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 (I[2]R), A*STAR, Singapore
[3]Sichuan Zhiqian Technology Co., Ltd [4]Chengdu 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, thus mitigating the negative impact of noisy labels. 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-

*Corresponding author: Peng Hu (penghu.ml@gmail.com).

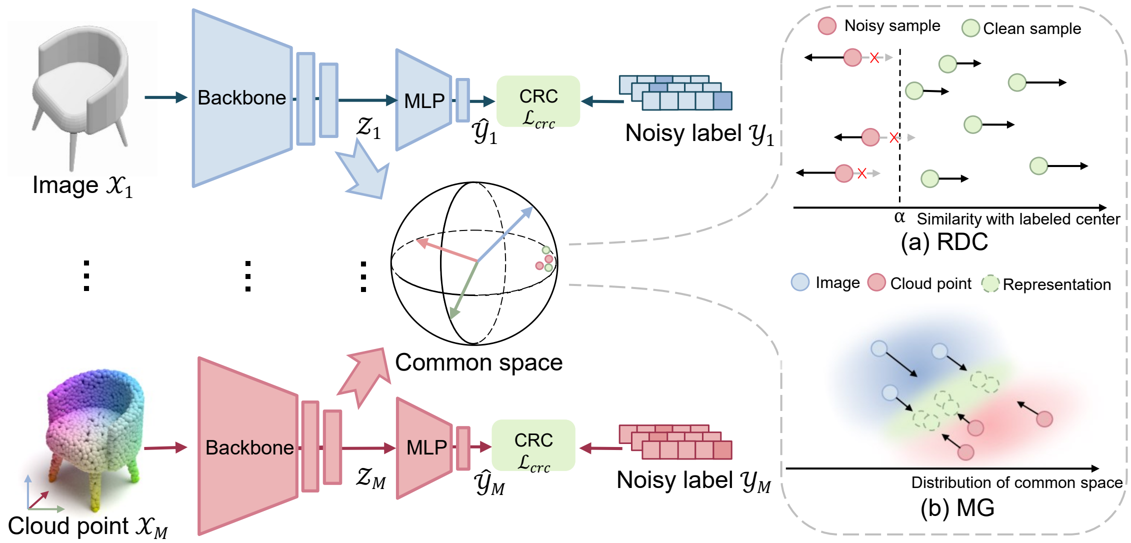


Figure 1. The pipeline of our robust 2D-3D retrieval framework (RONO). First, modality-specific extractors project different modalities $\{\mathcal{X}_j, \mathcal{Y}_j\}_{j=1}^{M}$ into a common space. Second, our Robust Discriminative Center Learning mechanism (RDCL) is conducted in the common space 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) $\mathcal{L}_{crc}$, thus endowing representations with modality-invariant discrimination.

nate the semantic discrepancy, thus encapsulating common discrimination into common representations. Our main contributions can be summarized as follows:

- 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 the best of our knowledge, this work could be one of the first attempts to learn with noisy labels for 2D-3D cross-modal retrieval.
- To mitigate the impact of noisy labels, a novel Robust Discriminative Center Learning mechanism (RDCL) is proposed to adaptively distinguish clean and noisy samples, and then provide them with positive and negative optimization directions, respectively.
- To construct discriminative and modality-invariant representations, a Shared Space Consistency Learning mechanism (SSCL) is presented to alleviate the intrinsic gaps across heterogeneity, representation, and label spaces.
- We theoretically and experimentally demonstrate the robustness of the proposed method under both synthetic symmetric/asymmetric and real-world noisy labels. Our RONO remarkably outperforms the state-of-the-art methods on 3D object benchmarks of different scales with noisy labels without bells and whistles.

