Supplementary Material

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

In this supplementary material, we provide complementary information on theory and experiments. Specifically, we mainly supplement the proofs of mathematical properties and lemmas of RDC in AppendixA. In AppendixB, we give the implementation details of our RONO. In AppendixC.1, supplementary experimental results are shown and we give some insightful observations. In AppendixC.2, we give an experiment-based parameter analysis.

A. Mathematical Proof of Robustness

In this section, we further supplement **Section 3.4** with clear and accessible proofs of Property 1, Property 2 and Lemma 1. Since DNNs are noise tolerant in the early training stage [2], we only discuss the robustness of our RDC during the latter training stage as the balanced parameter v in RDC has dynamically increased to 1.

A.1. Mathematical Properties of RDC

The theoretical results [7,30] show that symmetric loss is noise-tolerant under the symmetric label noise and asymmetric noise. It is defined as:

$$\sum_{i=1}^{K}\mathcal{L}(f(\boldsymbol{x}),i)=C,\forall\boldsymbol{x}\in\mathcal{X},\forall f, \tag{1}$$

where $\mathcal{L}(f(\boldsymbol{x}),y)$ means the loss function which is calculated from model calculation results $f(\boldsymbol{x})$ and labels y.

In the training stage after the memorization effect of the neural networks [2] passed, our RDC can be simplified as:

$$\mathcal{L}_{rdc}=\\-\frac{1}{MN}\sum_{i}^{N}\sum_{j}^{M}\left|\frac{\sum_{k}^{K}e^{\left(\boldsymbol{c}_{k\neq y_i^j}\right)^T\boldsymbol{z}_i^j}}{K-1}-e^{\left(\boldsymbol{c}_{k=y_i^j}\right)^T\boldsymbol{z}_i^j}+\alpha\right|, \tag{2}$$

We can obviously get the upper and lower definite bound of our RDC, defined as:

$$\mathcal{L}_{rdc}\in[-(2e+|\alpha|),0], \tag{3}$$

where $\alpha\in[-e,e]$.

We define any sample belonging to any modality $\boldsymbol{x}\in\mathcal{X}$, then obtain common representation $\boldsymbol{z}=f(\boldsymbol{x}),\forall f$. For our RDC, calculating Equation (1) yields:

$$\sum_{i=1}^{K}\mathcal{L}_{rdc}(f(\boldsymbol{x}),i)=-\sum_{i}^{K}\left|\frac{\sum_{j}^{K}e^{(\boldsymbol{c}_{j\neq i})^T\boldsymbol{z}}}{K-1}-e^{(\boldsymbol{c}_i)^T\boldsymbol{z}}+\alpha\right|. \tag{4}$$

Due to the memorization effect of the DNNs, the common representations are more similar to their real category centers. Thus, for noisy samples ($i\neq y^*$ in Equation (4), where y^* means the really true label of the $\boldsymbol{x}$.), we get:

$$\frac{\sum_{j}^{K}e^{(\boldsymbol{c}_{j\neq i})^T\boldsymbol{z}}}{K-1}-e^{(\boldsymbol{c}_i)^T\boldsymbol{z}}+\alpha>0, \tag{5}$$

The function of α is to separate the results within the absolute value obtained from the clean samples and the noise samples at 0.

And for clean samples ($i=y^*$ in Eq.(4)), we get

$$\frac{\sum_{j}^{K}e^{(\boldsymbol{c}_{j\neq i})^T\boldsymbol{z}}}{K-1}-e^{(\boldsymbol{c}_i)^T\boldsymbol{z}}+\alpha<0. \tag{6}$$

