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

Yanglin Feng1 Hongyuan Zhu2
1College of Computer Science, Sichuan University
2Institute for Infocomm Research (I2R), A*STAR, Singapore

Dezhong Peng1,3,4 Xi Peng1 Peng Hu1*
3Sichuan Zhiqian Technology Co., Ltd
4Chengdu Ruibei Yingte Information Technology Co., Ltd

Abstract

Recently, with the advent of Metaverse and AI Generated Content, cross-modal retrieval becomes popular with a burst of 2D and 3D data. However, this problem is challenging given the heterogeneous structure and semantic discrepancies. Moreover, imperfect annotations are ubiquitous given the ambiguous 2D and 3D content, thus inevitably producing noisy labels to degrade the learning performance. To tackle the problem, this paper proposes a robust 2D-3D retrieval framework (RONO) to robustly learn from noisy multimodal data. Specifically, one novel Robust Discriminative Center Learning mechanism (RDCL) is proposed in RONO to adaptively distinguish clean and noisy samples for respectively providing them with positive and negative optimization directions. Besides, we present a Shared Space Consistency Learning mechanism (SSCL) to capture the intrinsic information inside the noisy data by minimizing the cross-modal and semantic discrepancy between common space and label space simultaneously. Comprehensive mathematical analyses are given to theoretically prove the noise tolerance of the proposed method. Furthermore, we conduct extensive experiments on four 3D-model multimodal datasets to verify the effectiveness of our method by comparing it with 15 state-of-the-art methods. Code is available at https://github.com/penghu-cs/RONO.

1. Introduction

Point-cloud retrieval (PCR) is fundamental and crucial for processing and analyzing 3D data [14], which could provide the direct technical support of the 3D data search engine, thus embracing compelling application prospects and practical value in the fields of robotics [7, 32], autonomous driving [23, 31], virtual/augmented reality [9], and medicine [29], etc. Different from 2D images, 3D point clouds could depict the internal architecture and external appearance of objects from distinct views/modalities. Hence, PCR is often accompanied by retrieving across diverse modalities, termed 2D-3D cross-modal retrieval [19].

On the other hand, it is extremely expensive and labor-intensive to label such a huge amount of data points [17, 41], not to mention the additional challenges of the missing color and texture of the point clouds. In order to reduce the labeling cost, we could utilize open source or low-cost annotation tools (e.g., point-cloud-annotation-tool [18], LabelHub, etc.), hence it will inevitably introduce label noise due to the non-expert annotation. However, almost all existing works excessively rely on well-labeled data [19, 20, 44], thus making them vulnerable to noisy labels and leading to unavoidable performance degradation.

To address the aforementioned issues, we propose a robust 2D-3D retrieval framework (RONO) to robustly learn from noisy multimodal data as shown in Figure 1. Our RONO framework consists of two mechanisms: 1) a novel Robust Discriminative Center Learning mechanism (RDCL) to robustly and discriminatively tackle clean and noisy samples, and 2) a Shared Space Consistency Learning mechanism (SSCL) to alleviate and even eliminate the heterogeneity and semantic gaps across different modalities.

More specifically, RDCL is presented to adaptively divide the noisy data into clean and noisy samples based on the memorization effect of deep neural networks (DNNs) [3], and then endowing them with positive and negative optimization directions, respectively. In brief, RDCL could compact the clean points to the corresponding category centers while scattering the noisy ones apart away from the noisy centers in the common space, thus alleviating the interference of noisy labels. In addition, our SSCL aims at mitigating the inherent gaps in the common space, i.e., the heterogeneity and semantic gaps. On the one hand, to bridge the heterogeneity gap across different modalities, our SSCL enforces modality-specific samples from the same instance collapse into a single point in the common space, thus producing modality-invariant representations. On the other hand, our SSCL narrows the gap between the representation space and shared label space to explicitly elimi-
Figure 1. The pipeline of our robust 2D-3D retrieval framework (RONO). First, modality-specific extractors project different modalities \( \{X^j, Y^j\}_{j=1}^M \) into a common space. Second, our Robust Discriminative Center Learning mechanism (RDCL) is conducted in the common space to divide the clean and noisy data while rectifying the optimization direction of noisy ones, leading to robustness against noisy labels. Finally, RONO employs a Shared Space Consistency Learning mechanism (SSCL) to bridge the intrinsic gaps between common space and label space. To be specific, SSCL narrows the cross-modal gap by a Multimodal Gap loss (MG) while minimizing the semantic discrepancy between the common space and label space using a Common Representations Classification loss (CRC) \( L_{crc} \), thus endowing representations with modality-invariant discrimination.

