

Supplementary Materials

DAC: 2D-3D Retrieval with Noisy Labels via Divide-and-Conquer Alignment and Correction

Anonymous Authors

In this supplementary material, we provide complementary information on experiments and Objaverse-N200. Specifically, we give extra implementation details of our DAC in Appendix A. In Appendix B, our supplementary experimental results are shown and we conduct several insightful experiments to comprehensively investigate the impact of our multimodal loss distribution and self-correction. Finally, we give supplementary dataset information of Objaverse-N200 in Appendix C.

A IMPLEMENTATION DETAILS

Algorithm 1 The proposed DAC framework

Input: Training dataset $\{\mathcal{X}, \mathcal{Y}\}$, feature encoders $\{\phi_j\}_j^M$, shared classifier F_u , multimodal classifier F_m .

- 1: **while** $e < \text{MaxEpoch}$ **do**
- 2: Calculate multimodal loss distribution $l = \{l_i\}_{i=1}^N$
- 3: $\gamma_i = \text{GMM}(l_i | \{l_i\}_{i=1}^N)$ ▷ to obtain the credibility
- 4: $S_c = \{(x_i, y_i)_c | \gamma_i > \alpha, \forall (x_i, y_i) \in \{\mathcal{X}, \mathcal{Y}\}\}$ ▷ clean set
- 5: $S_n = \{(x_i, y_i)_n | \gamma_i \leq \alpha, \forall (x_i, y_i) \in \{\mathcal{X}, \mathcal{Y}\}\}$ ▷ noisy set
- 6: **for** $i = 1$ to num_iters **do**
- 7: $(x_i, y_i)_c \xleftarrow{\text{align}} \mathcal{L}_{sem}(\phi_0; \dots; \phi_M) \ \& \ \mathcal{L}_{inst}(\phi_0; \dots; \phi_M)$.
▷ apply semantic and instance alignment to S_c
- 8: $(x_i, \hat{y}_i)_n \xleftarrow{\text{refurbish}} F_m(\hat{z}_i)$
▷ refurbish label of sample in S_n with fused feature \hat{z}_i
- 9: $(x_i, \hat{y}_i)_n \xleftarrow{\text{align}} \mathcal{L}_{sem}(\phi_0; \dots; \phi_M) \ \& \ \mathcal{L}_{inst}(\phi_0; \dots; \phi_M)$.
▷ apply semantic and instance alignment to S_n
- 10: $(x_i, y_i)_c \ \& \ (x_i, \hat{y}_i)_n \xleftarrow{\text{align}} \mathcal{L}_{cls}(F_u; F_m)$.
▷ align samples in label space
- 11: $\mathcal{L} = \mathcal{L}_{sem} + \lambda_1 \mathcal{L}_{inst} + \lambda_2 \mathcal{L}_{cls}$ ▷ total loss
- 12: Update parameters of $\{\phi_j\}_j^M, F_u, F_m$. in backward process.
- 13: **end for**
- 14: **end while**

Output: The final feature encoders $\{\phi_j\}_j^M$.

In this work, we adopt two fully connected layers as the shared classifier F_u and multimodal classifier F_m for common representation classification. For all the datasets, the temperature parameter τ_c is set as 0.22, and τ_m is set as 1.0. Consistent with [4], we utilize the mean Average Precision (mAP) score as our primary metric to evaluate the performance of the models. It is worth noting that for the balance parameters λ_1 and λ_2 , we set them as different values for the clean set S_c and noisy set S_n . Specifically, for S_c , λ_1 and λ_2 are set as 0.1 and 0.1, respectively. For noisy set S_n , we set λ_1 and λ_2 as 10 and 1 to mitigate the negative impact of false-corrected labels. The optimization process of our framework is shown in Algorithm 1.

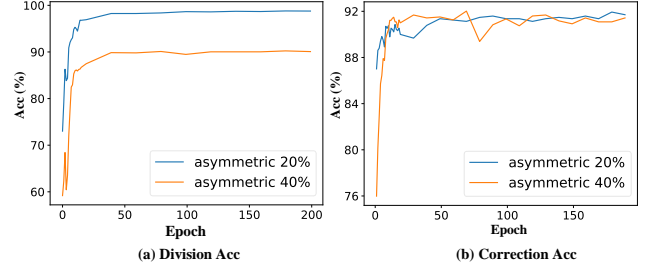


Figure 1: Division accuracy of MDD and correction accuracy of the corrected labels generated by self-correction strategy on ModelNet10 under asymmetric noise.