So we can remove the limit of absolute value:

$$\begin{aligned}&\sum_{i=1}^{K}\mathcal{L}_{rdc}(f(\boldsymbol{x}),i)\\&=\Big(\frac{\sum_{j}^{K}e^{(\boldsymbol{c}_{j\neq y^*})^T\boldsymbol{z}}}{K-1}-e^{(\boldsymbol{c}_{y^*})^T\boldsymbol{z}}+\alpha\Big)\\&\quad-\sum_{i\neq y^*}^{K}\Big(\frac{\sum_{j}^{K}e^{(\boldsymbol{c}_{j\neq i})^T\boldsymbol{z}}}{K-1}-e^{(\boldsymbol{c}_i)^T\boldsymbol{z}}+\alpha\Big)\\&=-(K-1)\Big(\frac{\sum_{j}^{K}e^{(\boldsymbol{c}_{j\neq y^*})^T\boldsymbol{z}}}{K-1}-e^{(\boldsymbol{c}_{y^*})^T\boldsymbol{z}}+\alpha\Big)+C'\\&=(K-1)\mathcal{L}_{rdc}(f(\boldsymbol{x}),y^*)+C\end{aligned} \tag{7}$$

where C', C is a constant, $\mathcal{L}(f(\boldsymbol{x}),y^*)$ represents the loss between $\boldsymbol{z}=f(\boldsymbol{x})$ and its real category center.

A.2. Symmetric Label Noise Tolerance of RDC

Assuming that our RDC is under symmetric or uniform label noise. The definitions of $R_{\mathcal{L}_{rdc}}(f)$ and $R_{\mathcal{L}_{rdc}}(f^*)$ are consistent with those of $R_{\mathcal{L}_{mae}}(f)$ and $R_{\mathcal{L}_{mae}}(f^*)$.

Recall that for $f(\boldsymbol{x}), \forall \boldsymbol{x}, \forall f$,

$$R_{\mathcal{L}_{rdc}}(f) = \mathbb{E}_{\boldsymbol{x},y}\mathcal{L}_{rdc}(f(\boldsymbol{x}), y) \tag{8}$$

For uniform noise, we have, for any f,

$$\begin{aligned}
&R^{\eta}_{\mathcal{L}_{rdc}}(f) \\
=&\mathbb{E}_{\boldsymbol{x},y}\mathcal{L}_{rdc}(f(\boldsymbol{x}), y) \\
=&\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y^*|\boldsymbol{x}}\mathbb{E}_{y|\boldsymbol{x},y^*}\mathcal{L}_{rdc}(f(\boldsymbol{x}), y) \\
=&\mathbb{E}_{\boldsymbol{x}}\mathbb{E}_{y^*|\boldsymbol{x}}[(1-\eta)\mathcal{L}_{rdc}(f(\boldsymbol{x}), y^*) \\
&+ \frac{\eta}{K-1}\sum_{i\neq y^*}\mathcal{L}_{rdc}(f(\boldsymbol{x}), i)] \\
=&(1-\eta)R_{\mathcal{L}_{rdc}}(f) \\
&+ \frac{\eta}{K-1}(((K-1)R_{\mathcal{L}_{rdc}}(f) + C) - R_{\mathcal{L}_{rdc}}(f)) \\
=&(1 - \frac{\eta}{K-1})R_{\mathcal{L}_{rdc}}(f) + C.
\end{aligned} \tag{9}$$

Thus, for any f,

$$\begin{aligned}
&R^{\eta}_{L}(f^*) - R^{\eta}_{\mathcal{L}_{rdc}}(f) \\
=&(1 - \frac{\eta}{K-1})(R_{\mathcal{L}_{rdc}}(f^*) - R_{\mathcal{L}_{rdc}}(f)) \leq 0
\end{aligned} \tag{10}$$

because $1 - \frac{\eta}{K-1} > 0$ and f^* is a minimizer of $R_{\mathcal{L}_{rdc}}$. This proves f^* is also minimizer of risk under uniform noise.

A.3. Asymmetric Label Noise Tolerance of RDC

Assuming that our RDC is under asymmetric or class-conditional label noise.