2. Related Work

2.1. Cross-modal Retrieval

In recent years, cross-modal retrieval has attracted more and more attention from academia and industry due to its flexible search \([17, 39, 40]\). The most popular solution is to project multimodal data into a common space, resulting in retrieving related content in the space across different modalities. They could be roughly classified into unsupervised and supervised methods. More specifically, 1) one typical kind of unsupervised method is Canonical Correlation Analysis (CCA) and its variants \([1, 35, 45]\). They aim to map multimodal inputs into a common space by maximizing the correlation across different modalities. 2) With the help of label information, supervised methods could encapsulate the discrimination into the shared space, thus suit- ing the downstream retrieval task better. To learn discriminative representations, some shallow methods utilize Fisher criterion \([6]\) to project different modalities into a latent common space \([20, 30]\). To capture the high nonlinearity in multimodal data, Deep Neural Networks (DNNs) are introduced to learn discriminative and modality-invariant representations \([28, 34, 44]\).

2.2. Learning with Noisy Labels

To tackle the ubiquitous imperfect annotations in the training data, numerous methods have been presented to im-
prove the robustness of DNNs against noisy labels. The existing works could be grouped into three main categories: 1) Architecture-based methods focus on modifying the DNN architecture to model the noise transition matrix [5, 8, 11]. However, it is still an open issue to accurately estimate noise transition due to the unpredictable and complex noise. 2) Samples-based methods generally learn from clean samples while weakening the adverse impact of noisy ones by re-weighting samples and refurbishing the labels [2, 25]. However, they require some additional well-labeled data, which is difficult and even impossible to satisfy in real-world applications. To address the problem, some methods adaptively divide the training data into clean and noisy sets based on the memorization effect of DNNs [3] for robust DNN training [13, 24]. 3) Loss-based methods mainly focus on designing robust optimization objectives to guide DNNs learning from noisy labels [10, 16, 26, 37, 43]. Although these robust loss functions could alleviate DNNs overfitting on noisy labels, they may lead DNNs to underfit the clean and hard samples.

3. The Proposed Method

3.1. Problem Formulation

First, some notations are defined for a clear presentation. Boldface lowercase \( x \) and plain letters \( M_N \) represent column vectors and scalars, respectively. Give a \( K \)-category multimodal dataset \( D = \{M_j\}_{j=1}^M = \{X_j, Y_j\}_{j=1}^M \), where \( M \) is the number of modalities, \( N \) is the number of samples in one modality, \( M_j = \{ (x_{ij}^j, y_{ij}^j) \}_{i=1}^N \) is the \( j \)-th modality, \( x_{ij}^j \in \mathbb{R}^{d_j} \) is the \( i \)-th sample from the \( j \)-th modality, \( d_j \) is the dimension of the \( j \)-th modality and \( y_{ij}^j \) is the corresponding label of \( x_{ij}^j \) that may be incorrect.