To evaluate the effectiveness of our DAC in real-world scenarios, we conduct experiments on our proposed realistic noisy dataset Objaverse-N200, where we set Objaverse-N200, which comprised 194,800 objects as the training set, and Objaverse-lvis with 46,205 objects [3] as the testing set.

B MORE RESULTS AND ANALYSIS

Due to space limitations in the main paper, we present supplementary experiments in this section.

B.1 More Comparative Experimental Results and Analysis

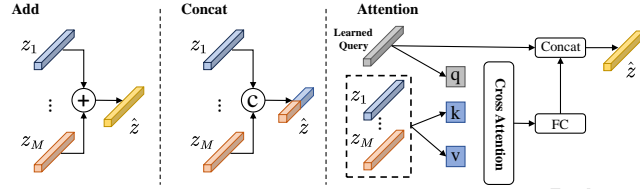
To comprehensively evaluate our model, we conducted three comparative experiments. 1) To fully verify the effectiveness of our model, we conduct comparative experiments on 3D MNIST [17] under symmetric and asymmetric noise as shown in Tab. 1 and Tab. 2. From the results, one can observe that our model achieves superior results compared to previous methods under various symmetric and asymmetric noise, which demonstrates the effectiveness of our method. 2) To thoroughly assess the robustness of our proposed MDD and the self-correction mechanism’s label correction capabilities under asymmetric noise conditions, we also conduct experiments on ModelNet10 [16] under 0.2 and 0.4 asymmetric label noise. The results, depicted in Fig. 1, demonstrate the robustness of our MDD strategy in effectively dividing noisy samples (Fig. 1 (a)) and the high quality of the corrected labels generated by self-correction (Fig. 1 (b)), under challenging asymmetric noise scenarios. 3) In addition, we conduct extensive retrieval experiments across three modalities (image, point cloud, and mesh) on ModelNet10 [16] under 0.2, 0.4, 0.6, and 0.8 symmetric noise, comparing it with the state-of-the-art RONO [4]. The results, presented in Tab. 3 demonstrate the robustness of our model DAC.

B.2 Investigation of Fusion Layer

To assess the impact of distinct fusion layers, we implemented three distinct fusion architectures: Add (feature adding), Concat

Table 1: Performance comparison in terms of mAP under the symmetric noise rates of 0.2, 0.4, 0.6, and 0.8 on the 3D MNIST datasets.

Method	3D MNIST [17]							
	Image \rightarrow Point Cloud				Point Cloud \rightarrow Image			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
CCA [7]	0.415	0.415	0.415	0.415	0.414	0.414	0.414	0.414
DCCA [1]	0.595	0.595	0.595	0.595	0.593	0.593	0.593	0.593
DCCAE [15]	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600
DGCPN [18]	0.792	0.792	0.792	0.792	0.783	0.783	0.783	0.783
UCCH [9]	0.791	0.791	0.791	0.791	0.790	0.790	0.790	0.790
GMA [14]	0.449	0.438	0.426	0.415	0.437	0.432	0.423	0.414
MvDA [11]	0.481	0.461	0.432	0.328	0.482	0.461	0.431	0.323
AGAH [6]	0.688	0.557	0.128	0.108	0.680	0.548	0.122	0.116
DADH [2]	0.735	0.632	0.403	0.290	0.727	0.614	0.382	0.286
DAGNN [13]	0.883	0.850	0.749	0.445	0.879	0.845	0.743	0.435
ALGCN [12]	0.874	0.840	0.757	0.401	0.868	0.831	0.748	0.385
DSCMR [19]	0.908	0.812	0.512	0.219	0.896	0.811	0.472	0.140
MRL [8]	0.955	0.937	0.918	0.785	0.944	0.931	0.905	0.791
CLF [10]	0.890	0.811	0.460	0.124	0.872	0.793	0.426	0.120
CLF +MAE [5]	0.810	0.812	0.501	0.122	0.809	0.811	0.483	0.122
RONO [4]	0.962	0.952	0.931	0.831	0.948	0.934	0.915	0.828
DAC (Ours)	0.965	0.964	0.954	0.846	0.950	0.940	0.939	0.838

