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 Appendix A. In Appendix B, we give the implementation details of our RONO. In Appendix C, supplementary experimental results are shown and we give some insightful observations. In Appendix C.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:

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

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

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

\[
\mathcal{L}_{rdc} = 
\frac{1}{MN} \sum_{i=1}^{N} \sum_{j=1}^{M} \left| \sum_{k} e^{(c_{k,x,y}^{i})} \frac{z_{k}^{i}}{K-1} - e^{(c_{k,y}^{i})} z_{k}^{i} + \alpha \right|. \tag{2}
\]

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 \(x \in \mathcal{X}\), then obtain common representation \(z = f(x), \forall f\). For our RDC, calculating Equation (1) yields:

\[
\sum_{i=1}^{K} \mathcal{L}_{rdc}(f(x), i) = -\sum_{i} \left| \sum_{j=1}^{K} e^{(c_{j,x}^{i})} z_{j}^{i} \frac{z_{j}^{i}}{K-1} - e^{(c_{j})} z_{j} + \alpha \right|. \tag{4}
\]

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 \(x\), we get:

\[
\frac{\sum_{j} e^{(c_{j,x}^{i})} z_{j}}{K-1} - e^{(c_{j})} z_{j} + \alpha > 0, \tag{5}
\]

The function of \(\alpha\) 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

\[
\frac{\sum_{j} e^{(c_{j,x}^{i})} z_{j}}{K-1} - e^{(c_{j})} z_{j} + \alpha < 0. \tag{6}
\]

So we can remove the limit of absolute value:

\[
\sum_{i=1}^{K} \mathcal{L}_{rdc}(f(x), i) = (\frac{\sum_{j} e^{(c_{j,x}^{i})} z_{j}}{K-1} - e^{(c_{j})} z_{j} + \alpha) \tag{7}
\]

\[
= - (K-1)(\frac{\sum_{j} e^{(c_{j,x}^{i})} z_{j}}{K-1} - e^{(c_{j})} z_{j} + \alpha) + C', \tag{7}
\]

where \(C', C\) is a constant, \(\mathcal{L}(f(x), y^{*})\) represents the loss between \(z = f(x)\) 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 \(\mathcal{L}_{rdc}(f)\) and \(\mathcal{L}_{rdc}(f^{*})\) are consistent with those of \(\mathcal{L}_{max}(f)\) and \(\mathcal{L}_{max}(f^{*})\).
Recall that for \( f(x), \forall x, \forall f \),
\[
R_{L_{rdc}}(f) = E_{x,y}L_{rdc}(f(x), y)
\]  \( (8) \)

For uniform noise, we have, for any \( f \),
\[
R^0_{L_{rdc}}(f) = E_{x,y}L_{rdc}(f(x), y)
\]
\[
= E_{x,y}E_{y^*}[E_{y|x,y^*}L_{rdc}(f(x), y)]
\]
\[
= E_{x,y}E_{y^*}[E_{y|x,y^*}[\eta(1-\eta)L_{rdc}(f(x), y^*) + \eta K^{-1} \sum_{y \neq y^*} L_{rdc}(f(x), i)]]
\]
\[
= (1-\eta)R_{L_{rdc}}(f) + \frac{\eta}{K-1}((K-1)R_{L_{rdc}}(f) + C) - R_{L_{rdc}}(f)
\]
\[
= (1-\eta)R_{L_{rdc}}(f) + \frac{\eta}{K-1}R_{L_{rdc}}(f) + C.
\]

Thus, for any \( f \),
\[
R^0_{L_{rdc}}(f^*) - R^0_{L_{rdc}}(f)
\]
\[
= (1 - \eta)R_{L_{rdc}}(f^*) - (1 - \eta)R_{L_{rdc}}(f)
\]
\[
= \frac{1}{K-1}((K-1)R_{L_{rdc}}(f^*) - \sum_{y \neq y^*} L_{rdc}(f^*(x), i)) \leq 0
\]
because \( 1 - \eta > 0 \) and \( f^* \) is a minimizer of \( R_{L_{rdc}} \). 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:
\[
\sum_{y \neq y^*} L_{rdc}(f(x), i) = (K-2)R_{L_{rdc}}(f(x), y^*) + C
\]
\( (11) \)

