RoMo: Robust Unsupervised Multimodal Learning With Noisy Pseudo Labels

Yongxiang Li, Yang Qin, *Graduate Student Member, IEEE*, Yuan Sun,
Dezhong Peng, *Senior Member, IEEE*, Xi Peng, *Senior Member, IEEE*,
and Peng Hu

***Abstract*— The rise of the metaverse and the increasing volume of heterogeneous 2D and 3D data have created a growing demand for cross-modal retrieval, enabling users to query semantically relevant data across different modalities. Existing methods heavily rely on class labels to bridge semantic correlations; however, collecting large-scale, well-labeled data is expensive and often impractical, making unsupervised learning more attractive and feasible. Nonetheless, unsupervised cross-modal learning faces challenges in bridging semantic correlations due to the lack of label information, leading to unreliable discrimination. In this paper, we reveal and study a novel problem: unsupervised cross-modal learning with noisy pseudo-labels. To address this issue, we propose a 2D-3D unsupervised multimodal learning framework that leverages multimodal data. Our framework consists of three key components: 1) Self-matching Supervision Mechanism (SSM) warms up the model to encapsulate discrimination into the representations in a self-supervised learning manner. 2) Robust Discriminative Learning (RDL) further mines the discrimination from the learned imperfect predictions after warming up. To tackle the noise in the predicted pseudo labels, RDL leverages a novel Robust Concentrating Learning Loss (RCLL) to alleviate the influence of the uncertain samples, thus embracing robustness against noisy pseudo labels. 3) Modality-invariance Learning Mechanism (MLM) minimizes the cross-modal discrepancy to enforce SSM and RDL to produce common representations. We conduct comprehensive experiments on four 2D-3D multimodal datasets, comparing our method against 14 state-of-the-art approaches, thereby demonstrating its effectiveness and superiority.**

***Index Terms*— Unsupervised cross-modal retrieval, self-matching supervision, robust discriminant learning.**

Manuscript received 27 October 2023; revised 4 April 2024; accepted 25 June 2024. Date of publication 27 August 2024; date of current version 16 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant U21B2040, Grant 62176171, Grant 62472295, Grant 62102274, and Grant 62372315; in part by the Sichuan Science and Technology Program Grant 2024NSFTD0038, Grant 2024NSFTD0047, Grant 2024NSFTD0049, and Grant 2023ZYD0143; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303 and Grant CJ202403; and in part by the SCU-Luzhou Science and Technology Cooperation Program under Grant 2023CDLZ-16. The associate editor coordinating the review of this article and approving it for publication was Dr. Zhao Zhang. *(Corresponding author: Peng Hu.)*

Yongxiang Li, Yang Qin, Yuan Sun, and Peng Hu are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: rhythmli.scu@gmail.com; penghu.ml@gmail.com).

Dezhong Peng is with the College of Computer Science, Sichuan University, Chengdu 610065, China, and also with Sichuan Newstrong UHD Video Technology Company Ltd., Chengdu 610095, China.

Xi Peng is with the College of Computer Science, Sichuan University, Chengdu 610065, China, and also with the State Key Laboratory of Hydraulics and Mountain River Engineering, Chengdu 610065, China.

This article has supplementary downloadable material available at https://doi.org/10.1109/TIP.2024.3426482, provided by the authors.

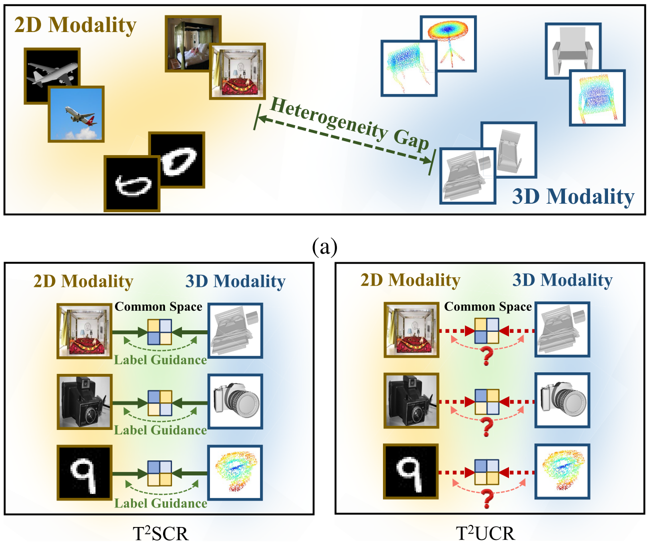
Digital Object Identifier 10.1109/TIP.2024.3426482

I. Introduction

With the rapid development of multimedia technology, cross-modal retrieval of 2D images (e.g., visible images, infrared images, sketch, etc.) and 3D data (e.g., point clouds, voxels, meshes, etc.) has increasingly become a crucial task for many multimedia applications, such as aviation, robotics, autonomous driving, metaverse, etc. [1] and [2]. For example, in visual localization for autonomous driving, matching a 2D image with various 3D models can improve the environmental perception and decision-making of unmanned ground vehicles [3], [4]. Hence, it is essential to develop effective and accurate 2D-3D cross-modal retrieval methods. However, this task faces a major challenge due to the significant discrepancy between 2D and 3D data, termed heterogeneity gap [5].

