Fine-grained Prototypical Voting with Heterogeneous Mixup for Semi-supervised 2D-3D Cross-modal Retrieval

Supplementary Material

1. Detailed Introduction of Datasets

In this section, we provide detailed information about the three datasets used for the experiments:

- **3D MNIST** [9] is an extension of the original classic MNIST dataset widely used in 3D computer vision tasks such as 3D shape recognition. It includes a diverse set of corresponding 3D point clouds generated from handwritten digits $0-9$. To ensure fair comparisons, we follow RONO [2] and select a subset of $6,000$ image-point cloud pairs for experiments. Pretrained ResNet-18 and DGCNN models are utilized to extract feature vectors from the images and point clouds, respectively.
- **ModelNet10** [8] is a computer vision dataset used for 3D object recognition and classification tasks. It is a subset of the ModelNet40 dataset and consists of 3D models from 10 different categories sourced from Google 3D Warehouse, Trimble 3D Warehouse, and other public resources. Each model is placed in a standardized 3D space and rendered from various angles to provide diverse viewpoints. Following CLF [4], we randomly select one image from the corresponding 180 images for each 3D model as the ground truth.
- **ModelNet40** [8] is an extended version of the ModelNet10 dataset, consisting of 3D models from 40 different categories. Its data sources are consistent with ModelNet10 as well. Similarly, each model in the ModelNet40 dataset is also placed in a standardized 3D space and rendered from multiple angles. We also follow the practice of randomly selecting one image from the corresponding 180 images for each 3D model as the ground truth.

2. Detailed Introduction of Baselines

We make comparisons with various state-of-the-art methods, including seven text-image retrieval approaches (MRL [3], DSCMR [10], ALGCN [5], DA-I-GCN [6], DA-P-GCN [6], DA-I-GAT [6], DA-P-GAT [6]) and two 2D-3D retrieval approaches (CLF [4] and RONO [2]). The detailed introductions of these baselines are as follows:

- **MRL** [3] is an effective cross-modal retrieval method against noisy labels. It maps various modalities into a shared latent space by robust multimodal learning techniques.
- **DSCMR** [10] is an early attempt to deep cross-modal retrieval. It concurrently minimizes both discrimination loss and modality invariance loss to acquire shared representations across diverse modalities.

Table 1. Comparison with more baselines.

Dataset	3D MNIST (200)		ModelNet10 (200)		ModelNet40 (800)	
Task	I2P	P2I	I2P	P2I	I2P	P2I
FixMatch	25.56	27.71	23.66	22.97	11.45	10.93
RONO+PL	87.99	86.84	79.61	76.98	67.36	63.77
M2CP	89.95	87.63	81.94	81.99	70.05	68.66
Ours	**91.68**	**89.92**	**83.97**	**83.22**	**71.32**	**70.24**

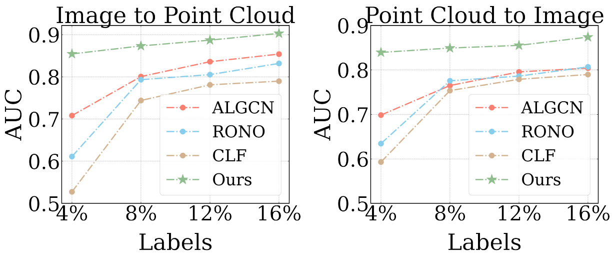

Figure 1. The AUC results on 3D MNIST.

- **ALGCN** [5] uncovers the semantics of labels and conserves semantic correlations across different modalities through the joint training of two different branches.
- **DA-I-GCN** [6] is based on Graph Neural Networks (GNNs) and Generative Adversarial Networks (GANs) with contrastive learning on multi-labels. It leverages Iterative GNN and employs Graph Convolutional Network (GCN) as GNN's backbone.
- **DA-P-GCN** [6] is adapted from DA-I-GCN. It leverages Probabilistic GNN and employs GCN as GNN's backbone.
- **DA-I-GAT** [6] is adapted from DA-I-GCN. It leverages Iterative GNN and employs Graph Attention Network (GAT) as GNN's backbone.
- **DA-P-GAT** [6] is adapted from DA-I-GCN. It leverages Probabilistic GNN and employs GAT as GNN's backbone.
- **CLF** [4] is a strong 2D-3D cross-modal retrieval method that obtains both modality-invariant and discriminative representations through a cross-modal center loss.
- **RONO** [2] tackles the problem of 2D-3D retrieval under label noise by proposing a consistency loss and a robust center learning strategy.