From Eq.(7), we could derive the following:

$$\sum_{i\neq y^*}\mathcal{L}_{rdc}(f(\boldsymbol{x}), i) = (K-2)\mathcal{L}_{rdc}(f(\boldsymbol{x}), y^*) + C \tag{11}$$

It then yields the following derivations:

$$\begin{aligned}
&R^{\eta}_{\mathcal{L}_{rdc}}(f) \\
=&\mathbb{E}_{\boldsymbol{x},y}(1-\eta_y)\mathcal{L}_{rdc}(f(\boldsymbol{x}), y^*) + \mathbb{E}_{\boldsymbol{x},y}\sum_{i\neq y^*}\eta_{yi}\mathcal{L}_{rdc}(f(\boldsymbol{x}), i) \\
=&\mathbb{E}_{\boldsymbol{x},y}(1-\eta_y)(\frac{\sum_{i\neq y^*}\mathcal{L}_{rdc}(f(\boldsymbol{x}), i) - C}{K-2}) \\
&+ \mathbb{E}_{\boldsymbol{x},y}\sum_{i\neq y^*}\eta_{yi}\mathcal{L}_{rdc}(f(\boldsymbol{x}), i) \\
=&C\mathbb{E}_{\boldsymbol{x},y}(\frac{\eta_y - 1}{K-2}) + \mathbb{E}_{\boldsymbol{x},y}\sum_{i\neq y^*}(\frac{1-\eta_y}{K-2} + \eta_{yi})\mathcal{L}_{rdc}(f(\boldsymbol{x}), i),
\end{aligned} \tag{12}$$

where $1-\eta_y$ is the probability of a label being correct, and the noise condition η_{yi} generally states that a sample $\boldsymbol{x}$ still has the highest probability of being in the correct category, though it has probability of η_{yi} being in an arbitrary noisy (incorrect) category $i \neq y^*$. Since f^*_η is the minimizer of $R^{\eta}_{\mathcal{L}_{rdc}}$, we have $R^{\eta}_{\mathcal{L}_{rdc}}(f^*_\eta) - R^{\eta}_{\mathcal{L}_{rdc}}(f^*) \leq 0$, and hance from Eq.(12) we have:

$$\mathbb{E}_{\boldsymbol{x},y}\sum_{i\neq y^*}(\frac{1-\eta_y}{K-2} + \eta_{yi})(\mathcal{L}(f^*_\eta(\boldsymbol{x}), i) - \mathcal{L}(f^*(\boldsymbol{x}), i)) \leq 0 \tag{13}$$

According to the characteristics of our RDC, when $\alpha \geq 0$, $R_{\mathcal{L}_{rdc}}(f^*(\boldsymbol{x}), i) = -(2e+\alpha)$ for $i \neq y^*$, which is the infimum of our RDC. As $\frac{1-\eta_y}{K-2} + \eta_{yi} > 0$ in Eq.(13), so $R_{\mathcal{L}_{rdc}}(f^*_\eta(\boldsymbol{x}), i) = -(2e+\alpha)$ for $i \neq y^*$, obtaining $R_{\mathcal{L}_{rdc}}(f^*(\boldsymbol{x}), i) = R_{\mathcal{L}_{rdc}}(f^*_\eta(\boldsymbol{x}), i)$ for $i \neq y^*$.

From Eq.(11), we can obtain:

$$\begin{aligned}
&\sum_{i\neq y^*}\mathcal{L}_{rdc}(f^*_\eta(\boldsymbol{x}), i) - \mathcal{L}_{rdc}(f^*(\boldsymbol{x}), i) \\
=&(K-2)(\mathcal{L}_{rdc}(f^*_\eta(\boldsymbol{x}), y^*) - \mathcal{L}_{rdc}(f^*(\boldsymbol{x}), y^*))
\end{aligned} \tag{14}$$

Therefore, we obtain $\mathcal{L}_{rdc}(f^*_\eta(\boldsymbol{x}), y^*) = \mathcal{L}_{rdc}(f^*(\boldsymbol{x}), y^*)$. On this account, we obtain $R_{\mathcal{L}_{rdc}}(f^*(\boldsymbol{x}), i) = R_{\mathcal{L}_{rdc}}(f^*_\eta(\boldsymbol{x}), i)$ for $i = 1, \cdots, K$, which can also be written as $R_{\mathcal{L}_{rdc}}(f^*) = R_{\mathcal{L}_{rdc}}(f^*_\eta)$, so we finally proof when $\alpha \geq 0$, our RDC is asymmetric noise tolerance.