**Figure 2: Structure comparison of different fusion layer ψ .**

(feature concatenation), and an Attention-based approach (leveraging a learned query for feature fusion through cross-attention) as shown in Fig. 2. The comparative experimental results of these multimodal fusion techniques are presented in Tab. 4. In the table, Ensemble denotes to naive ensemble of the unimodal loss distribution, which is set as the baseline. The findings reveal that a naive combination of unimodal loss distribution yields suboptimal performance as the significance of various modalities varies. Conversely, multimodal feature fusion enables the network to adaptively capture discriminative semantics from diverse modalities, which significantly boosts the discrimination of representation. Among these strategies, feature concatenation emerges as the most effective approach, as demonstrated in Tab. 4. Hence, we utilize Concat (feature concatenation) as our final fusion layer ψ .

B.3 Insightful Experiments of Multimodal Loss Distribution

To thoroughly investigate the impact of our multimodal loss distribution, we conduct several insightful visualization experiments on trimodal ModelNet40 [16] under both symmetric and asymmetric noise. The results are shown in Fig. 3 and Fig. 4. From the results, we can find that multimodal loss distribution exhibits a smaller overlap and larger loss value for False labeled samples compared to the unimodal loss distribution (Image, Point cloud, Mesh). Notably, for

Table 2: Performance comparison in terms of mAP under the asymmetric noise rates of 0.1, 0.2, and 0.4 on the 3D MNIST datasets.

Method	3D MNIST [17]							
	Image \rightarrow Point Cloud				Point Cloud \rightarrow Image			
	0	0.1	0.2	0.4	0	0.1	0.2	0.4
CCA [7]	0.415	0.415	0.415	0.415	0.415	0.415	0.415	0.415
DCCA [1]	0.595	0.595	0.595	0.595	0.593	0.593	0.593	0.593
DCCAE [15]	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600
DGCPN [18]	0.792	0.792	0.792	0.792	0.783	0.783	0.783	0.783
UCCH [9]	0.791	0.791	0.791	0.791	0.790	0.790	0.790	0.790
GMA [14]	0.514	0.444	0.436	0.415	0.500	0.435	0.417	0.396
MvDA [11]	0.530	0.472	0.407	0.370	0.508	0.472	0.397	0.352
AGAH [6]	0.967	0.730	0.611	0.519	0.961	0.729	0.589	0.512
DADH [2]	0.971	0.848	0.718	0.570	0.969	0.825	0.701	0.572
DAGNN [13]	0.927	0.894	0.871	0.684	0.927	0.893	0.864	0.691
ALGCN [12]	0.908	0.876	0.860	0.635	0.900	0.871	0.852	0.641
DSCMR [19]	0.963	0.914	0.869	0.711	0.945	0.906	0.862	0.704
MRL [8]	0.963	0.959	0.944	0.792	0.945	0.940	0.922	0.762
CLF [10]	0.983	0.945	0.924	0.809	0.958	0.932	0.920	0.802
CLF [19]+MAE[5]	0.971	0.942	0.921	0.796	0.951	0.930	0.918	0.783
RONO [4]	0.983	0.961	0.958	0.912	0.968	0.947	0.938	0.897
DAC(Ours)	0.974	0.965	0.965	0.924	0.957	0.949	0.948	0.899

Table 3: Performance comparison of RONO [4] and our DAC under the symmetric noise rates of 0, 0.2, 0.4, 0.6, and 0.8 on trimodal (Image, Mesh, Point cloud) ModelNet10 dataset.