To alleviate the influence of corrupted labels, a robust 2D-3D retrieval framework (RONO) is proposed to robustly learn discriminative features from noisy labels while bridging the cross-modal gap. In the framework, we present a Robust Discriminative Center Learning mechanism (RDCL) to reduce the negative impact of unreliable labels, and a Shared Space Consistency Learning mechanism (SSCL) to simultaneously narrow the semantic and heterogeneity gaps. The overall objective function is shown as follows:

\[
\mathcal{L} = \mathcal{L}_{rdc} + \beta_{mg} \mathcal{L}_{mg} + \beta_{crc} \mathcal{L}_{crc},
\]

where \( \mathcal{L}_{rdc} \) is the loss function adopted by RDCL (see Equation (4)), \( \mathcal{L}_{mg} \) and \( \mathcal{L}_{crc} \) are the loss functions employed by SSCL (see Equations (5) and (6)), and \( \beta_{mg} \) and \( \beta_{crc} \) are the trade-off parameters. Our RONO could be optimized by descending Equation (1) with stochastic gradient descent. In the following sections, we will elaborate on each component of our RONO.

3.2. Robust Discriminative Center Learning

To encapsulate the discrimination into the common space, we enforce the samples with the same category compact to the shared clustering centers, while escaping from other centering centers. First, we formulate a contrastive center error \( t \) to measure the semantic difference between the estimated representations and the clustering centers as follows:

\[
p_i^j = \frac{1}{K-1} \sum_k^K e^{\langle c_{k,v} - z_i^j \rangle^T z_i^j} - e^{\langle c_{k,v} - z_i^j \rangle^T z_i^j}, \tag{2}
\]

where \( z_i^j = f_j(x_i^j) \in \mathbb{R}^{d_c} \) is the \( d_c \)-D common representation of the \( x_i^j \), \( f_j : \mathbb{R}^{d_j} \rightarrow \mathbb{R}^{d_c} \) is the input space to common space mapping function, \( c_i^j = \frac{1}{|Z_i|} \sum_{z \in Z_i} z \) is the center of \( i \)-th category, and \( Z_i \) is the set of the \( i \)-th category in the common space. In Equation (2), the dot product is exploited to measure the similarity between a given point and a clustering center in the common space. Obviously, minimizing Equation (2) could maximize the within-class similarities while minimizing the between-class ones, thus endowing the common representations with discrimination. Thus, we could formulate a vanilla loss of RDCL as below:

\[
\mathcal{L}_{rdc} = \frac{1}{MN} \sum_{i}^M \sum_{j}^N t_i^j, \tag{3}
\]

However, such a learning paradigm of Equation (3) will utilize all samples to train networks indiscriminately, thus leading to overfitting on corrupted labels like traditional supervision methods, especially under high noise rates.

To investigate the impact of noisy labels on \( \mathcal{L}_{rdc} \), we conduct some visualized experiments on noisy labels and true labels as shown in Figure 2. From the figure, one could observe that although \( \mathcal{L}_{rdc} \) makes the networks overfitting to corrupted labels (see Figures 2b and 2f), it could capture the correct discrimination to compact the samples with corrupted labels to the correct clusters in the early training stage (see Figures 2c and 2g), which is well known as the memorization effect of DNNs [3]. That is to say, the similarities between points and the assigned clusters could be markedly distinguishable after a short training period for the samples with correct and corrupt labels. Inspired by the observation, we propose a Robust Discriminative Center loss (RDC) to adaptively discriminate the clean and noisy samples according to \( t \), and then apply a bias to endow them with positive and negative optimization direction. Thus, we rewrite Equation (3) as follows:

\[
\mathcal{L}_{rdc} = \frac{1}{MN} \sum_{i}^M \sum_{j}^N \left[ (1-v) t_i^j - v \left| t_i^j + \alpha \right| \right], \tag{4}
\]

where \( v \in [0,1] \) is a dynamically increasing balanced parameter increasing from 0 to 1 with the number of epoch lin-
Figure 2. (a), (b), (c) and (d) show the density vs. the similarity between common representations and noisy centers, and (e), (f), (g) and (h) show the density vs. the similarity between common representations and clean centers for the test set after the training under 0.4 symmetric noise. Moreover, (a) and (e) are trained with the Cross-Modal Center loss (CMC) [19], (b) and (f) are trained with vanilla RDCL loss $L_{rdc}'$ of Equation (3), (c) and (g) are trained for only 5 epochs with $L_{rdc}'$ and (d) and (h) are trained with the proposed Robust Discriminative Center loss (RDC). It can be found that after short training periods with vanilla RDCL loss, both true and false labeled samples are compacted to their real category centers, and false labeled samples are not similar to their mislabeled category centers. Adopting RDC for training could maintain and reinforce the robustness of the results, which makes the samples in noisy set adaptively avoid being similar to the mislabeled category centers and maintain their similarity to the real category centers, while CMC could hardly achieve.