B. Implementation Details

Algorithm 1 Main optimization process of our RONO

Input: The training K-category multimodal data $\mathcal{D} = \{\mathcal{M}_j\}_{j=1}^{M}$, where $\mathcal{M}_j = \{(\boldsymbol{x}^j_i, y^j_i)\}_{i=1}^{N}$, 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 $f_i(\boldsymbol{x}^j_i)$ 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:
$\Theta_i = \Theta_i - lr \cdot (\frac{\partial \mathcal{L}_{rdc}}{\partial \Theta_i} + \beta_{mg}\frac{\partial \mathcal{L}_{mg}}{\partial \Theta_i} + \beta_{crc}\frac{\partial \mathcal{L}_{crc}}{\partial \Theta_i})$,
$\Gamma = \Gamma - lr \cdot (\beta_{crc}\frac{\partial \mathcal{L}_{crc}}{\partial \Gamma})$,
$\boldsymbol{C} = \boldsymbol{C} - lr \cdot (\frac{\partial \mathcal{L}_{rdc}}{\partial \boldsymbol{C}})$, for $i, j = 1, \cdots, M$.
7: **end for**

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

In this work, we adopt the ResNet18 [9] as the backbone network for 2D image feature extraction, dynamic graph convolutional neural network (DGCNN) [25] for 3D point cloud feature extraction and MeshNet [5] for mesh feature

	ModelNet10 [26]						ModelNet40 [26]					
	Img→Pnt			Pnt→Img			Img→Pnt			Pnt→Img		
	0.1	0.2	0.4	0.1	0.2	0.4	0.1	0.2	0.4	0.1	0.2	0.4
CCA [12]	0.625	0.625	0.625	0.627	0.627	0.627	0.532	0.532	0.532	0.531	0.531	0.531
DCCA [1]	0.684	0.684	0.684	0.678	0.678	0.678	0.584	0.584	0.584	0.569	0.569	0.569
DCCAE [24]	0.703	0.703	0.703	0.693	0.693	0.693	0.593	0.593	0.593	0.572	0.572	0.572
DGCPN [29]	0.765	0.765	0.765	0.759	0.759	0.759	0.705	0.705	0.705	0.697	0.697	0.697
UCCH [14]	0.771	0.771	0.771	0.770	0.770	0.770	0.755	0.755	0.755	0.739	0.739	0.739
GMA [22]	0.649	0.634	0.619	0.617	0.615	0.599	0.512	0.510	0.498	0.503	0.498	0.486
MvDA [16]	0.586	0.539	0.483	0.550	0.500	0.455	0.421	0.402	0.367	0.409	0.391	0.370
AGAH [8]	0.821	0.805	0.756	0.827	0.801	0.743	0.817	0.778	0.778	0.800	0.779	0.761
DADH [3]	0.845	0.810	0.749	0.828	0.805	0.723	0.825	0.798	0.782	0.823	0.777	0.776
DAGNN [21]	0.814	0.756	0.672	0.807	0.738	0.671	0.840	0.822	0.763	0.837	0.810	0.751
ALGCN [20]	0.784	0.701	0.542	0.758	0.721	0.531	0.761	0.687	0.526	0.754	0.653	0.523
DSCMR [31]	0.851	0.838	0.675	0.825	0.810	0.661	0.831	0.811	0.656	0.819	0.804	0.651
MRL [13]	<u>0.869</u>	<u>0.867</u>	<u>0.859</u>	<u>0.865</u>	<u>0.861</u>	<u>0.854</u>	0.846	<u>0.838</u>	<u>0.811</u>	0.844	<u>0.838</u>	<u>0.799</u>
CLF [15]	0.856	0.803	0.741	0.840	0.798	0.743	<u>0.855</u>	0.820	0.757	<u>0.852</u>	0.813	0.758
CLF [15]+MAE [7]	0.848	0.794	0.771	0.841	0.791	0.754	0.837	0.811	0.761	0.832	0.798	0.763
Ours	**0.885**	**0.875**	**0.863**	**0.875**	**0.860**	**0.857**	**0.861**	**0.852**	**0.827**	**0.854**	**0.845**	**0.822**

Table 1. Performance comparison under the asymmetric noise rates of 0.1, 0.2 and 0.4 on the ModelNet10 and ModelNet40 datasets. The highest mAPs are shown in **bold** and the second highest mAPs are <u>underlined</u>.

extraction. And all the features are projected as 512D common representations by two fully connected layers. We adopt two fully connected layers as the common classifier $g(\Gamma)$ for our common representation classification. For our overall framework optimization, we employ ADAM [18] as our optimizer and the optimization process is shown in Algorithm1.