η	Qry	Img			Msh			Pnt		
	Retrv	Img	Msh	Pnt	Img	Msh	Pnt	Img	Msh	Pnt
0	RONO	0.913	0.906	0.898	0.896	0.919	0.904	0.895	0.903	0.892
	Ours	0.925	0.915	0.909	0.915	0.915	0.905	0.916	0.899	0.905
0.2	RONO	0.871	0.889	0.877	0.890	0.912	0.905	0.872	0.899	0.895
	Ours	0.919	0.912	0.906	0.913	0.914	0.903	0.897	0.899	0.897
0.4	RONO	0.866	0.888	0.878	0.883	0.911	0.900	0.865	0.897	0.895
	Ours	0.920	0.914	0.907	0.915	0.914	0.903	0.900	0.895	0.892
0.6	RONO	0.840	0.857	0.850	0.868	0.901	0.892	0.854	0.888	0.892
	Ours	0.900	0.899	0.891	0.900	0.904	0.894	0.889	0.890	0.891
0.8	RONO	0.826	0.859	0.849	0.858	0.898	0.887	0.842	0.880	0.885
	Ours	0.879	0.872	0.878	0.875	0.875	0.878	0.879	0.877	0.889

asymmetric noise, the extremely complex class conditional noise will degrade the performance of the memorization effect of DNNs, resulting in a large overlap for unimodal loss distribution. In contrast, our multimodal approach effectively captures discriminative semantics across modalities and leverages the complementary information from different modalities to model the multimodal loss distribution. As a result, our multimodal loss distribution exhibits a reduced overlap and a more distinct bimodal pattern, thereby facilitating efficient sample division. These results demonstrate the effectiveness of our multimodal loss distribution in managing both symmetric and asymmetric noise.

B.4 Insightful Experiments of Self-Correction.

To investigate the impact of our self-correction, we conduct visualization experiments to investigate the quality of representations of samples on ModelNet40 [16] under 0.4 symmetric noise, as shown in Fig. 5. The results reveal the following key insights:

Table 4: Investigation on the structure of fusion layer ψ .

Method	ModelNet40			
	Img \rightarrow Pnt		Pnt \rightarrow Img	
	0.4	0.8	0.4	0.8
Ensemble	0.877	0.832	0.866	0.827
Add	0.878	0.830	0.868	0.826
Concat	0.885	0.849	0.882	0.847
Attention	0.880	0.830	0.873	0.824

1) Improved representation compactness. By comparing Fig. 5 (a) and (b), we observed that our model with self-correction generated more compact representations for clean samples as the use of corrected labels from noisy samples facilitated enhanced network optimization. 2) Enhanced discrimination of noisy samples. The comparison between Fig. 5 (c) and (d) demonstrated that without self-correction, the model struggled to mine the true semantics of noisy samples, resulting in a less focused similarity distribution. In contrast, with self-correction, we can exploit the noise-free semantic information from the model's multimodal predictions, which allowed noisy samples to compact to their respective clean centers, thus boosting the semantic compactness and discrimination of their representations. These results highlight the high quality of the corrected labels and the effectiveness of our self-correction strategy.

C OBJAVERSE-N200

In this section, we present a comprehensive overview of our proposed realistic noisy 3D benchmark: Objaverse-N200. Fig. 6 illustrates the noisy samples within Objaverse-N200. From the figure, we could obtain the following observations: 1) **Diversity**. Derived from the extensive Objaverse [3], Objaverse-N200 reveals a significantly enhanced diversity in instances within a category. 2) **Uneven noise distribution**. A comparison between categories like Apple and Basketball reveals a stark contrast in noise ratios, with Basketball having approximately 90% noisy samples, while Apple has around 20%. This highlights the varying noise levels across different categories, posing a realistic challenge. 3) **Feature-dependent noise**. By comparing samples with true and false labels, it is evident that the noisy samples exhibit similar appearances. Furthermore, the zero-shot classification mechanism employed for label assignment introduces feature-dependent noise, making the dataset more challenging. These observations collectively demonstrate the complexity and practical relevance of Objaverse-N200, serving as a valuable resource for evaluating the robustness of 2D-3D retrieval methods in the presence of real-world noise.