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 suiting 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 improve 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 $\boldsymbol{x}$ and plain letters M, N represent column vectors and scalars, respectively. Give a K-category multimodal dataset $\mathcal{D} = \{\mathcal{M}_j\}_{j=1}^M = \{\mathcal{X}_j, \mathcal{Y}_j\}_{j=1}^M$, where M is the number of modalities, N is the number of samples in one modality, $\mathcal{M}_j = \{(\boldsymbol{x}_i^j, y_i^j)\}_{i=1}^N$ is the j-th modality, $\boldsymbol{x}_i^j \in \mathbb{R}^{d_j}$ is the i-th sample from the j-th modality, d_j is the dimension of the j-th modality and y_i^j is the corresponding label of $\boldsymbol{x}_i^j$ that may be incorrect.

To alleviate the influence of corrupted labels, a robust 2D-3D retrieval framework (RONO) is proposed to robustly learn discrimination 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:

$$\mathcal{L} = \underbrace{\mathcal{L}_{rdc}}_{RDCL} + \underbrace{\beta_{mg}\mathcal{L}_{mg} + \beta_{crc}\mathcal{L}_{crc}}_{SSCL}, \tag{1}$$

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 β_{mg} and β_{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:

$$t_i^j = \frac{\sum_k^K e^{\left(\boldsymbol{c}_{k \neq y_i^j}\right)^T \boldsymbol{z}_i^j}}{K-1} - e^{\left(\boldsymbol{c}_{k = y_i^j}\right)^T \boldsymbol{z}_i^j}, \tag{2}$$

where $\boldsymbol{z}_i^j = f_j(\boldsymbol{x}_i^j) \in \mathbb{R}^{d_c}$ is the d_c-D common representation of the $\boldsymbol{x}_i^j$, $f_j : \mathbb{R}^{d_j} \to \mathbb{R}^{d_c}$ is the input space to common space mapping function, $\boldsymbol{c}_i = \frac{1}{|\mathcal{Z}_i|}\sum_{\boldsymbol{z} \in \mathcal{Z}_i} \boldsymbol{z}$ is the center of i-th category, and $\mathcal{Z}_i$ 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:

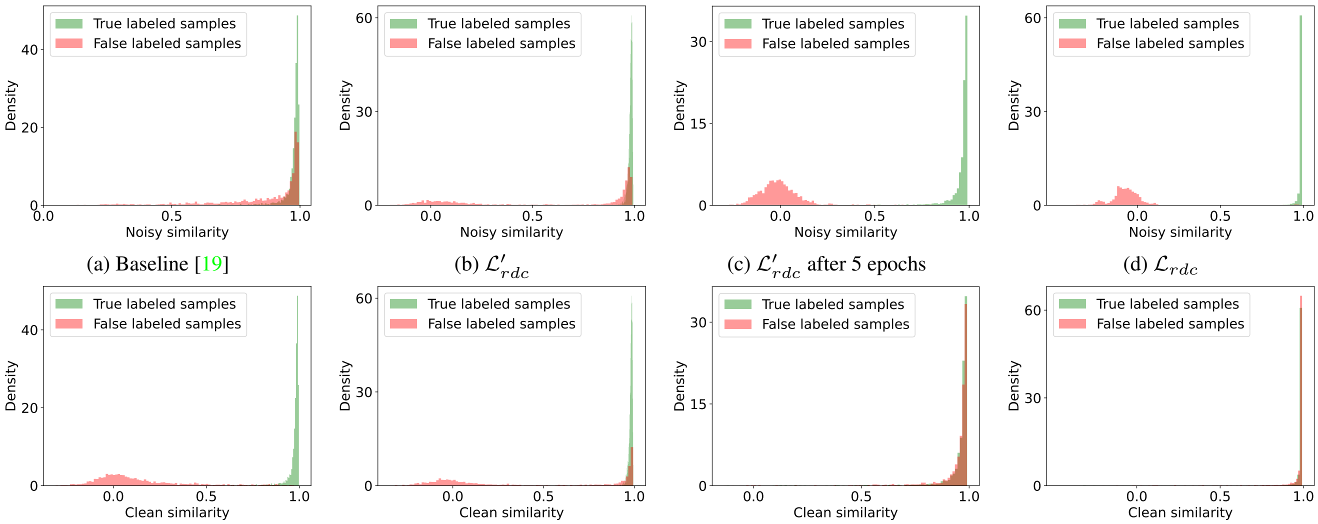
$$\mathcal{L}'_{rdc} = \frac{1}{MN}\sum_i^N \sum_j^M t_i^j, \tag{3}$$