First, in order to further alleviate or even eliminate the inherent gap across different modalities, we adopt a Multi-modal Gap loss (MG) to maximize the mutual information between different modalities from the instance-based perspective [16], which could be formulated as:

$$L_{mg} = -\frac{1}{MN} \sum_i \sum_j \log \left( \frac{\sum_k e^{\frac{1}{\tau} (z^k_i)^T z_j^l}}{\sum_l \sum_k e^{\frac{1}{\tau} (z^l_k)^T z_j^l}} \right),$$  \hspace{1cm} (5)

where $\tau$ is a temperature parameter. By minimizing Equation (5), the discrepancy across different modalities could be reduced to project cross-modal samples into a common space, thus narrowing the heterogeneity gap.

Second, in addition to the heterogeneity gap, the semantic discrepancy will also degrade the performance since distinct modalities intrinsically belong to the same label space. To eliminate the discrepancy, we propose a Common Representations Classification loss (CRC) to narrow the gap between the common space and label space. Specifically, a shared classifier is employed to bridge the common space and label space, and then minimize the difference between classification predictions and labels. To alleviate the influence of noisy labels, the robust MAE loss is utilized to minimize the difference as follows:

$$L_{crc} = \frac{1}{MN} \sum_i \sum_j |g(z^l_i, \Gamma) - y^l_j|,$$  \hspace{1cm} (6)

### 3.3. Shared Space Consistency Learning

Although our RDCL could achieve robustness against noisy labels, the cross-modal learning paradigm is frequently affected by the inherent gaps across different modalities. More specifically, due to the randomness of the category centers caused by the random initialization of DNNs, blindly increasing the discrimination between points and centers would lead to losing the intrinsic information inside the data, thus degrading the retrieval performance. To handle the problem, we present a Shared Space Consistency Learning mechanism (SSCL) to capture the intrinsic information by narrowing the heterogeneity and semantic gaps simultaneously.

<table>
<thead>
<tr>
<th>Density</th>
<th>Clean similarity</th>
<th>Noisy similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>15</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>15</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>30</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>30</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>30</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>40</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>40</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>40</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>60</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>60</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>60</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
where $g(z^2, \Gamma)$ is a common classifier with parameters $\Gamma$ shared by all modalities.

### 3.4. Theoretical Justification

Following previous works [10, 26], we could indicate that our joint loss function $L$ is robust against both symmetric and asymmetric (or class conditional) noisy labels. Since DNNs are noise tolerant in the early training stage [3], we only discuss the robustness of our method during the latter training stage as $v$ has dynamically increased to 1.

**Property 1** \( \exists \alpha \in [e, e] \), RDC satisfies:

\[
\sum_{i=1}^{K} L_{rdc}(f(x), i) = (K-1) L_{rdc}(f(x), y^*) + C \quad x \in \mathcal{X}, \forall i,
\]

where $L_{rdc}(f(x), y^*)$ is the loss function to compute $L_{rdc}$ with the common representations $f(x)$ and $K$-category true label $y^*$, and $C$ is a constant.

**Property 2** There exist upper and lower definite boundaries for $L_{rdc}$.

**Lemma 1** If the above properties are satisfied, RDC with appropriate $\alpha$ is noise tolerant against symmetric (or uniform) and asymmetric (or class conditional) noisy labels.