To address this challenge, numerous cross-modal methods have been proposed to learn a shared common space for different modalities under the guidance of label information [5], [6], [7], [8], which is also known as 2D-3D Supervised Cross-modal Retrieval (T^{2}SCR). These methods map each modality into one common discriminative space where samples with similar semantics across different modalities are close to each other while dissimilar samples are far from each other [9]. However, their performance heavily depends on a considerable amount of well-labeled data, which is often expensive and time-consuming to accurately and timely obtain in practice [10], [11]. This is especially true for 3D data that lack distinctive visual attributes (e.g., color and texture), such as 3D point clouds. Unfortunately, the annotation requirements for multiple modalities are more severe than for single-modal data. In contrast, unsupervised cross-modal retrieval methods could naturally avoid this problem, because they do not rely on labeled data while mining the inherent semantic correlations in the data [12]. These methods offer a potential solution for large-scale unlabeled 2D-3D cross-modal data retrieval. However, to the best of our knowledge, 2D-3D Unsupervised Cross-modal Retrieval (T^{2}UCR) is still a largely unexplored area. Furthermore, the lack of useful supervisory information makes it even harder to semantically bridge the gap between 2D and 3D data.

To tackle this problem, most existing unsupervised cross-modal retrieval methods aim to learn a unified semantic representation for each instance across different modalities by maximizing the co-occurrence information, such as image-text and audio-visual [13]. However, they often ignore the semantic



(a)

T^{2}SCR

(b)

T^{2}UCR

(c)

Fig. 1. (a) The heterogeneity gap of 2D and 3D modalities. (b) 2D-3D Supervised Cross-modal Retrieval (T^{2}SCR). (c) 2D-3D Unsupervised Cross-modal Retrieval (T^{2}UCR). In contrast to T^{2}SCR, T^{2}UCR does not rely on any label annotations for guidance to address the heterogeneity gap, making it more challenging to establish semantic correspondence in the common space.

relevance information that can be derived from the intrinsic relationships between instances. It can enhance the data understanding and model performance [14]. Following the idea of single-modal unsupervised learning, we can leverage pseudo labels to capture the semantic relevance between samples [15]. However, it is impossible to ensure perfect predictions, and inevitably introduces an amount of noise in the predicted pseudo labels, leading to the model overfitting false supervision and degrading performance. Therefore, it is crucial to alleviate the impact of pseudo label noise for performance improvement. To this end, we formulate T^{2}UCR as a cross-modal learning paradigm with noisy pseudo labels, which is a novel perspective for unsupervised cross-modal learning. The core of this paradigm is to tackle the challenges of bridging the heterogeneity gap and alleviating the adverse impact of pseudo label noise.

To address the challenges mentioned above, we propose a novel robust unsupervised multimodal learning framework (i.e. RoMo) for unsupervised 2D-3D cross-modal retrieval. As shown in Fig. 2, our proposed RoMo consists of three core components, i.e., Self-matching Supervision Mechanism (SSM), Robust Discriminative Learning (RDL), and Modality-invariance Learning Mechanism (MLM). Specifically, 1) SSM enhances the prediction confidence of the models to learn discrimination in a self-supervised manner, by minimizing the entropy between the current prediction and the sharpened prior prediction for each sample. The prior prediction is obtained from a maintained feature memory bank. In this way, the model can assign higher probabilities to more confident classes, and make the samples with similar semantics compact while dissimilar ones scatter. 2) RDL aims to further mine the discrimination by using the knowledge learned by SSM, i.e., the pseudo labels generated by the trained model. However, the pseudo labels inevitably contain lots of noise due to the lack of well-labeled ground truths. To tackle this problem, we design a novel Robust Concentrating Learning Loss (RCLL) in RDL, which alleviates the optimization attention on noisy data and focuses more on clean data. 3) MLM forces modality-specific samples from the same instance to converge to a single point in the common space, thus producing modality-invariant representations. During the training process, we initially warm up the model by SSM and MLM, generating the pseudo supervision for all data. Subsequently, we employ RDL and MLM to robustly learn from the noisy pseudo labels, thereby addressing the T^{2}UCR problem effectively.

The main novelties and contributions of this paper are summarized as follows:

- We propose a novel unsupervised multimodal learning framework for 2D-3D cross-modal retrieval, named RoMo. To the best of our knowledge, RoMo addresses the issue of noise in pseudo labels for unsupervised cross-modal learning, which is rarely studied before and can remarkably improve performance.
- RoMo includes three key mechanisms, i.e. SSM, RDL and MLM. SSM warms up the model to encapsulate discrimination into the representations in a self-supervised learning manner. RDL further mines the discrimination from the learned imperfect predictions after warming up. MLM minimizes the cross-modal discrepancy to enforce SSM and RDL to produce common representations.
- To tackle the noise in the predicted pseudo labels, we leverage a novel robust loss RCLL to alleviate the influence of the uncertain samples, thus embracing robustness against predicted pseudo label noise.
- We provide theoretical derivation and extensive experiments to comprehensively verify the effectiveness of our RoMo. It also shows remarkably superior performance over the current state-of-the-art methods in various comparison experiments.

