**88.61**	90.48	**91.61**
Point	78.08	**79.86**	83.67	**85.29**	87.08	**88.49**
Mesh	77.16	**78.81**	85.43	**86.51**	86.04	**88.29**

Table 1: Performance comparison of our proposed M2CP (X_p) loss with CMC (X_c) loss on the ModelNet40 dataset. Our proposed M2CP loss consistently outperforms the CMC loss in all three tasks.

References

[1] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss. *arXiv preprint arXiv:2008.03561*, 2020.

[2] 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.