JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Paper Tag: WED-AM-322

RONO: Robust Discriminative Learning with Noisy Labels for 2D-3D Cross-Modal Retrieval

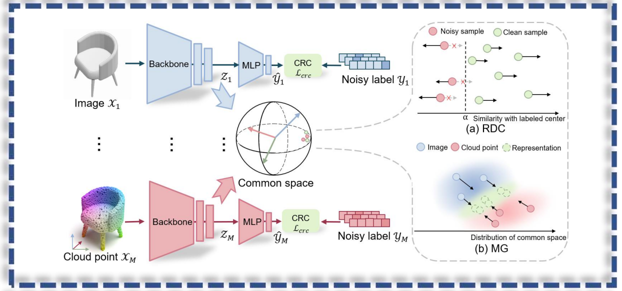
Yanglin Feng[1] Hongyuan Zhu[2] Dezhong Peng[1,3,4] Xi Peng[1] Peng Hu[1,*]

[1]College of Computer Science, Sichuan University
[2]Institute for Infocomm Research (I2R), A*STAR, Singapore
[3]Sichuan Zhiqian Technology Co., Ltd
[4]Chengdu Ruibei Yingte Information Technology Co., Ltd

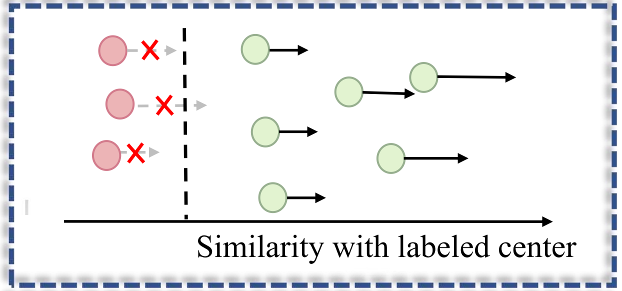
Agency for Science, Technology and Research
SINGAPORE

Overview of Our Work

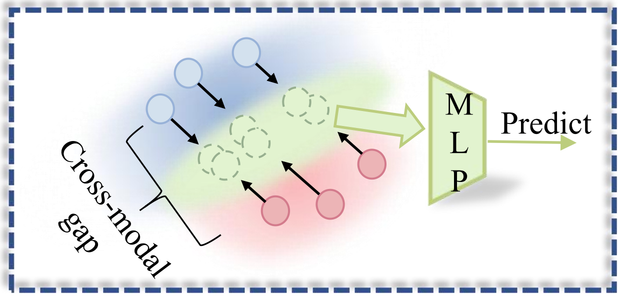
❖ We propose a robust 2D-3D cross-modal retrieval framework (RONO) to robustly learn the common discriminative and modality-invariant representations from noisy labels.



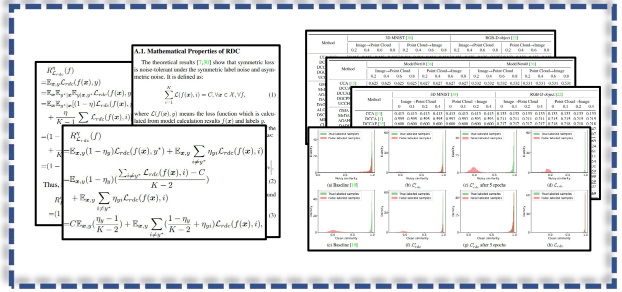
❖ To mitigate the impact of noisy labels, a novel Robust Discriminative Center Learning mechanism (RDCL) is proposed.



❖ To construct discriminative and modality-invariant representations, a Shared Space Consistency Learning mechanism (SSCL) is proposed.