This article is organized as follows. In Section II, we conduct a review of related works. Section III offers a comprehensive explanation of our proposed framework and its implementation. Then, Section IV encompasses the description and analysis of the datasets, benchmarks, evaluation metrics, and experimental results. Finally, in Section V, we draw our conclusions and look forward to the future.

II. Related Works

A. Learning From 2D-3D Data

Learning from 2D and 3D data is a crucial step for semantic understanding [16]. With the rapid development of deep neural networks, 2D feature representation has achieved remarkable results, such as VGG, ResNet, Vision Transformer (ViT), etc. Moreover, there are also various effective methods for feature representation of 3D data [17]. These methods can deal with different 3D formats, such as point clouds, multi-view images, meshes, volumetrics, etc. Among them, PointNet, as the pioneer of a 3D feature representation model based

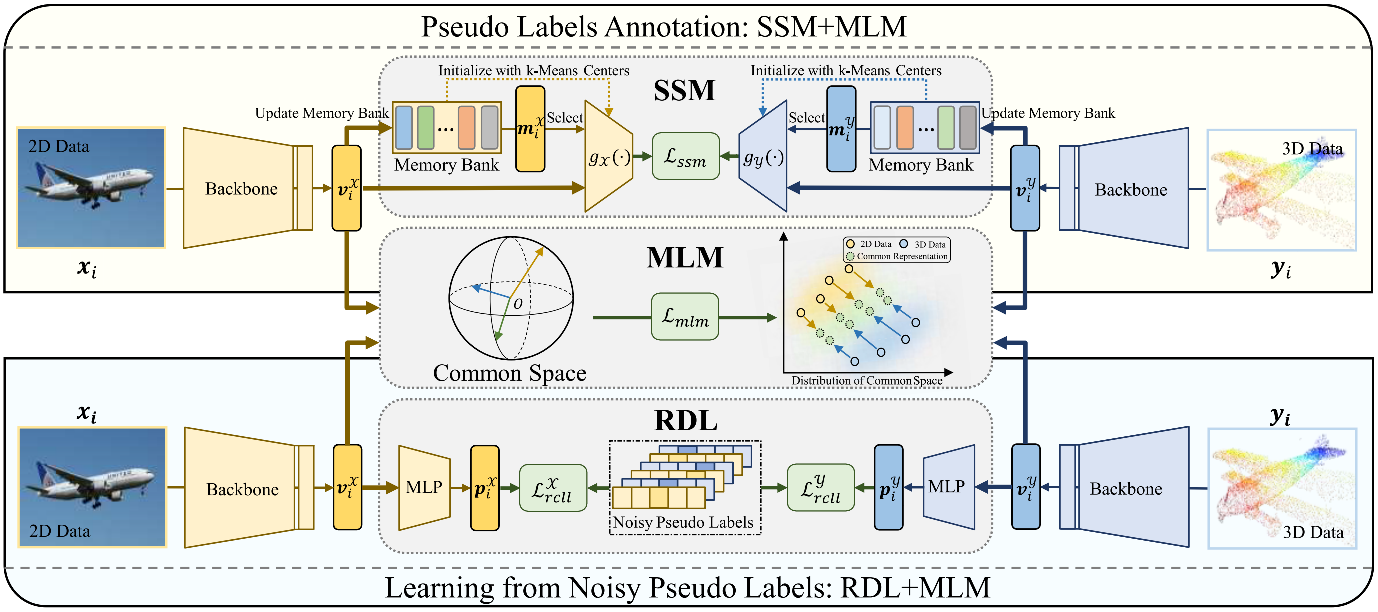


Fig. 2. The pipeline of our RoMo for unsupervised 2D-3D cross-modal retrieval. During the training process, we first carry on Pseudo Labels Annotation by SSM and MLM, to obtain pseudo labels with cross-modal semantic consistency for all data. Subsequently, we conduct Learning from Noisy Pseudo Labels through RDL and MLM to robustly learn from the noisy pseudo labels and effectively address the associated challenges of T^{2}UCR.

on deep learning, can directly extract features from unordered point cloud data [18]. DGCNN extracts features by using a dynamic graph convolutional neural network of k-Nearest Neighbors (KNN) [19]. MeshNet and MeshCNN learn features from mesh data by modeling the geometric relationship of object mesh faces [20]. For multimodal learning, these existing single-modal feature representation networks can usually be used to map 2D and 3D data to the feature space, thereby accomplishing various subtasks of multimodal learning [5]. These methods provide much guidance for our work.