It then yields the following derivations:
\[
R^0_{L_{rdc}}(f^*) - R^0_{L_{rdc}}(f)
\]
\[
= \frac{1}{K-1}((K-1)R_{L_{rdc}}(f^*) - \sum_{y \neq y^*} L_{rdc}(f^*(x), i)) \leq 0
\]
\( (12) \)

\( \eta \) is the probability of a label being correct, and the noise condition \( \eta_{yi} \) generally states that a sample \( x \) still has the highest probability of being in the correct category, though it has probability of \( \eta_{yi} \) being in an arbitrary noisy (incorrect) category \( i \neq y^* \). Since \( f^*_i \) is the minimizer of \( R^0_{L_{rdc}} \), we have \( R^0_{L_{rdc}}(f^*_i) \leq 0 \), and hence from Eq.(12) we have:
\[
E_{x,y} \sum_{i \neq y^*} \left( \frac{1 - \eta_y}{K-2} + \eta_{yi} \right)(L(f^*_i(x), i) - L(f^*(x), i)) \leq 0
\]

According to the characteristics of our RDC, when \( \alpha \geq 0 \), \( R_{L_{rdc}}(f^*(x), i) = -2(\alpha + \gamma) \) for \( i \neq y^* \), which is the infimum of our RDC. As \( 1 - \frac{\eta_y}{K-2} + \gamma_{yi} > 0 \) in Eq.(13), so \( R_{L_{rdc}}(f^*_i(x), i) = -2(\alpha + \gamma) \) for \( i \neq y^* \), obtaining \( R_{L_{rdc}}(f^*(x), i) = R_{L_{rdc}}(f^*_i(x), i) \) for \( i \neq y^* \).

From Eq.(11), we can obtain:
\[
\sum_{y \neq y^*} L_{rdc}(f^*_i(x), i) - L_{rdc}(f^*(x), i)
\]
\[
= (K-2)(L_{rdc}(f^*_i(x), y^*) - L_{rdc}(f^*(x), y^*))
\]
Therefore, we obtain \( L_{rdc}(f^*_i(x), y^*) = L_{rdc}(f^*(x), y^*) \). On this account, we obtain \( R_{L_{rdc}}(f^*_i(x), i) = R_{L_{rdc}}(f^*(x), i) \), which can also be written as \( R_{L_{rdc}}(f^*_i(x), i) = R_{L_{rdc}}(f^*(x), i) \), so we finally prove 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 \( D = \{M_j\}_{j=1}^{M} \), where \( M_j = \{(x^j_i, y^j_i)\}_{i=1}^{N_j} \) maximal epoch number \( N_e \), and learning rate \( lr \).

1. Randomly initialize the center of each category in the common space \( C = \{c_1, \ldots, c_K\} \).
2. for \( i = 1, 2, \ldots, N_e \) do
3. \quad Calculate the common representations \( f_i(x^j_i) \) 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 \( g() \).
4. \quad Normalize the \( C = \{c_1, \ldots, c_K\} \).
5. \quad Calculate RDC, MG and CRC on the batch.
6. \quad Update the network parameters \( \{\Theta_i\}_{i=1}^{M} \), \( \Gamma \) and \( C \) by minimizing the loss \( L \) with descending their stochastic gradient:
   \( \Theta_i = \Theta_i - lr \cdot (\frac{\partial L}{\partial \Theta_i} + \beta_{mg} \frac{\partial L}{\partial g} + \beta_{crc} \frac{\partial L}{\partial \Theta_i}) \),
   \( \Gamma = \Gamma - lr \cdot (\gamma \frac{\partial L}{\partial \Gamma}) \),
   \( C = C - lr \cdot (\frac{\partial L}{\partial C}) \), for \( i, j = 1, \ldots, M \).
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 (DGCCN) [25] for 3D point cloud feature extraction and MeshNet [5] for mesh feature
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 underlined.