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:

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

where $v \in [0, 1]$ is a dynamically increasing balanced parameter increasing from 0 to 1 with the number of epoch lin-



(a) Baseline [19] (b) $\mathcal{L}'_{rdc}$ (c) $\mathcal{L}'_{rdc}$ after 5 epochs (d) $\mathcal{L}_{rdc}$

(e) Baseline [19] (f) $\mathcal{L}'_{rdc}$ (g) $\mathcal{L}'_{rdc}$ after 5 epochs (h) $\mathcal{L}_{rdc}$

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 $\mathcal{L}'_{rdc}$ of Equation (3), (c) and (g) are trained for only 5 epochs with $\mathcal{L}'_{rdc}$ 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.

early within up to 30 epochs, and $\alpha \in [-e, e]$ is a threshold for clean-noise separation. In our training stage, v gradually increases from 0 to 1. More specifically, our RDC will focus on learning simple patterns before fitting on the corrupted labels in the early stage. With further training, our RDC will make the second item dominant to explicitly separate samples into clean and noisy sets, and assign them with correct optimization directions. Surprisingly, our RDC could accurately cluster noisy samples into their correct centers regardless of the corrupted supervision as shown in the visualization results of Figures 2d and 2h, thus embracing robustness against noisy labels.

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.

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:

$$\mathcal{L}_{mg} = -\frac{1}{MN}\sum_{i}^{N}\sum_{j}^{M} log\left(\frac{\sum_{k}^{M} e^{\frac{1}{\tau}(\boldsymbol{z}_i^k)^T \boldsymbol{z}_i^j}}{\sum_{l}^{N}\sum_{m}^{M} e^{\frac{1}{\tau}(\boldsymbol{z}_l^m)^T \boldsymbol{z}_i^j}}\right), \quad (5)$$

where τ 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:

$$\mathcal{L}_{crc} = \frac{1}{MN}\sum_{i}^{N}\sum_{j}^{M} |g(\boldsymbol{z}_i^j, \Gamma) - y_i^j|, \quad (6)$$

where $g(\boldsymbol{z}_i^j, \Gamma)$ is a common classifier with parameters Γ shared by all modalities.

3.4. Theoretical Justification

Following previous works [10, 26], we could indicate that our joint loss function $\mathcal{L}$ is robust against both symmetric and asymmetric label noise. 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:*

$$\sum_{i=1}^{K} \mathcal{L}_{rdc}(f(\boldsymbol{x}), i) = (K-1)\mathcal{L}_{rdc}(f(\boldsymbol{x}), y^*) + C \quad \boldsymbol{x} \in \mathcal{X}, \forall f, \tag{7}$$

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

Property 2 *There exist upper and lower definite boundaries for $\mathcal{L}_{rdc}$.*

Lemma 1 *If the above properties are satisfied, RDC with appropriate α 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 $R_{\mathcal{L}_{mae}}(f)$, f^* is the global minimizers of $R_{\mathcal{L}_{mae}}(f)$, $\mathcal{L}_{mae}(f(\boldsymbol{x}), y)$ is the calculation of MAE, when $R_{\mathcal{L}_{mae}}(f^*) = 0$, $0 \leq \mathcal{L}_{mae}(f(\boldsymbol{x}), i) \leq \frac{C}{K-1} = 2, \forall i$ and any category-wide noise rate is less than the rate of being clean $\eta_{ij} \leq 1 - \eta_y$, MAE is robust against asymmetric (or class conditional) noisy labels.*