3. Further Qualitative Comparisons

As depicted in Figure 2 and Figure 3, we make further qualitative comparisons by plotting the Precision and Recall curves with respect to Top N returned samples with various

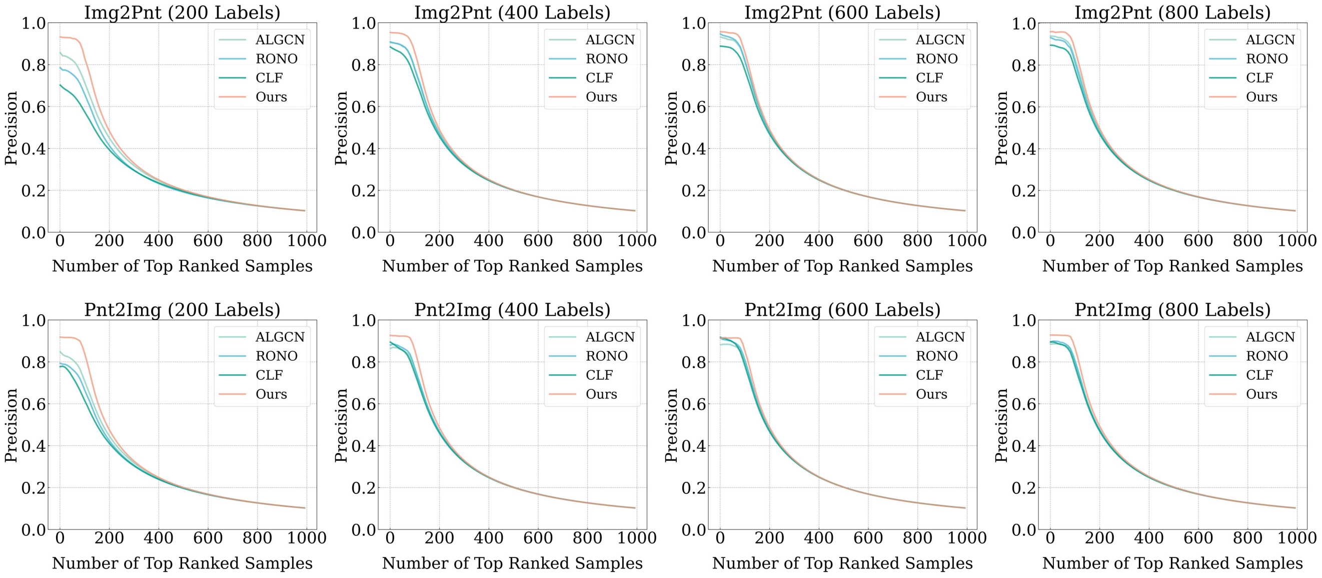

Figure 2. The Precision-Top N curve with various amounts of labels on the 3D MNIST dataset. 2D-to-3D results are plotted in the first row, and 3D-to-2D results are plotted in the second row.