B. Unsupervised Cross-Modal Retrieval Methods

Unsupervised methods are more desirable than well-labeled cross-modal retrieval techniques, as they do not require any costly sample annotation [21]. Traditional shallow unsupervised methods start from Canonical Correlation Analysis (CCA) that learn two linear transformations to maximize the correlation between different modalities [11]. To overcome the constraints of linear techniques, KCCA employs kernel methodologies, aiming to optimize correlation within the Reproducing Kernel Hilbert Space (RKHS) through the application of nonlinear transformations [11]. In [13], a Collective Matrix Factorization Hashing (CMFH) method is proposed to learn a common Hamming space by using collective matrix factorization with a latent factor model. Liu et al. propose the Fusion Similarity Hashing (FSH) method which embeds the graph-based fusion similarity between distinct modalities into the common hash representations [22].

However, the aforementioned methods fail to capture the highly nonlinear semantics present in multimodal data. To address this issue, some Deep Neural Network (DNN) based methods have been proposed recently. Deep Canonical Correlation Analysis (DCCA) exemplifies a deep cross-modal model. It effectively maps two modalities into a shared latent subspace, yielding representations that exhibit strong linear correlations through intricate nonlinear transformations [23]. Unsupervised Generative Adversarial Cross-modal Hashing (UGACH) exploits the underlying manifold structure of cross-modal data with maximum margin ranking loss [14]. Recently, an improved cross-modal hashing technique was proposed that made the hash operations learnable in contrastive learning and utilized the differentiation from all pairs rather than just hard negative pairs [24]. Chen et al. combined information theory and adversarial learning to narrow the semantic gap [25]. However, these approaches tend to overlook the semantic information present in the original cross-modal data, making them less directly applicable to T^{2}UCR.

C. Learning With Noisy Labels

Noisy labels are a common challenge in the training process, as they can mislead the learning direction [26]. To effectively deal with the noise in the label annotation, various strategies have been proposed. Current research mainly optimizes from three main aspects. The first category of methods mainly focuses on learning from clean samples and mitigates the negative impact of noisy labels by reweighting samples or correcting labels [27], [28]. Some researchers have adopted adaptive training strategies, aiming to automatically select samples with true labels for learning [29]. However, these methods often require additional high-quality annotated data, which may not be easily available in practice. The second category mainly revises the network structure and constructs a specific noise transition matrix [30]. Due to the complexity and variability of the noise, it is often difficult to accurately and reliably model the noise. The last category focuses on designing robust optimization objectives, guiding the model to learn discriminative features from noisy labels [31], [32]. These methods have a relatively wide range of applications. However, while their robust loss functions can effectively reduce the overfitting of deep neural networks on noisy labels, they may also affect the ability of the model to fit complex and hard samples. In general, the robust loss function has significant advantages of low-cost and easy training in adapting noisy labels, and also irreplaceability in practical applications.

III. Proposed Method

A. Problem Formulation and Notations

Without loss of generality, we focus on 2D data (including RGB and grayscale images) and 3D data (including point clouds and meshes) in our paper. Let $\boldsymbol{\mathcal{D}} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ be a 2D-3D dataset, where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ denote the samples of 2D and 3D data respectively, and N represents the number of 2D-3D pairs in the dataset. The sets of 2D and 3D modalities are denoted as $\mathcal{X}$ and $\mathcal{Y}$, respectively. Note that $\mathcal{C} \in \{\mathcal{X}, \mathcal{Y}\}$ in the following content. More importantly, the class labels are unavailable in the data for T^{2}UCR.

Cross-modal retrieval aims to learn two modality-specific projectors $f_{\mathcal{X}}(\cdot; \Theta_{bb}^{\mathcal{X}})$ and $f_{\mathcal{Y}}(\cdot; \Theta_{bb}^{\mathcal{Y}})$ to project different modalities into a common space, where $\Theta_{bb}^{\mathcal{X}}$ and $\Theta_{bb}^{\mathcal{Y}}$ represent the learnable parameters of the corresponding projectors, respectively. The common representations $\{\left(\boldsymbol{v}_i^{\mathcal{X}}, \boldsymbol{v}_i^{\mathcal{Y}}\right)\}_{i=1}^{N}$ from different modalities can be obtained by

$$\boldsymbol{v}_i^{\mathcal{X}} = f_{\mathcal{X}}(\boldsymbol{x}_i; \Theta_{bb}^{\mathcal{X}}) \in \mathbb{R}^L, \tag{1}$$

$$\boldsymbol{v}_i^{\mathcal{Y}} = f_{\mathcal{Y}}(\boldsymbol{y}_i; \Theta_{bb}^{\mathcal{Y}}) \in \mathbb{R}^L, \tag{2}$$

where L denotes the dimensionality of the common space. Intuitively, $\boldsymbol{v}_i^{\mathcal{X}}$ and $\boldsymbol{v}_i^{\mathcal{Y}}$ should be as close as possible if they share the same semantics, otherwise they should be far away from each other in the semantic similarity distance.