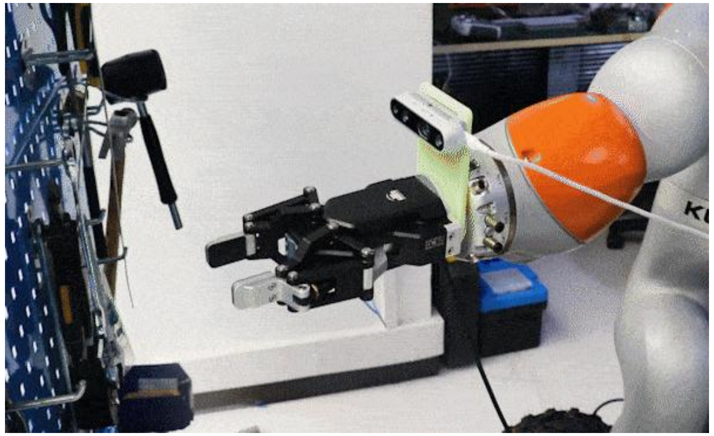


❖ We theoretically and experimentally demonstrate the robustness of our RONO under both synthetic symmetric/asymmetric and real-world noisy labels.



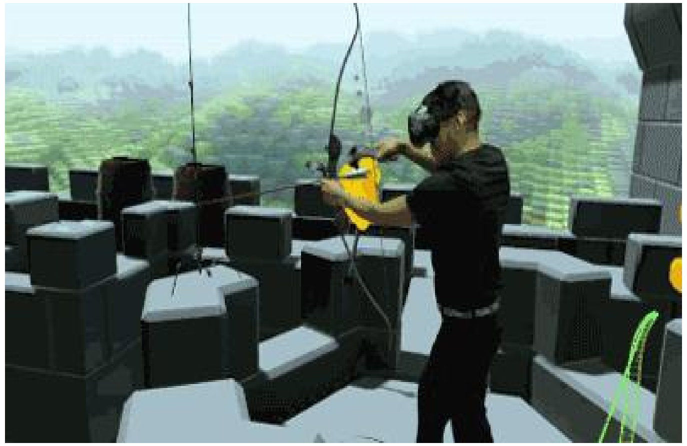
Background

- ❖ Point-cloud retrieval (PCR) is fundamental and crucial for processing and analyzing 3D data, which could provide the direct technical support of the 3D data search engine.



Robotics

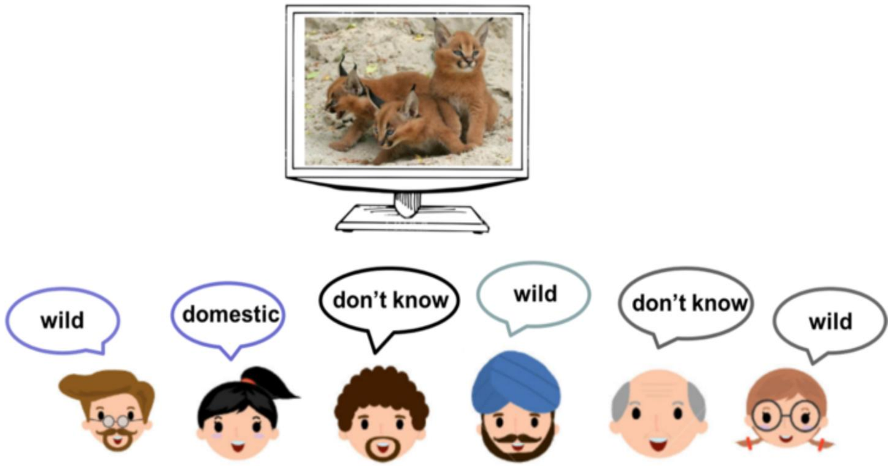
Autonomous driving

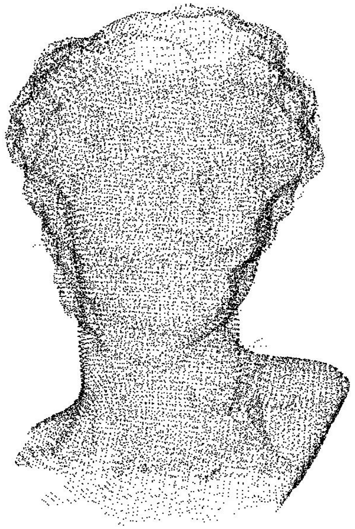


Virtual/augmented reality

- ❖ PCR is often accompanied by retrieving across diverse modalities, termed 2D-3D cross-modal retrieval.