**Lemma 2** [10] When symmetric noise rate $\eta < 1 - \frac{1}{K}$, MAE is robust against symmetric (or uniform) noisy labels. Defining the risk of the classifier is $R_{L_{mae}}(f)$, $f^*$ is the global minimizers of $R_{L_{mae}}$, $L_{mae}(f(x), y)$ is the calculation of MAE, when $R_{L_{mae}}(f^*) = 0, 0 \leq L_{mae}(f(x), i) \leq \frac{C}{1-\eta}$ and any category-wise noise rate is less than the rate of being clean $\eta_{ij} < 1 - \eta_y$, MAE is robust against asymmetric (or class conditional) noisy labels.

**Lemma 3** [21, 26] Assuming there are $n$ noise tolerant loss functions $\{L_i\}_{i=1}^{n}$, and $n$ trade-off hyperparameters $\{\gamma_i\}_{i=1}^{n}$, then $L = \sum_{i=1}^{n} \gamma_i L_i$ is noise tolerant.

The elaborate proofs for Property 1, Property 2, and Lemma 1 can be found in our Complementary Materials. Property 1 and Property 2 can be obtained by mathematical derivation. For Lemma 1, we denote the risk of representation extractor $f$ with clean data $\{x, y^*\}$ on our RDC as $R(f) = \mathbb{E}_{x,y^*} L_{rdc}$, and the risk with noisy labels as $R^n(f) = \mathbb{E}_{x,y} L_{rdc}$. If $f^*$ and $f^*_n$ are respectively the global minimizers of $R(f)$ and $R^n(f)$, we require to prove $f^*$ is also a global minimizer of noisy risk $R^n(f)$ for $L_{rdc}$ that is robust against noisy labels. For Lemma 2 and Lemma 3, the theoretical proofs are provided in [10] and [21, 26], respectively.

According to Lemma 1 and Lemma 2, our RDC and CRC are theoretically robust against both symmetric and asymmetric noisy labels. Moreover, our MG is also noise tolerant with noisy labels since it is unsupervised and independent of labels. Therefore, according to Lemma 3, we could draw the conclusion that our joint loss function $L$ is noise tolerant against noisy labels.

### 4. Experiments

To evaluate our RONO, we conduct extensive comparison experiments on four 3D multimodal datasets with different scales, i.e., 3D MNIST [38], RGB-D object [22], ModelNet10 [36] and ModelNet40 [36] datasets.

#### 4.1. Experimental Settings

In this work, our RONO is implemented in PyTorch and its optimization process could be found in our Complementary Materials. All the experiments are carried out on GeForce RTX 1080Ti GPUs. Four widely-used 3D multimodal datasets are utilized for evaluation, which are briefly introduced below:

- **3D MNIST** [38]: It is a small-scale 3D model dataset collected in Kaggle which contains 6000 image-point cloud pairs. The point cloud samples are generated from the MNIST dataset. We divide the dataset into 2 subsets: 5000 and 1000 pairs for training and testing sets, respectively.

- **RGB-D object** [22]: It is a large-scale dataset containing 300 common family objects belonging to 51 categories. The dataset has 207,621 image-point cloud pairs, which of each has a 640×480 image and a point-cloud object with 1000 to 5000 points. We split the dataset into 200,000 and 7,621 pairs for training and testing sets, respectively.

- **ModelNet10** [36]: It is a 10-categories 3D CAD object benchmark. We divide the dataset into 2 subsets: 2,468 and 908 for training and testing sets, respectively.

- **ModelNet40** [36]: It is a 40-categories 3D CAD object benchmark. We divide the dataset into 2 subsets: 9,840 and 3,991 for training and testing sets, respectively.

In our experiments, we compare our RONO with 15 state-of-the-art methods that include 5 unsupervised methods (i.e., CCA [15], DCCA [1], DCCAE [35], UCCH [17] and DGCPN [42]) and 10 supervised ones (i.e., GMA [30], MvDA [20], AGAH [12], DADH [4], DAGNN [28], ALGCN [27], DSCMR [44], MRL [16], CLF [19] and CLF [19]+MAE [10]). Note that, CLF+MAE is a variant of CLF [19] with a robust MAE [10].