Lemma 3 *[21, 26] Assuming there are n noise tolerant loss functions $\{\mathcal{L}_i\}_{i=1}^n$, and n trade-off hyperparameters $\{\gamma_i\}_{i=1}^n$, then $\mathcal{L} = \sum_{i=1}^n \gamma_i \mathcal{L}_i$ 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 $\{\boldsymbol{x}, y^*\}$ on our RDC as $R(f) = \mathbb{E}_{\boldsymbol{x}, y^*}\mathcal{L}_{rdc}$, and the risk with noisy labels as $R^\eta(f) = \mathbb{E}_{\boldsymbol{x}, y}\mathcal{L}_{rdc}$. If f^* and f_η^* are respectively the global minimizers of $R(f)$ and $R^\eta(f)$, we require to prove f^* is also a global minimizer of noisy risk $R^\eta(f)$ for $\mathcal{L}_{rdc}$ 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 $\mathcal{L}$ 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 → Point Cloud), using point-cloud samples as queries to retrieve the images (Point Cloud → 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

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**

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.

Method	ModelNet10 [36]								ModelNet40 [36]							
	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]	0.625	0.625	0.625	0.625	0.627	0.627	0.627	0.627	0.532	0.532	0.532	0.532	0.531	0.531	0.531	0.531
DCCA [1]	0.684	0.684	0.684	0.684	0.677	0.677	0.677	0.677	0.584	0.584	0.584	0.584	0.569	0.569	0.569	0.569
DCCAE [35]	0.703	0.703	0.703	0.703	0.693	0.693	0.693	0.693	0.593	0.593	0.593	0.593	0.572	0.572	0.572	0.572
DGCPN [42]	0.765	0.765	0.765	0.765	0.759	0.759	0.759	0.759	0.705	0.705	0.705	0.705	0.697	0.697	0.697	0.697
UCCH [17]	0.771	0.771	0.771	0.771	0.770	0.770	0.770	0.770	0.755	0.755	0.755	0.755	0.739	0.739	0.739	0.739
GMA [30]	0.650	0.617	0.585	0.521	0.646	0.579	0.550	0.497	0.525	0.485	0.466	0.448	0.515	0.487	0.459	0.437
MvDA [20]	0.494	0.458	0.449	0.390	0.464	0.428	0.406	0.328	0.420	0.373	0.329	0.316	0.412	0.370	0.300	0.271
AGAH [12]	0.853	0.736	0.583	0.425	0.837	0.699	0.549	0.408	0.809	0.732	0.687	0.568	0.783	0.736	0.664	0.554
DADH [4]	0.860	0.772	0.670	0.554	0.838	0.768	0.658	0.553	0.818	0.743	0.676	0.581	0.782	0.748	0.657	0.586
DAGNN [28]	0.844	0.800	0.754	0.422	0.836	0.810	0.763	0.448	0.802	0.723	0.635	0.402	0.798	0.728	0.643	0.412
ALGCN [27]	0.788	0.597	0.426	0.282	0.797	0.589	0.440	0.269	0.766	0.538	0.426	0.298	0.763	0.537	0.403	0.279
DSCMR [44]	0.849	0.758	0.666	0.324	0.836	0.732	0.637	0.307	0.824	0.788	0.687	0.328	0.811	0.785	0.694	0.339
MRL [16]	0.876	0.870	0.863	0.832	0.861	0.857	0.848	0.823	0.833	0.829	0.828	0.818	0.824	0.826	0.820	0.817
CLF [19]	0.849	0.782	0.620	0.365	0.838	0.764	0.595	0.387	0.822	0.778	0.624	0.315	0.815	0.771	0.587	0.295
CLF [19]+MAE[10]	0.853	0.752	0.679	0.343	0.838	0.716	0.659	0.373	0.827	0.758	0.651	0.384	0.816	0.749	0.640	0.372
Ours	**0.892**	**0.877**	**0.870**	**0.836**	**0.890**	**0.875**	**0.861**	**0.830**	**0.877**	**0.858**	**0.838**	**0.823**	**0.872**	**0.854**	**0.838**	**0.821**