B. Overview of the Proposed Framework RoMo

Our proposed framework RoMo involves two crucial steps: Pseudo Labels Annotation and Learning from Noisy Pseudo Labels. To be specific, we foremost warm up the model to acquire pseudo labels, which contain coarse semantic information, and then utilize the pseudo labels to retrain the model. However, it is unavoidable to introduce inaccurate predictions in the pseudo labels (termed noisy pseudo labels) due to the immature model, especially in unsupervised learning. Similar to noisy labels, noisy pseudo labels will also cause the model to overfit unreliable discrimination.

Pseudo Labels Annotation aims at building the initial semantic relationships of the multimodal data, which consists of a Self-matching Supervision Mechanism (SSM) and a Modality-invariance Learning Mechanism (MLM). Inspired by the memorization effect [33], SSM utilizes the noisy initial pseudo labels to warm up the model to learn a classifier with the self-matching loss $\mathcal{L}_{ssm}$, thus acquiring applicable pseudo labels. Moreover, MLM attempts to minimize the cross-modal discrepancy in the common space to learn modality-invariant representations. The overall loss $\mathcal{L}_{PLA}$ can be formulated as follows:

$$\mathcal{L}_{PLA} = \lambda_1 \mathcal{L}_{ssm} + (1 - \lambda_1)\mathcal{L}_{mlm}, \tag{3}$$

where $\lambda_1 \in [0, 1]$ balances the contribution between $\mathcal{L}_{ssm}$ and $\mathcal{L}_{mlm}$.

Learning from Noisy Pseudo Labels is leveraged to learn the common discriminative representations from the noisy pseudo labels obtained by Pseudo Labels Annotation. To mitigate the negative interference of noise in the pseudo labels, we present a novel Robust Concentrating Learning Loss (RCLL) $\mathcal{L}_{rcll}$ to make the model reduce the focus on the hard (likely mislabeled) samples. Similar to Pseudo Labels Annotation, we also leverage the $\mathcal{L}_{mlm}$ to narrow the cross-modal heterogeneity gap. The overall loss $\mathcal{L}_{LNP}$ could be formulated as:

$$\mathcal{L}_{LNP} = \lambda_2 \mathcal{L}_{rcll} + (1 - \lambda_2)\mathcal{L}_{mlm}, \tag{4}$$

where $\lambda_2 \in [0, 1]$ determines the relative influence of $\mathcal{L}_{rcll}$ and $\mathcal{L}_{mlm}$.

C. Self-Matching Supervision Mechanism

Inspired by contrastive learning [34], we regard the unsupervised instance-level discrimination as a metric learning problem. Here, distances (similarities) between instances are computed directly from their features in a non-parametric manner. Specifically, two memory banks $\{\boldsymbol{\mathcal{M}}^{\mathcal{C}}\}_{\mathcal{C}\in\{\mathcal{X},\mathcal{Y}\}}$ are exploited to learn discrimination from the two modalities in an unsupervised manner. Each modality-specific memory bank $\boldsymbol{\mathcal{M}}^{\mathcal{C}}$ involves the representations of all instances, which can be formulated as:

$$\boldsymbol{\mathcal{M}}^{\mathcal{C}} = [\boldsymbol{m}_1^{\mathcal{C}}, \cdots, \boldsymbol{m}_N^{\mathcal{C}}], \tag{5}$$

where $\boldsymbol{m}_i^{\mathcal{C}} \in \mathbb{R}^L$ are the representations of samples (i.e. $\boldsymbol{x}_i$ or $\boldsymbol{y}_j$). First, $f_{\mathcal{C}}(\cdot; \Theta_{bb}^{\mathcal{C}})$ is initialized by corresponding pre-trained 2D model and 3D model. Second, the memory banks are initialized with the features extracted by $f_{\mathcal{C}}(\cdot; \Theta_{bb}^{\mathcal{C}})$. Third, we combine the memory banks of different modalities and perform k-means on the combined features to obtain global clustering centers. Finally, the modality-specific k-means, whose centers are initialized by the obtained centers, are conducted to obtain the centers to initialize the modality-specific trainable classifiers $g_{\mathcal{C}}(\cdot)$ for the 2D (i.e. $g_{\mathcal{X}}(\cdot)$) and 3D (i.e. $g_{\mathcal{Y}}(\cdot)$) modalities, respectively. In particular, we adopt the over-clustering strategy to determine the number of clusters in k-means [15]. Due to the unknown number of clusters, this strategy allows clustering the instances into more clusters than the truth, which helps relax the restrictions on parameter dependencies.

Then, the trainable parameters of models are updated with gradient descent while the memory banks are updated with momentum iteratively. Specifically, the features in $\boldsymbol{\mathcal{M}}^{\mathcal{C}}$ are updated by $\boldsymbol{v}_i^{\mathcal{C}}$ with momentum η after each iteration,

$$\boldsymbol{m}_i^{\mathcal{C}} = \eta \boldsymbol{m}_i^{\mathcal{C}} + (1 - \eta)\boldsymbol{v}_i^{\mathcal{C}}. \tag{6}$$

To avoid excessive memory consumption, we follow [34] to randomly sample N' features from the memory bank for each mini-batch. Accordingly, the features of the memory banks are updated in a batch-by-batch manner.