Background





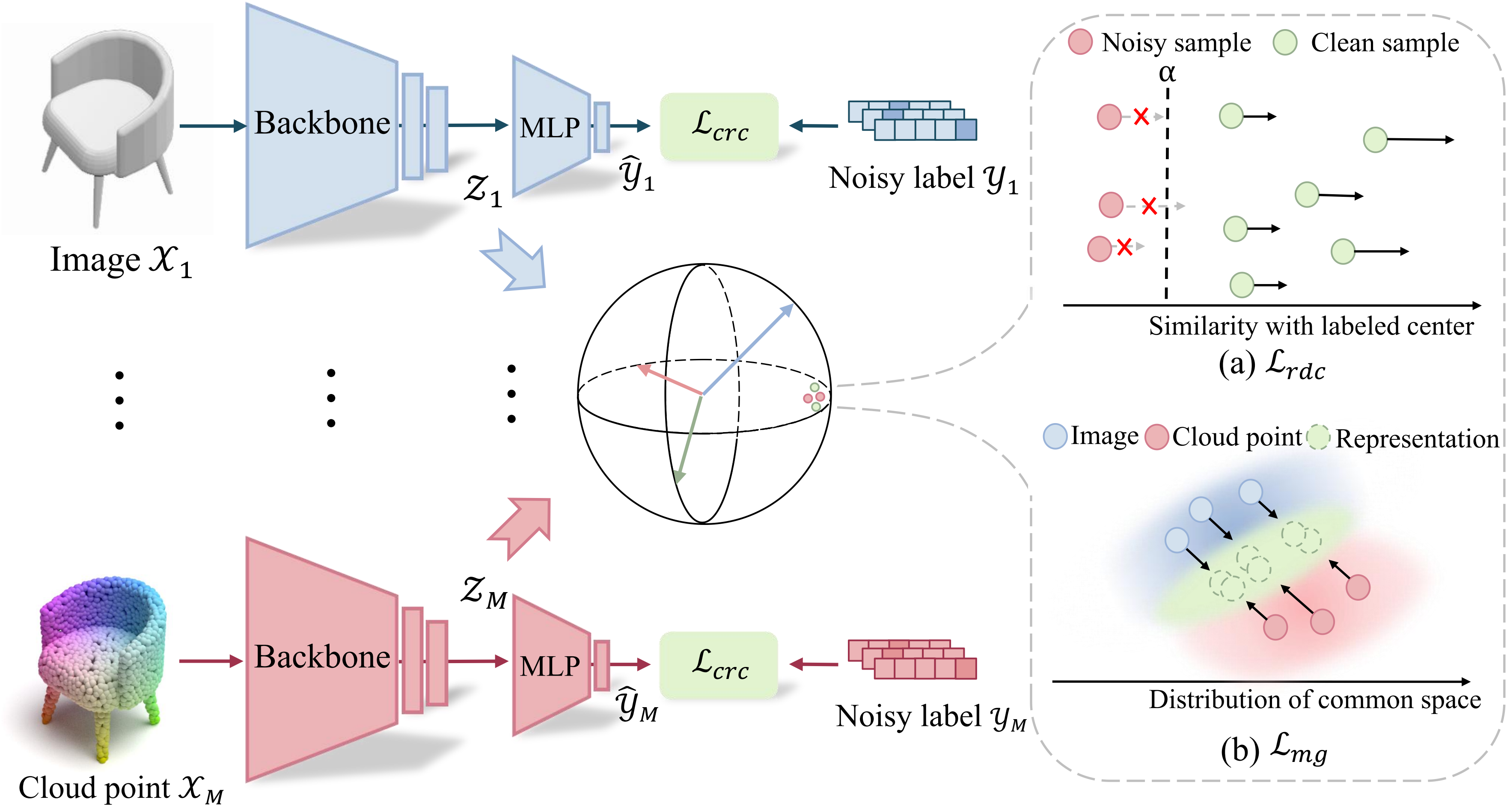
❖ Labeling is time-consuming and labor-intensive.