Most of the experiments are conducted on bimodal settings to evaluate two cross-modal tasks: using images as queries to retrieve the point-cloud samples (Image $\rightarrow$ Point Cloud), using point-cloud samples as queries to retrieve the images (Point Cloud $\rightarrow$ Image). Without loss of generality, several experiments are conducted across three modalities (i.e., Image, Mesh, and Point cloud) on ModelNet10 and ModelNet40. We use the widely-used mean Average Precision (mAP) score to evaluate the retrieval performance. We
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 and RGB-D object datasets. The highest mAPs are shown in **bold** and the second highest mAPs are underlined.

<table>
<thead>
<tr>
<th>Method</th>
<th>3D MNIST [38]</th>
<th>RGB-D object [22]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image→Point Cloud</td>
<td>Point Cloud→Image</td>
</tr>
<tr>
<td></td>
<td>0.2 0.4 0.6 0.8</td>
<td>0.2 0.4 0.6 0.8</td>
</tr>
<tr>
<td>CCA [15]</td>
<td>0.625 0.625 0.625 0.625</td>
<td>0.532 0.532 0.532 0.532</td>
</tr>
<tr>
<td>DCCA [1]</td>
<td>0.684 0.684 0.684 0.684</td>
<td>0.584 0.584 0.584 0.584</td>
</tr>
<tr>
<td>DCCAE [35]</td>
<td>0.703 0.703 0.703 0.703</td>
<td>0.593 0.593 0.593 0.593</td>
</tr>
<tr>
<td>DGCNN [28]</td>
<td>0.765 0.765 0.765 0.765</td>
<td>0.705 0.705 0.705 0.705</td>
</tr>
<tr>
<td>UCC [17]</td>
<td>0.771 0.771 0.771 0.771</td>
<td>0.755 0.755 0.755 0.755</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.892 0.877 0.870 0.836</strong></td>
<td><strong>0.890 0.875 0.861 0.830</strong></td>
</tr>
</tbody>
</table>

Table 2. Performance comparison under the symmetric noise rates of 0.2, 0.4, 0.6, and 0.8 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></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image→Point Cloud</td>
<td>Point Cloud→Image</td>
</tr>
<tr>
<td></td>
<td>0.2 0.4 0.6 0.8</td>
<td>0.2 0.4 0.6 0.8</td>
</tr>
<tr>
<td>CCA [15]</td>
<td>0.650 0.617 0.585 0.521</td>
<td>0.525 0.485 0.466 0.448</td>
</tr>
<tr>
<td>DCCA [1]</td>
<td>0.684 0.684 0.684 0.684</td>
<td>0.584 0.584 0.584 0.584</td>
</tr>
<tr>
<td>DCCAE [35]</td>
<td>0.703 0.703 0.703 0.703</td>
<td>0.593 0.593 0.593 0.593</td>
</tr>
<tr>
<td>DGCNN [28]</td>
<td>0.765 0.765 0.765 0.765</td>
<td>0.705 0.705 0.705 0.705</td>
</tr>
<tr>
<td>UCC [17]</td>
<td>0.771 0.771 0.771 0.771</td>
<td>0.755 0.755 0.755 0.755</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.892 0.877 0.870 0.836</strong></td>
<td><strong>0.890 0.875 0.861 0.830</strong></td>
</tr>
</tbody>
</table>

4.2. Comparison with the State-of-the-Arts

We apply 2D-3D cross-modal retrieval on four 3D model multimodal datasets to evaluate the robustness of our RONO and the baselines. The experimental results under symmetric noise are reported in Tables 1 and 2, and ones under asymmetric noise are reported in Table 3. From these experimental results, we could obtain the following observations:

- Noisy labels could remarkably reduce the retrieval per-
formance of supervised methods. Their performance will be degraded rapidly as the noise rate increases. Especially, when the noise rate is high, supervised methods tend to perform worse than unsupervised ones, or even fail to fit the data.

- For the symmetric noise, our RONO achieves remarkably better results than supervised (e.g., CLF, DAGNN, DSCMR, etc.), which demonstrates the robustness of our method against noisy labels. Besides, our RONO could utilize noisy labels to overcome the unsupervised methods (e.g., UCCH, DGCNP, etc.), thus indicating that additional label information could improve performance even if it contains noise.