For all datasets, the learning rate is initialized with 0.0001, batch size is set as 128 and temperature parameter in MG is set as 1. We use a maximal epoch number of 100 for 3D MNIST [27] dataset and 400 for RGB-D object [19], ModelNet10 [26] and ModelNet40 [26] datasets. It is worth noting that for the optimal selection of the hyperparameters α, β_{mg} and β_{crc}, we have used experiments in AppendixC.2for analysis.

C. More Experimental Results

Due to space limitations, a portion of the experiments we conducted could not be shown completely in the main body of our paper, so they will be shown in this section additionally.

C.1. More Comparative Experimental Results and Analysis

We have added a total of four comparative experiments: 1) To fully demonstrate the effectiveness of RONO under asymmetric noise, we have also conducted experiments on ModelNet10 and ModelNet40 datasets under 0.1, 0.2 and 0.4 asymmetric label noise. 2) In addition, we have conducted comparative experiments on four datasets we used without any synthetic noise (part of the experimental results have been shown in the main body of our paper). 3) We have not only conducted 2D-3D cross-modal retrieval experiments across three modal (i.e., Image, Mesh, and Point cloud) on ModelNet40 dataset, but also on ModelNet10 dataset under 0, 0.2, 0.4, 0.6 and 0.8 label noise, by comparing RONO with state-of-the-art CLF [15]. 4) To verify our RONO is superior in each domain, we conducted in-domain retrieval experiments on ModelNet40 dataset without synthetic noise, by comparing our RONO with several in-domain methods, (i.e., MVCNN [23], GIFT [4], SPNet [28], TCL [10], VNN [11], DGCNN [25], DLAN [6], SPH [17] and MeshNet [5]) which are taken from the image, mesh and point cloud domains, respectively.

The experimental results are shown in Table1, Table2, Table3and Table4, respectively, and we could draw the following observations:

- Despite such complicated conditions as asymmetric noise, our RONO remains superiority by virtue of noise robustness.
- Our RONO shows superior even without the addition of synthetic label noise in four datasets, further demonstrating that well-annotated datasets also contain noise impacting the performance of each non-robust method.
- Our RONO is not only superior in 2D-3D cross-modal

Methods	3D MNIST [27]		RGB-D object [19]		ModelNet10 [26]		ModelNet40 [26]	
	Img→Pnt	Pnt→Img	Img→Pnt	Pnt→Img	Img→Pnt	Pnt→Img	Img→Pnt	Pnt→Img
CCA [12]	0.415	0.415	0.135	0.133	0.625	0.627	0.532	0.531
DCCA [1]	0.595	0.593	0.211	0.215	0.684	0.678	0.584	0.569
DCCAE [24]	0.600	0.600	0.217	0.218	0.703	0.693	0.593	0.572
DGCPN [29]	0.792	0.783	0.138	0.142	0.765	0.759	0.705	0.697
UCCH [14]	0.791	0.790	0.309	0.307	0.771	0.770	0.755	0.739
GMA [22]	0.514	0.500	0.126	0.121	0.673	0.658	0.558	0.530
MvDA [16]	0.530	0.508	0.188	0.199	0.557	0.527	0.457	0.444
AGAH [8]	0.967	0.961	0.652	0.628	0.862	0.867	0.807	0.799
DADH [3]	0.971	**0.969**	0.772	0.761	0.889	0.884	0.836	0.824
DAGNN [21]	0.927	0.927	0.741	0.724	0.867	0.864	0.825	0.820
ALGCN [20]	0.908	0.900	0.717	0.691	0.815	0.799	0.785	0.791
DSCMR [31]	0.963	0.959	0.774	0.768	0.849	0.842	0.867	0.866
MRL [13]	0.963	0.945	0.723	0.719	0.887	0.871	0.848	0.843
CLF [15]	**0.983**	0.958	0.772	0.766	0.884	0.867	0.871	0.878
CLF [15]+MAE [7]	0.971	0.951	0.752	0.741	0.877	0.853	0.864	0.853
Ours	**0.983**	0.968	**0.779**	**0.771**	**0.892**	**0.892**	**0.883**	**0.881**

Table 2. Performance comparison in terms of mAP from image to point cloud (Img → Pnt) and from point cloud to image (Pnt→Img) retrieval without noise on the 3D MNIST, RGB-D object, ModelNet10 and ModelNet40 datasets. The highest mAPs are shown in **bold** and the second highest mAPs are underlined.