❖ 3D point cloud is sophisticated.

❖ Disorientation of 3D objects with multiple views.

RONO

Framework:



Robust Discriminative Center Learning (RDCL)

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

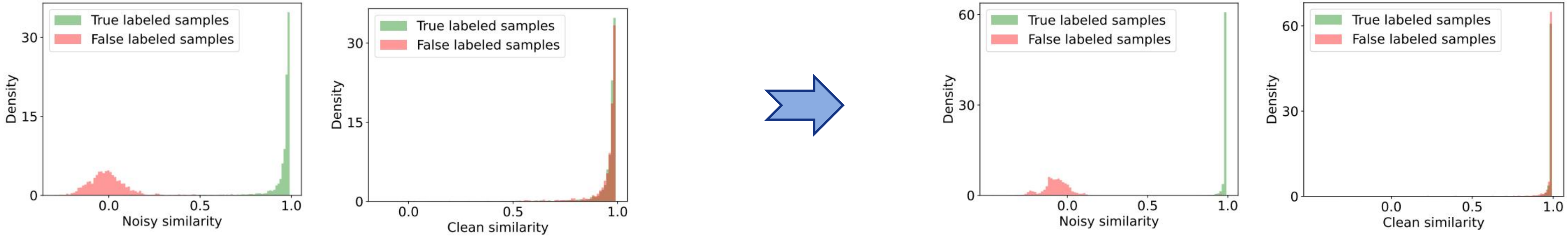
❖ Vanilla Robust Discriminative Center Loss $\mathcal{L}'_{rdc}$

Enforcing the samples with the same category compact to the shared clustering centers, while escaping from other centering centers.

The density vs. the similarity between common representations and noisy centers / clean centers

❖ Robust Discriminative Center Loss $\mathcal{L}_{rdc}$

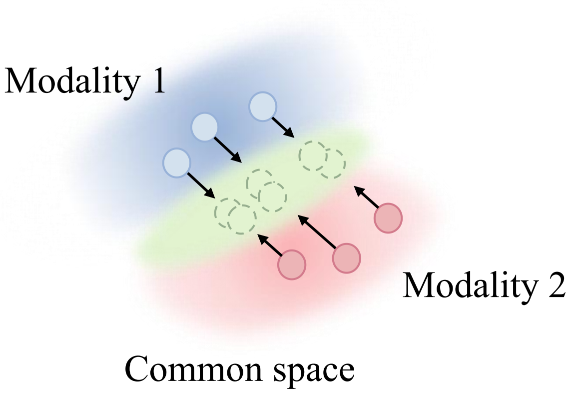
Exploiting the memory effect of neural networks, we enforce the clean samples with the same category compact to the centers, while escaping from other centering centers and enforce the noisy samples escaping from noisy centers.



The density vs. the similarity between common representations and noisy centers / clean centers

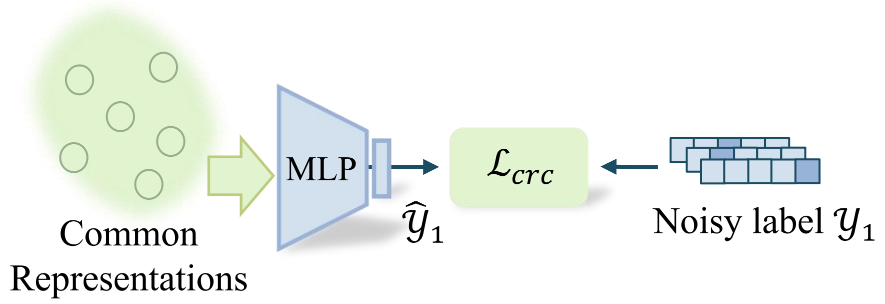
Shared Space Consistency Learning (SSCL)

❖ Multimodal Gap loss $\mathcal{L}_{mg}$



We adopt a Multimodal Gap loss to maximize the mutual information between different modalities from the instance-based perspective.

❖ Common Representations Classification loss $\mathcal{L}_{crc}$



We propose a Common Representations Classification loss (CRC) to narrow the gap between the common space and label space.