REFERENCES

- [1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International conference on machine learning*. PMLR, 1247–1255.
- [2] Cong Bai, Chao Zeng, Qing Ma, Jinglin Zhang, and Shengyong Chen. 2020. Deep adversarial discrete hashing for cross-modal retrieval. In *Proceedings of the 2020 international conference on multimedia retrieval*. 525–531.
- [3] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13142–13153.
- [4] Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, and Peng Hu. 2023. 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*. 11610–11619.
- [5] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. 2017. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 31.
- [6] Wen Gu, Xiaoyan Gu, Jingzi Gu, Bo Li, Zhi Xiong, and Weiping Wang. 2019. Adversary guided asymmetric hashing for cross-modal retrieval. In *Proceedings of the 2019 international conference on multimedia retrieval*. 159–167.
- [7] Harold Hotelling. 1992. Relations between two sets of variates. In *Breakthroughs in statistics: methodology and distribution*. Springer, 162–190.
- [8] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. 2021. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5403–5413.
- [9] Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. 2022. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3877–3889.
- [10] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. 2021. Cross-Modal Center Loss for 3D Cross-Modal Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3141–3150.
- [11] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. 2015. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence* 38, 1 (2015), 188–194.
- [12] Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. 2021. Adaptive label-aware graph convolutional networks for cross-modal retrieval. *IEEE Transactions on Multimedia* 24 (2021), 3520–3532.
- [13] Shengsheng Qian, Dizhan Xue, Huaiwen Zhang, Quan Fang, and Changsheng Xu. 2021. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2440–2448.
- [14] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. 2012. Generalized multiview analysis: A discriminative latent space. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2160–2167.
- [15] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. 2015. On deep multi-view representation learning. In *International conference on machine learning*. PMLR, 1083–1092.
- [16] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1912–1920.
- [17] Xiaofan Xu, Alireza Dehghani, David Corrigan, Sam Caulfield, and David Moloney. 2016. Convolutional neural network for 3d object recognition using volumetric representation. In *2016 first international workshop on sensing, processing and learning for intelligent machines (SPLINE)*. IEEE, 1–5.
- [18] Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. 2021. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 4626–4634.
- [19] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10394–10403.

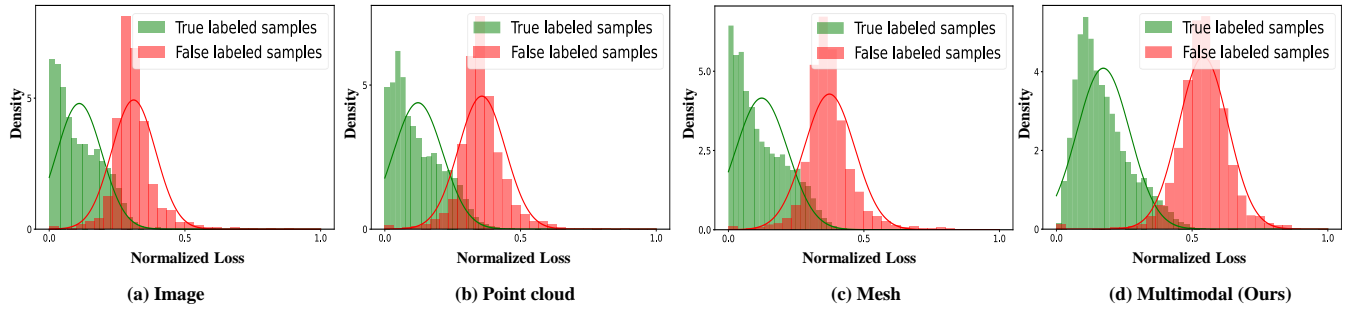


Figure 3: Comparison of the different loss distribution on trimodal ModelNet40 under 0.4 symmetric noise at epoch 2.

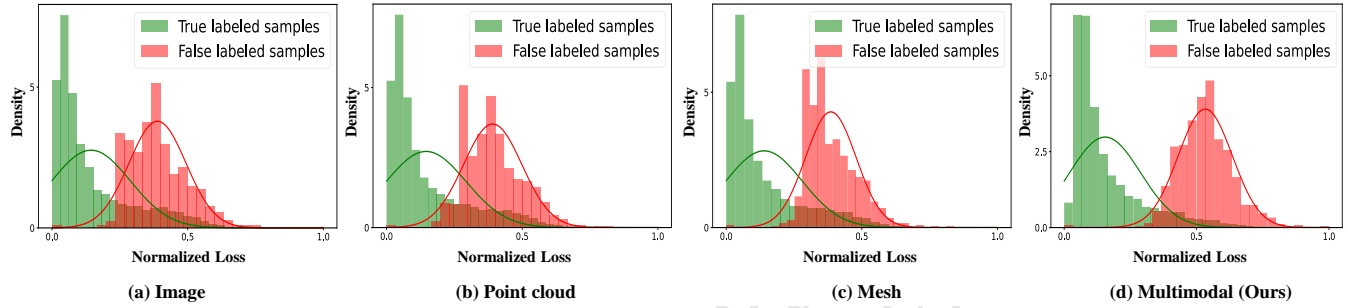


Figure 4: Comparison of the different loss distribution on trimodal ModelNet40 under 0.2 asymmetric noise at epoch 4.