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.

conduct the comparison experiments under both synthetic symmetric and asymmetric label noise for comprehensive evaluation. Due to the space limitation, we only evaluate the robustness against asymmetric label noise on 3D MNIST and RGB-D object datasets. More extra experimental results could be found in our Complementary Materials. For a comprehensive evaluation, the symmetric noise rates are set as 0.2, 0.4, 0.6, and 0.8, and the asymmetric ones are set as 0.1, 0.2, and 0.4.

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-

Method	3D MNIST [38]								RGB-D object [22]							
	Image→Point Cloud				Point Cloud→Image				Image→Point Cloud				Point Cloud→Image			
	0	0.1	0.2	0.4	0	0.1	0.2	0.4	0	0.1	0.2	0.4	0	0.1	0.2	0.4
CCA [15]	0.415	0.415	0.415	0.415	0.415	0.415	0.415	0.415	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]	0.514	0.444	0.436	0.415	0.500	0.435	0.417	0.396	0.126	0.089	0.085	0.073	0.121	0.085	0.081	0.069
MvDA [20]	0.530	0.472	0.407	0.370	0.508	0.472	0.397	0.352	0.188	0.188	0.159	0.142	0.199	0.168	0.139	0.124
AGAH [12]	0.967	0.730	0.611	0.519	0.961	0.729	0.589	0.512	0.652	0.603	0.444	0.356	0.628	0.559	0.436	0.361
DADH [4]	0.971	0.848	0.718	0.570	**0.969**	0.825	0.701	0.572	0.772	0.723	0.589	0.524	0.761	0.703	0.572	0.511
DAGNN [28]	0.927	0.894	0.871	0.684	0.927	0.893	0.864	0.691	0.741	0.704	0.646	0.563	0.724	0.689	0.631	0.554
ALGCN [27]	0.908	0.876	0.860	0.635	0.900	0.871	0.852	0.641	0.717	0.685	0.617	0.526	0.691	0.678	0.598	0.531
DSCMR [44]	0.963	0.914	0.869	0.711	0.945	0.906	0.862	0.704	0.774	0.755	0.711	0.673	0.768	0.738	0.700	0.659
MRL [16]	0.963	0.959	0.944	0.792	0.945	0.940	0.922	0.762	0.723	0.655	0.635	0.602	0.719	0.652	0.636	0.599
CLF [19]	**0.983**	0.945	0.924	0.809	0.958	0.932	0.920	0.802	0.772	0.771	0.734	0.674	0.766	0.759	0.721	0.659
CLF [19]+MAE[10]	0.971	0.942	0.921	0.796	0.951	0.930	0.918	0.783	0.752	0.758	0.720	0.651	0.741	0.750	0.712	0.637
Ours	**0.983**	**0.961**	**0.958**	**0.912**	0.968	**0.947**	**0.938**	**0.897**	**0.779**	**0.773**	**0.728**	**0.681**	**0.771**	**0.756**	**0.721**	**0.669**

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.

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, DGCPN, 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.

RDCL	SSCL		3D MNIST [38]				ModelNet40 [36]			
$\mathcal{L}_{rdc}$	$\mathcal{L}_{mg}$	$\mathcal{L}_{crc}$	Img→ Pnt 0.4	0.8	Pnt→ Img 0.4	0.8	Img→Pnt 0.4	0.8	Pnt→Img 0.4	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

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

4.3. Ablation Study

In this section, we investigate the contribution of each proposed component (i.e., loss $\mathcal{L}_{rdc}$, $\mathcal{L}_{mg}$ and $\mathcal{L}_{crc}$) 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., $\mathcal{L}_{rdc}$) with the vanilla loss of RDCL (i.e., $\mathcal{L}'_{rdc}$) will result in remarkable performance degradation, thus demonstrating the effectiveness of differentiated optimization for clean and noisy data, especially for high noise rates.