Optimization

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Robust Discriminative Center Learning (RDCL)

$\mathcal{L}_{rdc}$

Shared Space Consistency Learning (SSCL)

$\mathcal{L}_{crc}$ $\mathcal{L}_{mg}$

Final loss:

$$\mathcal{L} = \underbrace{\mathcal{L}_{rdc}}_{RDCL} + \underbrace{\beta_{mg}\mathcal{L}_{mg} + \beta_{crc}\mathcal{L}_{crc}}_{SSCL}$$

where β_{mg} and β_{crc} are the trade-off parameters. By minimizing the joint loss function, our RONO can be iteratively optimized in a batch-by-batch manner as the right algorithm.

Algorithm 1 Main optimization process of our RONO

Input: The training K-category multimodal data $\mathcal{D} = \{\mathcal{M}_j\}_{j=1}^{M}$, where $\mathcal{M}_j = \{(\boldsymbol{x}_i^j, y_i^j)\}_{i=1}^{N}$, maximal epoch number N_e and learning rate lr.

1: Randomly initialize the center of each category in the common space $\boldsymbol{C} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_K\}$.

2: **for** i = 1, 2, $\cdots$, N_e **do**

3: Calculate the common representations $f_i(\boldsymbol{x}_i^j)$ for all samples of the batch through the modality-specific extractors $\{f_i(\Theta_i)\}_{i=1}^{M}$, and use them for classification through a common classifier $g(\Gamma)$.

4: Normalize the $\boldsymbol{C} = \{\boldsymbol{c}_1, \cdots, \boldsymbol{c}_K\}$.

5: Calculate RDC, MG and CRC on the batch.

6: Update the network parameters $\{\Theta_i\}_{i=1}^{M}$, Γ and $\boldsymbol{C}$ by minimizing the loss $\mathcal{L}$ with descending their stochastic gradient:

$$\Theta_i = \Theta_i - lr \cdot (\frac{\partial \mathcal{L}_{rdc}}{\partial \Theta_i} + \beta_{mg}\frac{\partial \mathcal{L}_{mg}}{\partial \Theta_i} + \beta_{crc}\frac{\partial \mathcal{L}_{crc}}{\partial \Theta_i}),$$

$$\Gamma = \Gamma - lr \cdot (\beta_{crc}\frac{\partial \mathcal{L}_{crc}}{\partial \Gamma}),$$

$$\boldsymbol{C} = \boldsymbol{C} - lr \cdot (\frac{\partial \mathcal{L}_{rdc}}{\partial \boldsymbol{C}}), \text{ for } i, j = 1, \cdots, M.$$

7: **end for**

Output: Optimized network parameter $\{\Theta_i\}_{i=1}^{M}$.

Experiment Results

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

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 and RGB-D object datasets. The highest mAPs are shown in **bold** and the second highest mAPs are underlined.