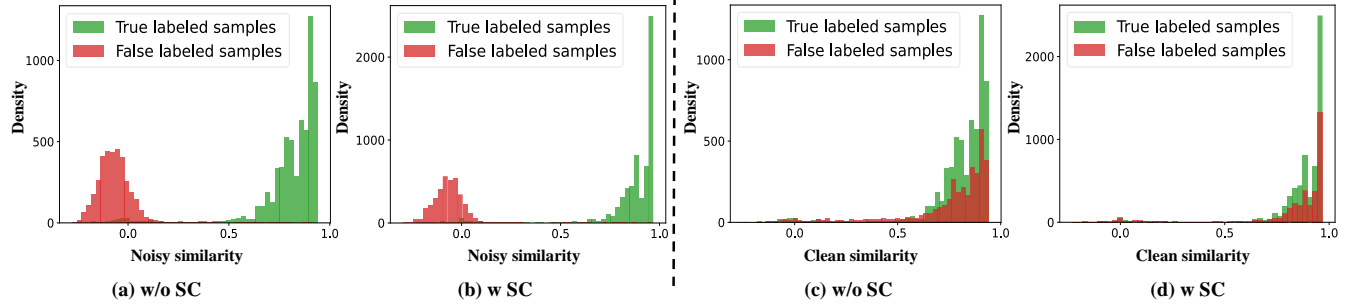


Figure 5: Investigation of Self-Correction (SC) on ModelNet40 under 0.4 symmetric noise. (a), (b), (c), and (d) show the similarity distribution of the training dataset, which describes the cosine similarity between the centers in our contrastive center loss \mathcal{L}_{sem} and common representations of samples. Noisy similarity denotes the similarity between common representations and the centers corresponding to their noisy labels. Clean similarity denotes the similarity between common representations and the centers corresponding to their ground-truth clean labels.

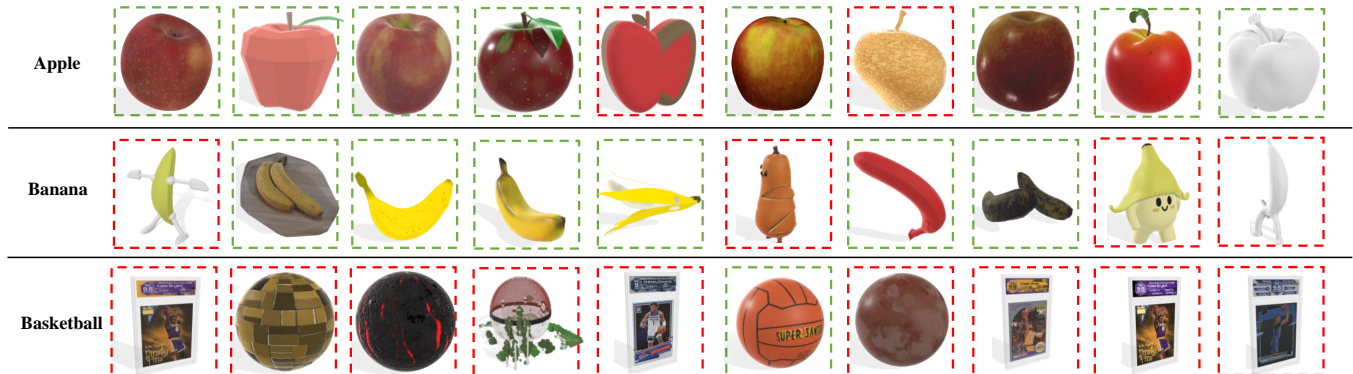


Figure 6: Samples in Objaverse-N200. Green/red boxes indicate True/False labels