- Our RONO is superior to a strong baseline (i.e., MRL) that is a noise-tolerate cross-modal method. Especially, our method could achieve 0.703 in terms of mAP under 0.8 noise on the large-scale RGB-D object dataset, which is higher than MRL (0.493) by 0.210, thus demonstrating the effectiveness of our adaptive optimization strategy for clean and noisy data.

- For asymmetric noise, the extremely perplexing class conditional noise will degrade the performance of the memorization effect of DNNs, however, our RONO still achieves superior robustness against noisy labels.

- Our RONO shows superiority even without the addition of synthetic label noise, demonstrating that well-annotated datasets also contain noise impacting the performance of each non-robust method.

Table 3. Performance comparison under the asymmetric noise rates of 0, 0.1, 0.2, and 0.4 on the 3D MNIST and RGB-D object datasets. The highest mAPs are shown in **bold** and the second highest mAPs are *underlined*.

<table>
<thead>
<tr>
<th>Method</th>
<th>3D MNIST [38]</th>
<th>RGB-D object [22]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image→Point Cloud</td>
<td>Point Cloud→Image</td>
</tr>
<tr>
<td></td>
<td>0.01 0.2 0.4</td>
<td>0.01 0.2 0.4</td>
</tr>
<tr>
<td>CCA [15]</td>
<td>0.415 0.415 0.415</td>
<td>0.415 0.415 0.415</td>
</tr>
<tr>
<td>DCCA [1]</td>
<td>0.595 0.595 0.595</td>
<td>0.595 0.595 0.595</td>
</tr>
<tr>
<td>DCMAE [35]</td>
<td>0.600 0.600 0.600</td>
<td>0.600 0.600 0.600</td>
</tr>
<tr>
<td>DGCNP [42]</td>
<td>0.792 0.792 0.792</td>
<td>0.792 0.792 0.792</td>
</tr>
<tr>
<td>UCCH [17]</td>
<td>0.791 0.791 0.791</td>
<td>0.791 0.791 0.791</td>
</tr>
<tr>
<td>GMA [30]</td>
<td>0.514 0.444 0.436</td>
<td>0.514 0.444 0.436</td>
</tr>
<tr>
<td>MvDA [20]</td>
<td>0.530 0.472 0.407</td>
<td>0.508 0.472 0.397</td>
</tr>
<tr>
<td>AGAH [12]</td>
<td>0.967 0.730 0.611</td>
<td>0.961 0.729 0.589</td>
</tr>
<tr>
<td>DADH [4]</td>
<td>0.971 0.848 0.718</td>
<td>0.969 0.825 0.701</td>
</tr>
<tr>
<td>DAGNN [28]</td>
<td>0.927 0.894 0.871</td>
<td>0.927 0.893 0.864</td>
</tr>
<tr>
<td>ALGCN [27]</td>
<td>0.908 0.876 0.860</td>
<td>0.900 0.871 0.852</td>
</tr>
<tr>
<td>DSCMR [44]</td>
<td>0.963 0.914 0.869</td>
<td>0.945 0.906 0.862</td>
</tr>
<tr>
<td>MRL [16]</td>
<td>0.963 0.959 0.944</td>
<td>0.945 0.940 0.922</td>
</tr>
<tr>
<td>CLF [19]</td>
<td>0.983 0.945 0.924</td>
<td>0.958 0.932 0.920</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>0.983</strong> 0.961 0.958</td>
<td><strong>0.983</strong> 0.947 0.938</td>
</tr>
</tbody>
</table>

Table 4. Ablation studies for RONO on the 3D MNIST and ModelNet40 datasets with 0.4 symmetric noise. ✓ stands for use.