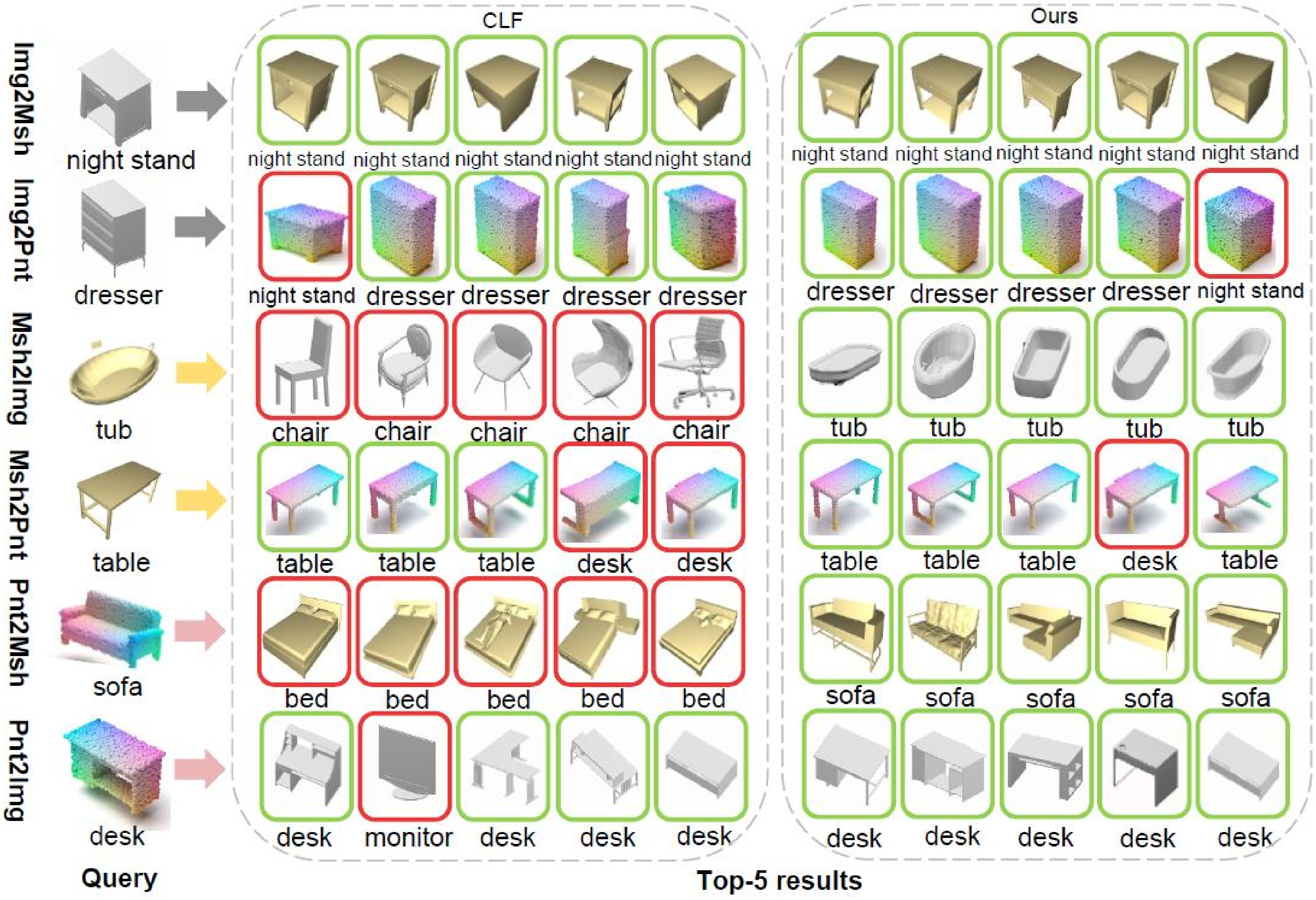
Method	3D MNIST [38]								RGB-D object [22]							
	Image→Point Cloud				Point Cloud→Image				Image→Point Cloud				Point Cloud→Image			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
CCA [15]	0.415	0.415	0.415	0.415	0.414	0.414	0.414	0.414	0.135	0.135	0.135	0.135	0.133	0.133	0.133	0.133
DCCA [1]	0.595	0.595	0.595	0.595	0.593	0.593	0.593	0.593	0.211	0.211	0.211	0.211	0.215	0.215	0.215	0.215
DCCAE [35]	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.600	0.217	0.217	0.217	0.217	0.218	0.218	0.218	0.218
DGCPN [42]	0.792	0.792	0.792	0.792	0.783	0.783	0.783	0.783	0.138	0.138	0.138	0.138	0.142	0.142	0.142	0.142
UCCH [17]	0.791	0.791	0.791	0.791	0.790	0.790	0.790	0.790	0.309	0.309	0.309	0.309	0.307	0.307	0.307	0.307
GMA [30]	0.449	0.438	0.426	0.415	0.437	0.432	0.423	0.414	0.090	0.085	0.088	0.089	0.087	0.083	0.087	0.086
MvDA [20]	0.481	0.461	0.432	0.328	0.482	0.461	0.431	0.323	0.133	0.132	0.128	0.112	0.132	0.133	0.109	0.102
AGAH [12]	0.688	0.557	0.128	0.108	0.680	0.548	0.122	0.116	0.608	0.379	0.195	0.090	0.601	0.380	0.194	0.090
DADH [4]	0.735	0.632	0.403	0.290	0.727	0.614	0.382	0.286	0.626	0.334	0.136	0.062	0.618	0.326	0.135	0.062
DAGNN [28]	0.883	0.850	0.749	0.445	0.879	0.845	0.743	0.435	0.707	0.637	0.520	0.315	0.715	0.635	0.513	0.313
ALGCN [27]	0.874	0.840	0.757	0.401	0.868	0.831	0.748	0.385	0.673	0.501	0.398	0.204	0.670	0.501	0.414	0.200
DSCMR [44]	0.908	0.812	0.512	0.219	0.896	0.811	0.472	0.140	0.727	0.671	0.532	0.290	0.731	0.673	0.523	0.275
MRL [16]	0.955	0.937	0.918	0.785	0.944	0.931	0.905	0.791	0.711	0.646	0.610	0.498	0.709	0.623	0.598	0.487
CLF [19]	0.890	0.811	0.460	0.124	0.872	0.793	0.426	0.120	0.723	0.661	0.346	0.111	0.703	0.634	0.343	0.106
CLF [19]+MAE [10]	0.810	0.812	0.501	0.122	0.809	0.811	0.483	0.122	0.705	0.626	0.426	0.187	0.703	0.624	0.399	0.167
Ours	**0.962**	**0.952**	**0.931**	**0.831**	**0.948**	**0.934**	**0.915**	**0.828**	**0.774**	**0.737**	**0.736**	**0.706**	**0.771**	**0.730**	**0.729**	**0.700**