η	Qry	Img			Msh			Pnt		
	Retrv	Img	Msh	Pnt	Img	Msh	Pnt	Img	Msh	Pnt
0	CLF	0.903	0.898	0.883	0.891	0.873	0.882	0.887	0.883	0.881
	Ours	**0.911**	**0.901**	**0.891**	**0.899**	**0.901**	**0.883**	**0.891**	**0.894**	**0.891**
0.2	CLF	0.825	0.830	0.825	0.831	0.850	0.848	0.829	0.850	0.851
	Ours	**0.874**	**0.872**	**0.881**	**0.883**	**0.891**	**0.889**	**0.875**	**0.884**	**0.890**
0.4	CLF	0.683	0.749	0.761	0.737	0.816	0.829	0.746	0.828	0.848
	Ours	**0.858**	**0.876**	**0.863**	**0.862**	**0.881**	**0.875**	**0.859**	**0.875**	**0.875**
0.6	CLF	0.560	0.589	0.627	0.569	0.681	0.729	0.604	0.725	0.794
	Ours	**0.842**	**0.853**	**0.851**	**0.857**	**0.857**	**0.862**	**0.843**	**0.868**	**0.872**
0.8	CLF	0.311	0.316	0.361	0.323	0.392	0.400	0.365	0.385	0.559
	Ours	**0.828**	**0.842**	**0.842**	**0.842**	**0.868**	**0.866**	**0.841**	**0.864**	**0.868**

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.

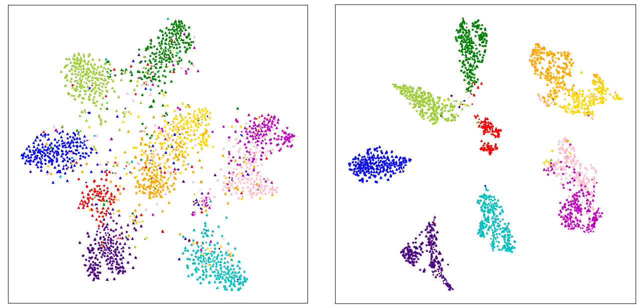
4.4. Further Comparison with CLF

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.



(a) CLF (b) Ours

Figure 3. The representation visualization on the testing set of ModelNet40 by using t-SNE method [33]. 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.

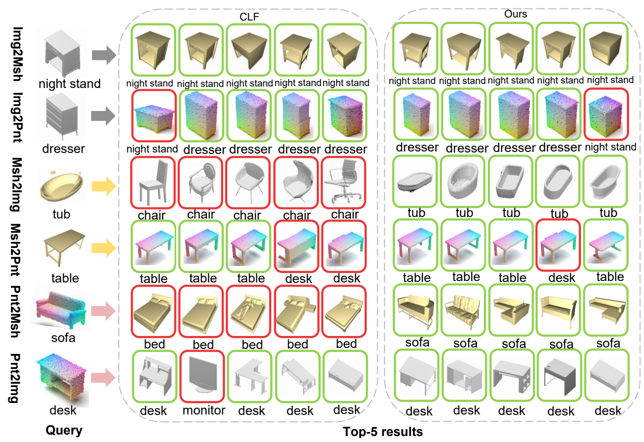


Figure 4. Top-5 retrieved results of CLF and our RONO under 0.4 label noise on the tri-modal ModelNet40 dataset. Green boxes indicate correct retrieval, while red boxes indicate wrong retrieval.

Furthermore, we conduct extensive experiments compared to 15 state-of-the-art methods on four 3D model multi-modal 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-STDP.