Unsupervised learning could be achieved by maximizing the mutual information between the network prediction and its corresponding counterparts (e.g., proxies in memory bank), thus encapsulating the intrinsic discrimination into the representations [35]. However, existing contrastive learning methods maximize mutual information only at the representation level [36], not the semantic level, which leads to a task gap between upstream representation learning and downstream semantic retrieval, resulting in performance degradation. To address this issue, we present to maximize mutual information in the predicted label space, thereby ensuring consistency between the upstream training and downstream retrieval. To this end, we present the mutual information I between predictions and supervised signals at the semantic level as follows.

$$I=\sum_{i=1}^{N'}\sum_{j=1}^{N'}P\left(I_{\boldsymbol{m}_i^{\mathcal{C}}},I_{\boldsymbol{v}_j^{\mathcal{C}}}\right)\log\left(\frac{P\left(I_{\boldsymbol{m}_i^{\mathcal{C}}},I_{\boldsymbol{v}_j^{\mathcal{C}}}\right)}{P\left(I_{\boldsymbol{m}_i^{\mathcal{C}}}\right)P\left(I_{\boldsymbol{v}_j^{\mathcal{C}}}\right)}\right),\tag{7}$$

where $I_{\boldsymbol{m}_i^{\mathcal{C}}}=\sigma\left(g_{\mathcal{C}}\left(\boldsymbol{m}_i^{\mathcal{C}}\right)\right)$ and $I_{\boldsymbol{v}_j^{\mathcal{C}}}=\sigma\left(g_{\mathcal{C}}\left(\boldsymbol{v}_j^{\mathcal{C}}\right)\right)$. $\sigma(\cdot)$ denotes the softmax function. Actually, $\sigma\left(g_{\mathcal{C}}\left(\boldsymbol{m}_i^{\mathcal{C}}\right)\right)$ could be seen as a guessed label using the predictions of the proxy $\boldsymbol{m}_i^{\mathcal{C}}$. Thanks to semantic contrastive learning, we can conduct label refining in the predicted label space to improve the performance further. To this end, we employ a sharpening function $S_{\mathcal{C}}(\cdot,\tau_1)=(\sigma(g_{\mathcal{C}}(\cdot)))^{1/\tau_1}$ to lower the entropy of the guessed label distribution, where τ_1 is a temperature parameter. When $\tau_1\rightarrow 0$, the outputs of $S_{\mathcal{C}}(\cdot,\tau_1)$ tends to approach a Dirac distribution. By reducing the value of τ_1, the model is encouraged to yield lower-entropy predictions, thus making the model more confident in its decisions. Lower-entropy predictions enhance model performance by reducing the uncertainty of model predictions, thereby improving stability and reliability. Lower entropy indicates more certain and reliable predictions, aiding the model in accurately learning cross-modal data features and patterns, thus enhancing self-learning accuracy. By introducing the sharpening function to Equation (7), the loss function $\mathcal{L}_{ssm}^{\mathcal{C}}$ can be written as:

$$\mathcal{L}_{ssm}^{\mathcal{C}}=-\sum_{i=1}^{N'}\sum_{j=1}^{N'}P\left(S_{\mathcal{C}}\left(\boldsymbol{m}_i^{\mathcal{C}},\tau_1\right),I_{\boldsymbol{v}_j^{\mathcal{C}}}\right)\log\left(\frac{P\left(S_{\mathcal{C}}\left(\boldsymbol{m}_i^{\mathcal{C}},\tau_1\right),I_{\boldsymbol{v}_j^{\mathcal{C}}}\right)}{P\left(S_{\mathcal{C}}\left(\boldsymbol{m}_i^{\mathcal{C}},\tau_1\right)\right)P\left(I_{\boldsymbol{v}_j^{\mathcal{C}}}\right)}\right).\tag{8}$$

Finally, the self-matching loss $\mathcal{L}_{ssm}$ for all modalities could be formulated as follows:

$$\mathcal{L}_{ssm}=\sum_{\mathcal{C}\in\{\mathcal{X},\mathcal{Y}\}}\mathcal{L}_{ssm}^{\mathcal{C}}.\tag{9}$$

D. Robust Discriminative Learning

After the warm-up stage (i.e., Pseudo Labels Annotation), we could exploit the pseudo labels predicted by the trained model to improve the unsupervised cross-modal learning. Although pseudo labels could provide some extra semantic information, it is inevitable to introduce unreliable discrimination into the pseudo labels due to the lack of ground-truth supervision, leading to performance degradation. To tackle this issue, we propose a robust discriminative learning approach to alleviate the adverse impact of noisy pseudo labels. The pseudo labels are denoted as $\mathcal{Z}=\{\left(z_i^{\mathcal{X}},z_i^{\mathcal{Y}}\right)\}_{i=1}^{N}$, where $z_i^{\mathcal{X}}=g_{\mathcal{X}}(f_{\mathcal{X}}(\boldsymbol{x}_i))$ and $z_i^{\mathcal{Y}}=g_{\mathcal{Y}}(f_{\mathcal{Y}}(\boldsymbol{y}_i))$.