Ablation Study

Ablation studies for RONO on the 3D MNIST and ModelNet40 datasets with 0.4 symmetric noise. ✓ stands for use.

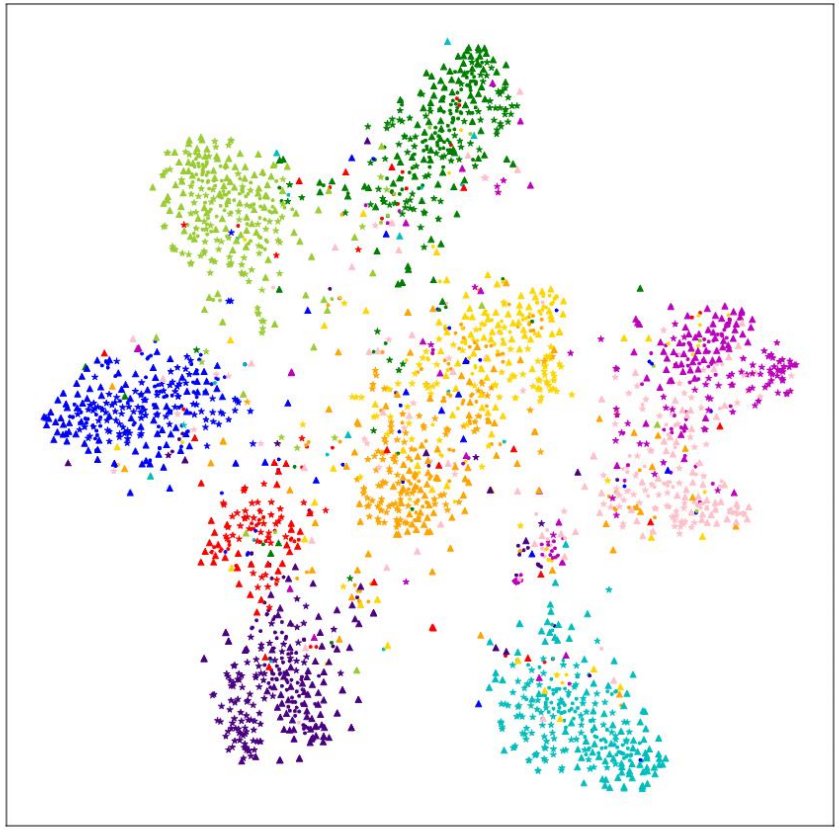
RDCL	SSCL		3D MNIST [37]				ModelNet40 [36]			
$\mathcal{L}_{rdc}$	$\mathcal{L}_{mg}$	$\mathcal{L}_{crc}$	Image→Point Cloud 0.4	Image→Point Cloud 0.8	Point Cloud→Image 0.4	Point Cloud→Image 0.8	Image→Point Cloud 0.4	Image→Point Cloud 0.8	Point Cloud→Image 0.4	Point Cloud→Image 0.8
✓	✓	✓	**0.952**	**0.831**	**0.934**	**0.828**	**0.858**	**0.824**	**0.854**	**0.821**
$\mathcal{L}'_{rdc}$	✓	✓	0.891	0.342	0.899	0.314	0.815	0.670	0.814	0.647
	✓	✓	0.841	0.675	0.832	0.671	0.821	0.777	0.820	0.770
✓		✓	0.929	0.615	0.927	0.564	0.793	0.551	0.791	0.536
✓	✓		0.930	0.709	0.914	0.706	0.804	0.673	0.806	0.653
✓			0.641	0.423	0.635	0.429	0.735	0.608	0.708	0.588
	✓		0.661	0.661	0.660	0.660	0.444	0.444	0.444	0.444
		✓	0.527	0.348	0.537	0.275	0.725	0.484	0.712	0.498