RDCL | SSCL | 3D MNIST [38] | ModelNet40 [36] |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_{rdc}$</td>
<td>$L_{mg}$</td>
<td>$L_{crc}$</td>
<td>Img→Pnt 0.4</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>0.952 0.831</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>0.891 0.342</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>0.841 0.675</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>0.929 0.615</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>0.930 0.709</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>0.641 0.423</td>
</tr>
<tr>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>✓ ✓</td>
<td>0.527 0.348</td>
</tr>
</tbody>
</table>

4.3 Ablation Study

In this section, we investigate the contribution of each proposed component (i.e., loss $L_{rdc}$, $L_{mg}$ and $L_{crc}$) to 2D-3D retrieval with noisy labels. For a comprehensive comparison, we ablate each component from the framework and conduct the variants with the same experimental settings on two distinct datasets (i.e., 3D MNIST and ModelNet40). The results are shown in Table 4. From the table, one could draw the following observation: 1) RONO with/without any component will improve/drop retrieval performance, which indicates that each component contributes to our framework. 2) Replacing RCD (i.e., $L_{rdc}$) with the vanilla loss of RDCL (i.e., $L_{rdc}'$) will result in remarkable performance degradation, thus demonstrating the effectiveness of differentiated optimization for clean and noisy data, especially for high noise rates.
Table 5. Performance comparison of CLF [19] and our RONO under the symmetric noise rates of 0, 0.2, 0.4, 0.6 and 0.8 on tri-modal ModelNet40 dataset [36]. The highest mAPs are shown in bold. For a convenience presentation, we abbreviate Image, Mesh, Point cloud, Query, and Retrieval to Img, Msh, Pnt, Qry, and Retrv, respectively.

4.4. Further Comparison with CLF

To comprehensively evaluate our RONO, we conduct various 2D-3D cross-modal retrieval and visualization experiments across three modalities (i.e., Image, Mesh, and Point cloud) on ModelNet10 and ModelNet40 [36], by comparing RONO with state-of-the-art CLF [19]. The comparison results are shown in Table 5 and Figures 3 and 4. Due to space limitation, more cross-modal and in-domain comparison results could be found in our Complementary Materials.

From the experimental results, one could observe that: 1) Regardless of noise rates, our RONO show stronger robustness against noisy labels across three modalities. 2) Our RONO could obtain more discriminative clusters compared to CLR, which demonstrates the robustness of our RONO. 3) Our RONO could achieve more correct retrieved results while CLF fails on the same queries, indicating that our method has stronger robustness and is consistent with our quantitative results.

5. Conclusion

In this paper, we propose a novel 2D-3D cross-modal retrieval framework to robustly learn discriminative and modality-invariant representations with noisy labels, termed RONO. To be specific, our RONO employs a novel Robust Discriminative Center Learning mechanism (RDCL) to endow clean and noisy samples with correct optimization directions, while a Shared Space Consistency Learning mechanism (SSCL) to guarantee the cross-modal and semantic consistency across different modalities in the common space. Comprehensive mathematical analyses are provided to theoretically prove the noise tolerance of our RONO. Furthermore, we conduct extensive experiments compared to 15 state-of-the-art methods on four 3D model multimodal datasets to demonstrate the robustness of the proposed method against synthetic and real label noise.

Acknowledgments

This work is supported by the National Key R&D Program of China under Grant 2020YFB1406702, the National Natural Science Foundation of China (Grants No. 62102274, 62176171, and U19A2078), Sichuan Science and Technology Planning Project (Grants No. 2021YFS0389, 2021YFG0317, 2021YFG0301, 2022YFQ0014, and 2022YFH0021), Fundamental Research Funds for the Central Universities, A*STAR AME Programmatic Funding A18A2b0046, RobotHTPO Seed Fund under Project C211518008 and EDB Space Technology Development Grant under Project S22-19016-STD.
References


Yuncheng Li, Jianchao Yang, Yale Song, Hal Daume, and Steven CH Hoi. Dividemix: A deep representation for volumetric shapes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1912–1920, 2015. 5, 6, 7, 8


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. 5, 6, 7


Xu Yang, Cheng Deng, Tongliang Liu, and Dacheng Tao. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020. 1