η	Qry	Img			Msh			Pnt		
	Retrv	Img	Msh	Pnt	Img	Msh	Pnt	Img	Msh	Pnt
0	CLF	0.903	**0.907**	0.895	0.889	0.916	0.900	0.887	0.893	0.885
	Ours	**0.913**	0.906	**0.898**	**0.896**	**0.919**	**0.904**	**0.895**	**0.903**	**0.892**
0.2	CLF	0.829	0.848	0.847	0.841	0.871	0.866	0.838	0.865	0.873
	Ours	**0.871**	**0.889**	**0.877**	**0.890**	**0.912**	**0.905**	**0.872**	**0.899**	**0.895**
0.4	CLF	0.762	0.790	0.788	0.786	0.810	0.795	0.772	0.785	0.825
	Ours	**0.866**	**0.888**	**0.878**	**0.883**	**0.911**	**0.900**	**0.865**	**0.897**	**0.895**
0.6	CLF	0.572	0.572	0.633	0.567	0.583	0.617	0.606	0.578	0.749
	Ours	**0.840**	**0.857**	**0.850**	**0.868**	**0.901**	**0.892**	**0.854**	**0.888**	**0.892**
0.8	CLF	0.315	0.218	0.237	0.258	0.304	0.246	0.258	0.212	0.449
	Ours	**0.826**	**0.859**	**0.849**	**0.858**	**0.898**	**0.887**	**0.842**	**0.880**	**0.885**

Table 3. Performance comparison of CLF [15] and our RONO under the symmetric noise rates of 0, 0.2, 0.4, 0.6 and 0.8 on tri-modal (Image, Mesh, Point cloud) ModenNet10 dataset [26]. Under each noise condition, the highest mAPs are shown in **bold**.

Domain	Method	mAP
Img	MVCNN [23]	0.802
	GIFT [4]	0.819
	SPNet [28]	0.852
	TCL [10]	0.880
	VNN [11]	0.893
	Ours	**0.911**
Pnt	DGCNN [25]	0.848
	DLAN [6]	0.850
	Ours	**0.891**
Msh	SPH [17]	0.333
	MeshNet [5]	0.819
	Ours	**0.901**

Table 4. Comparison with the state-of-the-art in-domain retrieval methods on tri-modal ModelNet40 Dataset without noise. In each domain, the highest mAPs are shown in **bold**.

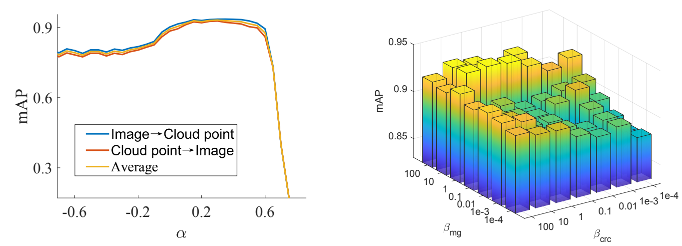
retrieval but also maintains its superiority in in-domain retrieval by making full use of the mutual information between modalities.

C.2. Parameter Analysis

To investigate the parameter sensitivity of our method, we plot the average mAP scores of cross-modal retrieval versus different hyper-parameters (i.e., α, β_{mg} and β_{crc}) on the test sets of 3D MNIST as shown in Figure1. From Figure1a, one could see that our RONO could achieve stable superior performance when α is in the range of 0.1 to 0.5, thus indicating that our method is insensitive to α in the range. From Figure1b, one could find that each component of $\mathcal{L}$ contributes to the model which is consistent with our ablation study. To be specific, our method could achieve stable comparable performance when β_{mg} is in the range of 10 to 100 and β_{crc} is in the range of 0.1 to 10.