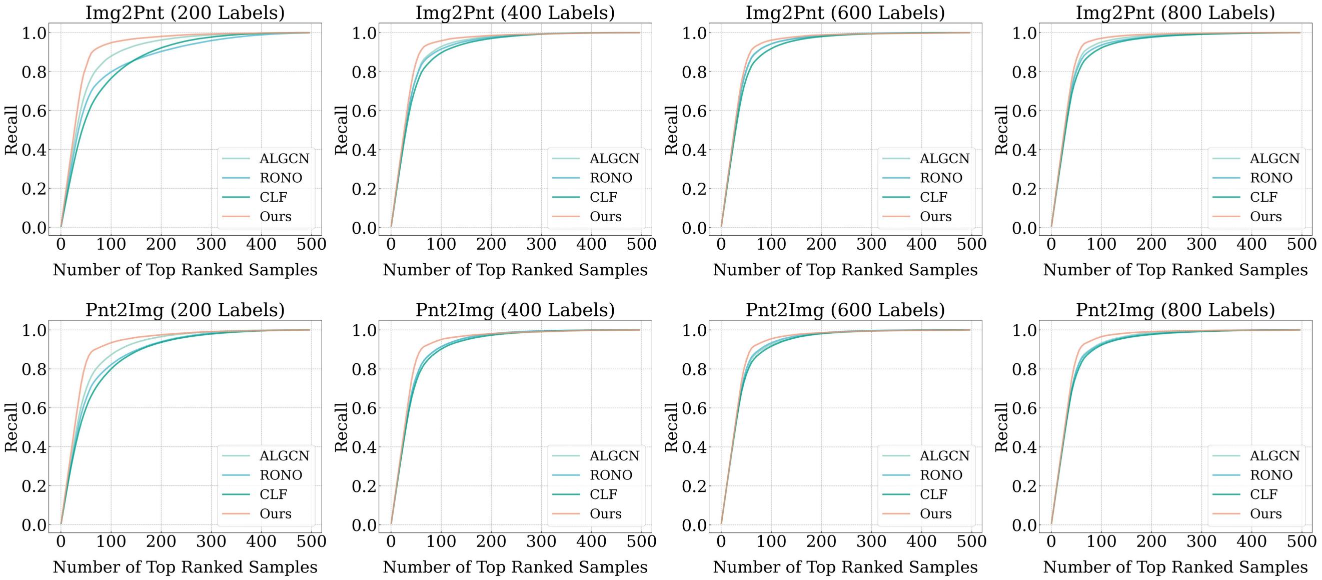

Figure 3. The Recall-Top N curve with various amounts of labels on the 3D MNIST dataset. 2D-to-3D results are plotted in the first row, and 3D-to-2D results are plotted in the second row.

amounts of labels.

From these curves, it is evident that the precision value decreases as the number of returned samples increases, while the recall value exhibits an increase with a rise in the number of returned samples. These two metrics are contradictory to each other. What remains consistent is that under both metrics, one method demonstrates superior performance compared to others, as indicated by its curve being above the other curves. It can be seen that in scenarios with multiple label quantities, our curve consistently remains at the top. This represents the outstanding performance and robustness of our FIVE, indicating our successful endeavor in leveraging labeled and unlabeled data for 2D-3D retrieval.

4. More Experimental Results

4.1. Comparison with More Baselines

We add three baselines, i.e., FixMatch [7], M2CP [1], and RONO+PL for comparisons. RONO+PL additionally utilizes pseudo-labeling to annotate unlabeled data. The compared results in Table 1 show our method continues to outperform other baselines. FixMatch performs poorly due to a missing similarity learning module for cross-modal discrepancy reduction.

4.2. Comparison on AUC

Besides the aforementioned curves, we also include AUC results for clarity. As depicted in Figure 1, our method is significantly superior to the existing baselines.

References

[1] Zhimin Chen, Longlong Jing, Yang Liang, YingLi Tian, and Bing Li. Multimodal semi-supervised learning for 3d objects. *arXiv preprint arXiv:2110.11601*, 2021. 3

[2] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, and Peng Hu. Rono: Robust discriminative learning with noisy labels for 2d-3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11610–11619, 2023. 1

[3] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5403–5413, 2021. 1

[4] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. Cross-modal center loss for 3d cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3142–3151, 2021. 1

[5] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Adaptive label-aware graph convolutional networks for cross-modal retrieval. *IEEE Transactions on Multimedia*, 24: 3520–3532, 2021. 1

[6] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Integrating multi-label contrastive learning with dual adversarial graph neural networks for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4794–4811, 2022. 1

[7] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proceedings of the Conference on Neural Information Processing Systems*, 2020. 3

[8] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 1912–1920. IEEE Computer Society, 2015. 1

[9] Xiaofan Xu, Alireza Dehghani, David Corrigan, Sam Caulfield, and David Moloney. Convolutional neural network for 3d object recognition using volumetric representation. In *International Workshop on Sensing, Processing and Learning for Intelligent Machines*, pages 1–5. IEEE, 2016. 1

[10] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019. 1