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

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, Table3 and 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
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.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Img→Pnt</td>
<td>Pnt→Img</td>
<td>Img→Pnt</td>
<td>Pnt→Img</td>
</tr>
<tr>
<td>CCA [12]</td>
<td>0.415</td>
<td>0.415</td>
<td>0.135</td>
<td>0.133</td>
</tr>
<tr>
<td>DCCA [1]</td>
<td>0.595</td>
<td>0.593</td>
<td>0.211</td>
<td>0.215</td>
</tr>
<tr>
<td>DCCAE [24]</td>
<td>0.600</td>
<td>0.600</td>
<td>0.217</td>
<td>0.218</td>
</tr>
<tr>
<td>DGCNN [29]</td>
<td>0.792</td>
<td>0.783</td>
<td>0.138</td>
<td>0.142</td>
</tr>
<tr>
<td>UCCH [14]</td>
<td>0.791</td>
<td>0.790</td>
<td>0.309</td>
<td>0.307</td>
</tr>
<tr>
<td>GMA [22]</td>
<td>0.514</td>
<td>0.500</td>
<td>0.126</td>
<td>0.121</td>
</tr>
<tr>
<td>MvDA [16]</td>
<td>0.530</td>
<td>0.508</td>
<td>0.188</td>
<td>0.199</td>
</tr>
<tr>
<td>AGAH [8]</td>
<td>0.967</td>
<td>0.961</td>
<td>0.652</td>
<td>0.628</td>
</tr>
<tr>
<td>DADH [3]</td>
<td>0.971</td>
<td><strong>0.969</strong></td>
<td>0.772</td>
<td>0.761</td>
</tr>
<tr>
<td>DAGNN [21]</td>
<td>0.927</td>
<td>0.927</td>
<td>0.741</td>
<td>0.724</td>
</tr>
<tr>
<td>ALGCN [20]</td>
<td>0.908</td>
<td>0.900</td>
<td>0.717</td>
<td>0.691</td>
</tr>
<tr>
<td>DSCMR [31]</td>
<td>0.963</td>
<td>0.959</td>
<td>0.774</td>
<td>0.768</td>
</tr>
<tr>
<td>MRL [13]</td>
<td>0.963</td>
<td>0.945</td>
<td>0.723</td>
<td>0.719</td>
</tr>
<tr>
<td>CLF [15]</td>
<td><strong>0.983</strong></td>
<td>0.958</td>
<td>0.772</td>
<td>0.766</td>
</tr>
<tr>
<td>CLF [15]+MAE [7]</td>
<td>0.971</td>
<td>0.951</td>
<td>0.752</td>
<td>0.741</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.983</strong></td>
<td><strong>0.968</strong></td>
<td><strong>0.779</strong></td>
<td><strong>0.771</strong></td>
</tr>
</tbody>
</table>

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) ModelNet10 dataset [26]. Under each noise condition, the highest mAPs are shown in bold.

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.

Table 5. Parameter analysis of our method, we plot the average mAP scores of cross-modal retrieval versus different hyper-parameters (i.e., $\alpha$, $\beta_{mg}$ and $\beta_{crc}$) on the test sets of 3D MNIST as shown in Figure 1. From Figure 1a, one could see that our RONO could achieve stable superior performance when $\alpha$ is in the range of 0.1 to 0.5, thus indicating that our method is insensitive to $\alpha$ in the range. From Figure 1b, one could find that each component of $L$ contributes to the model which is consistent with our ablation study. To be specific, our method could achieve stable comparable performance when $\beta_{mg}$ is in the range of 10 to 100 and $\beta_{crc}$ is in the range of 0.1 to 10.

**References**

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