References

[1]Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255.



(a) Impact of different α (b) Impact of different β_{crc}, β_{mg}

Figure 1. Cross-modal retrieval performance of our RONO in terms of MAP versus different values of α, β_{mg} and β_{crc} on the test sets of the 3D MNIST datasets. The noise rate is 0.4.

PMLR, 2013.3,4

[2]Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.1

[3]Cong Bai, Chao Zeng, Qing Ma, Jinglin Zhang, and Shengyong Chen. Deep adversarial discrete hashing for cross-modal retrieval. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 525–531, 2020.3,4

[4]Song Bai, Xiang Bai, Zhichao Zhou, Zhaoxiang Zhang, and Longin Jan Latecki. Gift: A real-time and scalable 3d shape search engine. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5023–5032, 2016.3,4

[5]Yutong Feng, Yifan Feng, Haoxuan You, Xibin Zhao, and Yue Gao. Meshnet: Mesh neural network for 3d shape representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8279–8286, 2019.2,3,4

[6]Takahiko Furuya and Ryutarou Ohbuchi. Deep aggregation of local 3d geometric features for 3d model retrieval. In *BMVC*, volume 7, page 8, 2016.3,4

[7]Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.1,3,4

[8]Wen Gu, Xiaoyan Gu, Jingzi Gu, Bo Li, Zhi Xiong, and Weiping Wang. Adversary guided asymmetric hashing for cross-modal retrieval. In *Proceedings of the 2019 on international conference on multimedia retrieval*, pages 159–167, 2019.3,4

[9]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.2

[10]Xinwei He, Yang Zhou, Zhichao Zhou, Song Bai, and Xiang Bai. Triplet-center loss for multi-view 3d object retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1945–1954, 2018.3,4

[11]Xinwei He, Tengteng Huang, Song Bai, and Xiang Bai. View n-gram network for 3d object retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7515–7524, 2019.3,4

[12]Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992. 3,4

[13]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 (CVPR)*, pages 5403–5413, June 2021.3,4

[14]Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.3,4

[15]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, 2021.3,4

[16]Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):188–194, 2015.3,4

[17]Michael Kazhdan, Thomas Funkhouser, and Szymon Rusinkiewicz. Rotation invariant spherical harmonic representation of 3 d shape descriptors. In *Symposium on geometry processing*, volume 6, pages 156–164, 2003.3,4

[18]Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.3

[19]Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *2011 IEEE International Conference on Robotics and Automation*, pages 1817–1824, 2011. doi: 10.1109/ICRA.2011.5980382.3,4

[20]Shengsheng Qian, Dizhan Xue, Quan Fang, and Changsheng Xu. Adaptive label-aware graph convolutional networks for cross-modal retrieval. *IEEE Transactions on Multimedia*, pages 1–1, 2021. doi: 10.1109/TMM.2021.3101642.3,4

[21]Shengsheng Qian, Dizhan Xue, Huaiwen Zhang, Quan Fang, and Changsheng Xu. Dual adversarial graph neural networks for multi-label cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2440–2448, 2021.3,4

[22] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2160–2167. IEEE, 2012. 3, 4

[23] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 3, 4

[24] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015. 3, 4

[25] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019. 2, 3, 4

[26] 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 conference on computer vision and pattern recognition*, pages 1912–1920, 2015. 3, 4

[27] Xiaofan Xu, Alireza Dehghani, David Corrigan, Sam Caulfield, and David Moloney. Convolutional neural network for 3d object recognition using volumetric representation. In *2016 First International Workshop on Sensing, Processing and Learning for Intelligent Machines (SPLINE)*, pages 1–5, 2016. doi: 10.1109/SPLIM.2016.7528403. 3, 4

[28] Mohsen Yavartanoo, Eu Young Kim, and Kyoung Mu Lee. Spnet: Deep 3d object classification and retrieval using stereographic projection. In *Asian conference on computer vision*, pages 691–706. Springer, 2018. 3, 4

[29] Jun Yu, Hao Zhou, Yibing Zhan, and Dacheng Tao. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4626–4634, 2021. 3, 4

[30] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018. 1

[31] 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. 3, 4