Different from the stage of Pseudo Labels Annotation, our RDL trains new models $f_{\mathcal{C}}(\cdot;\Theta_{bb}^{\mathcal{C}})$ in Learning from Noisy Pseudo Labels, where $\Theta_{bb}^{\mathcal{C}}$ is the trainable parameters of neural networks for 2D (i.e. $\Theta_{bb}^{\mathcal{X}}$) and 3D (i.e. $\Theta_{bb}^{\mathcal{Y}}$) modalities, respectively. Similar to Pseudo Labels Annotation, modality-specific classifier $h_{\mathcal{C}}(\cdot;\Theta_{mlp}^{\mathcal{C}})$ is employed to bridge the common space and pseudo label space, where $\Theta_{mlp}^{\mathcal{C}}$ is the trainable parameters of the classifiers. Then, we can obtain the classification predictions as below:

$$\boldsymbol{p}_i^{\mathcal{C}}=h_{\mathcal{C}}(\boldsymbol{v}_i^{\mathcal{C}};\Theta_{mlp}^{\mathcal{C}}),\tag{10}$$

where $\boldsymbol{p}_i^{\mathcal{C}}$ represents the predicted probability belonging to different pseudo classes for the samples (i.e. $\boldsymbol{x}_i$ or $\boldsymbol{y}_i$).

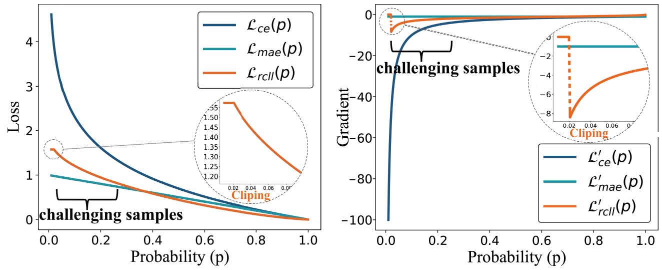
Previous studies [37] have shown that widely used supervised loss functions, such as Cross Entropy (CE), are prone to overfitting noisy labels. This may occur because CE-like losses always focus more on hard samples, which are often mislabeled in the presence of noise, leading to overfitting and performance degradation [33]. To mitigate this problem, Mean Absolute Error (MAE) is employed to enhance robustness, as theoretically demonstrated in [32]. However, MAE lacks the ability to focus on more challenging samples, treating all samples equally, resulting in underfitting and insufficient learning, as shown in Fig. 5.

To address the aforementioned dilemma, we introduce a novel loss function RCLL (i.e. $\mathcal{L}_{rcll}$) to mitigate the adverse impact of noisy labels. Specifically, compared to CE, our RCLL will reduce its focus on hard samples, which are likely to be mislabeled, thus alleviating the overfitting issue encountered by CE. On the other hand, compared to MAE, our RCLL will discard the high-risk samples and pay more attention to the challenging samples, embracing performance improvement. This loss function is formulated as follows:

$$\mathcal{L}_{rcll}^{\mathcal{C}}=\frac{1}{N}\sum_{i=1}^{N}\left(-\log\left(\boldsymbol{p}_i^{\mathcal{C}}\right)\right)^{\alpha}\left(1-\boldsymbol{p}_i^{\mathcal{C}}\right),\tag{11}$$

where $\alpha\in(0,1)$ is a hyper-parameter.

Besides, we employ a clipping strategy to prevent extremely difficult or high-risk samples from producing excessive gradient values. The clipping threshold is empirically set as $1/K$, where K is the number of clusters. This clipping operation plays a crucial role in stabilizing the training process and enhancing the robustness of our algorithm, particularly in the presence of noisy or outlier data points. Specifically, the clipping operation imposes a constraint on the gradients of the loss function, limiting their magnitude to a predefined threshold. This mechanism effectively prevents excessively large gradient updates that may arise from noisy samples, thereby mitigating the risk of unstable training or divergence. By constraining the optimization updates for such samples, the clipping operation ensures that the model focuses on learning from informative data points while reducing the influence of pseudo label noise.



(a) Loss Value vs Probability. (b) Gradient Value vs Probability.

Fig. 3. Comparison among CE, MAE, and RCLL with $\alpha = 0.35$. Compared to CE, RCLL aims to decrease focus on the hard samples, which are probably mislabeled. On the contrary, compared with CE, RCLL pays more attention to the challenging samples for performance improvement.