References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013. 2, 5, 6, 7

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*, pages 312–321. PMLR, 2019. 3

[3] 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, 5

[4] 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. 5, 6, 7

[5] Alan Joseph Bekker and Jacob Goldberger. Training deep neural-networks based on unreliable labels. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2682–2686. IEEE, 2016. 3

[6] Peter N. Belhumeur, Joao P Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on pattern analysis and machine intelligence*, 19(7):711–720, 1997. 2

[7] Paolo Bellandi, Franco Docchio, and Giovanna Sansoni. Roboscan: a combined 2d and 3d vision system for improved speed and flexibility in pick-and-place operation. *The International Journal of Advanced Manufacturing Technology*, 69 (5):1873–1886, 2013. 1

[8] Xinlei Chen and Abhinav Gupta. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1431–1439, 2015. 3

[9] Belen Jiménez Fernández-Palacios, Daniele Morabito, and Fabio Remondino. Access to complex reality-based 3d models using virtual reality solutions. *Journal of cultural heritage*, 23:40–48, 2017. 1

[10] 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. 3, 5, 6, 7

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[12] 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. 5, 6, 7

[13] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018. 3

[14] 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. 1

[15] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992. 5, 6, 7

[16] 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, 5, 6, 7

[17] 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. 1, 2, 5, 6, 7

[18] Muhammad Ibrahim, Naveed Akhtar, Michael Wise, and Ajmal Mian. Annotation tool and urban dataset for 3d point cloud semantic segmentation. *IEEE Access*, 9:35984–35996, 2021. doi: 10.1109/ACCESS.2021.3062547. 1

[19] 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. 1, 4, 5, 6, 7, 8

[20] 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. 1, 2, 5, 6, 7

[21] Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 101–110, 2019. 5

[22] 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. 5, 6, 7

[23] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 1

[24] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 3

[25] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017. 3

[26] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pages 6543–6553. PMLR, 2020. 3, 5

[27] 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. 5, 6, 7

[28] 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. 2, 5, 6, 7

[29] Yu Qian, Xiaohong Gao, Martin Loomes, Richard Comley, Balbir Barn, Rui Hui, and Zenmin Tian. Content-based retrieval of 3d medical images. In *The Third International Conference on eHealth, Telemedicine, and Social Medicine (eTELEMED 2011)*, pages 7–12, 2011. 1

[30] 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. 2, 5, 6, 7

[31] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1

[32] Guofeng Tong, Ran Liu, and Jindong Tan. 3d information retrieval in mobile robot vision based on spherical compound eye. In *2011 IEEE International Conference on Robotics and Biomimetics*, pages 1895–1900, 2011. doi: 10.1109/ROBIO.2011.6181567. 1

[33] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 8

[34] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 154–162, 2017. 2

[35] 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. 2, 5, 6, 7

[36] 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. 5, 6, 7, 8

[37] Tianyuan Xu, Xueliang Liu, Zhen Huang, Dan Guo, Richang Hong, and Meng Wang. Early-learning regularized contrastive learning for cross-modal retrieval with noisy labels. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 629–637, 2022. 3

[38] 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

[39] Erkun Yang, Mingxia Liu, Dongren Yao, Bing Cao, Chunfeng Lian, Pew-Thian Yap, and Dinggang Shen. Deep bayesian hashing with center prior for multi-modal neuroimage retrieval. *IEEE transactions on medical imaging*, 40(2): 503–513, 2020. 2

[40] Erkun Yang, Dongren Yao, Tongliang Liu, and Cheng Deng. Mutual quantization for cross-modal search with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7551–7560, 2022. 2

[41] 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

[42] 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. 5, 6, 7

[43] 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. 3

[44] 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. 1, 2, 5, 6, 7

[45] Wenming Zheng, Xiaoyan Zhou, Cairong Zou, and Li Zhao. Facial expression recognition using kernel canonical correlation analysis (kcca). *IEEE transactions on neural networks*, 17(1):233–238, 2006. 2