Visualization Experiments

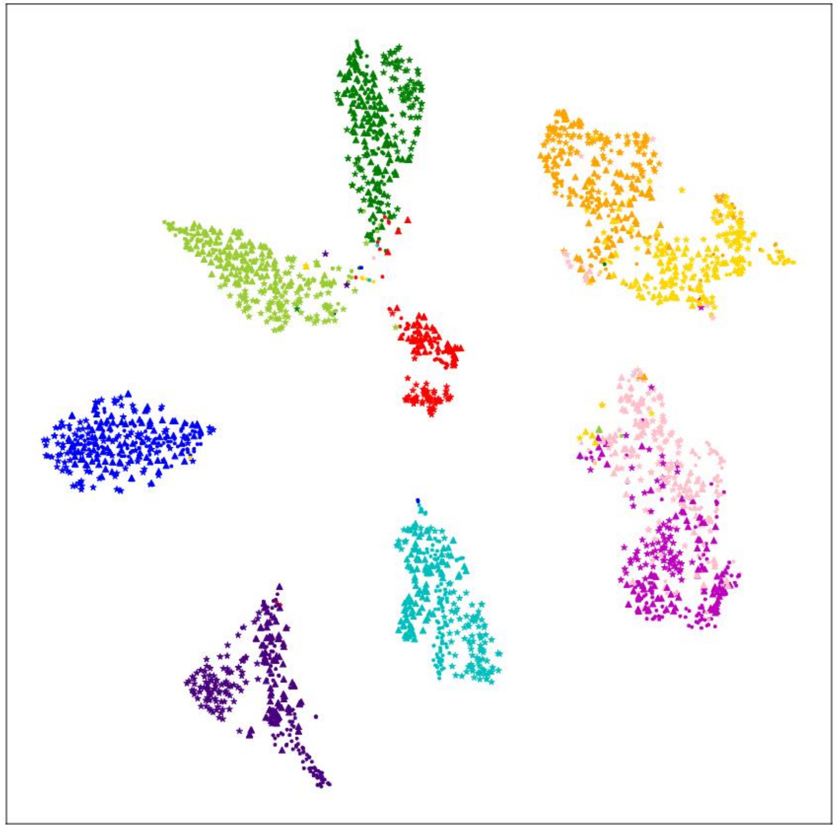


Top-5 retrieved results of CLF and our RONO under 0.4 label noise on the tri-modal ModelNet40 dataset.

Visualization Experiments



vs.



The representation visualization on the testing set of ModelNet40 by using t-SNE method. CLF and our RONO are trained under 0.4 symmetric label noise. Samples from the same category are rendered with the same color, and ones from the same modality are rendered with the same marker.

Thanks for watching!