Finally, the overall loss function $\mathcal{L}_{rcll}$ can be written as:

$$\mathcal{L}_{rcll} = \begin{cases} \sum\limits_{\mathcal{C} \in \{\mathcal{X}, \mathcal{Y}\}} \mathcal{L}^{\mathcal{C}}_{rcll}, & \boldsymbol{p}^{\mathcal{C}}_i > \dfrac{1}{K} \\ T, & \text{otherwise}, \end{cases} \tag{12}$$

where T is a constant, which is equal to the value of $\sum\limits_{\mathcal{C} \in \{\mathcal{X}, \mathcal{Y}\}} \mathcal{L}^{\mathcal{C}}_{rcll}$ when $\boldsymbol{p}^{\mathcal{C}}_i = 1/K$.

To visually study Equation (12), we draw comparison curves in Fig. 3. Compared with CE, one can see that our RCLL remarkably reduces the loss values and gradients on hard samples, which probably are mislabeled, thus preventing models from overfitting against the noisy labels.

Furthermore, we conduct theoretical analyses to investigate the property of RCLL against noisy labels as follows. The detailed proof process can be found in the APPENDIX.

Property 1: $\mathcal{L}^{\mathcal{C}}_{rcll}$ is equivalent to MAE when $\alpha \to 0$.

Property 2: For any input (e.g. $\boldsymbol{x}_i$ or $\boldsymbol{y}_i$) and $\alpha \in (0, 1)$, $\mathcal{L}^{\mathcal{C}}_{rcll}$ seek eclectic focus on the challenging samples to mitigate overfitting and underfitting problems.

According to Properties 1 and 2, one can draw the conclusion that RCLL is robust against noisy labels, which could effectively address the inevitable labeling noise produced by the Pseudo Labels Annotation.

E. Modality-Invariance Learning Mechanism

In addition to learning discrimination from each unlabeled modality through SSM and RDL, T^{2}UCR also needs to eliminate the cross-modal discrepancy between 2D and 3D modalities. To this end, we present a Modality-invariance Learning Mechanism (MLM) to reduce the heterogeneity gap by maximizing the mutual information between different modalities. In other words, MLM aims at maximizing the cross-modal consistency to effectively capture the shared information between distinct modalities.

More specifically, we first define the probability of sample $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ belonging to the i-th instance as $P(i|\boldsymbol{x}_i)$ and $P(i|\boldsymbol{y}_i)$, respectively. These probabilities are equivalent to $P(i|\boldsymbol{v}^{\mathcal{X}}_i)$ and $P(i|\boldsymbol{v}^{\mathcal{Y}}_i)$, where $\boldsymbol{v}^{\mathcal{X}}_i$ and $\boldsymbol{v}^{\mathcal{Y}}_i$ are the representations of $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$. The formulation of $P(i|\boldsymbol{v}^{\mathcal{C}}_i)$ denotes as:

$$P(i|\boldsymbol{v}^{\mathcal{C}}_i) = \frac{\sum\limits_{\mathcal{E} \in \{\mathcal{X}, \mathcal{Y}\}} \exp\left((\boldsymbol{v}^{\mathcal{E}}_i)^T \boldsymbol{v}^{\mathcal{C}}_i / \tau_2\right)}{\sum\limits_{\mathcal{E} \in \{\mathcal{X}, \mathcal{Y}\}} \sum\limits_{t=1}^{N} \exp\left((\boldsymbol{v}^{\mathcal{E}}_t)^T \boldsymbol{v}^{\mathcal{C}}_i / \tau_2\right)}, \tag{13}$$

where τ_2 is a temperature parameter. To mitigate the inherent cross-modal discrepancies, we enforce the representations of the same instance to be compact while ones of different instances are scattered in the common space. By minimizing $\mathcal{L}_{mlm}$, the outputs of multimodal networks are contrasted with each other at the instance level, thereby encapsulating shared discrimination in the common space. The MLM loss $\mathcal{L}_{mlm}$ could be formulated as follow:

$$\mathcal{L}_{mlm} = -\frac{1}{N} \sum_{\mathcal{C} \in \{\mathcal{X}, \mathcal{Y}\}} \sum_{i=1}^{N} \log(P(i|\boldsymbol{v}^{\mathcal{C}}_i)). \tag{14}$$

F. The Training Strategy of Proposed RoMo

In this section, we present the training procedure for RoMo outlined in Algorithm 1.

IV. Experiments and Analyses

A. Datasets

To confirm the effectiveness of our proposed framework, we carried out extensive experiments on four benchmark datasets: 3D MNIST [38], ModelNet10 [1], ModelNet40 [1], and MI3DOR [39]. The detailed experimental settings related to the dataset are provided in Table I.

1) 3D MNIST [38]: The 3D MNIST dataset is a 3D extension of the classic MNIST dataset, which is collected in Kaggle. In this dataset, the data of the 3D modality is Point Clouds, while the data of the 2D modality includes RGB Images and Gray Images. We divide the dataset into two subsets: 5, 000 2D-3D pairs for the training set and 1, 000 for the testing set.

2) ModelNet10 [1]: The ModelNet10 dataset is a widely-utilized benchmark dataset within the realm of 3D object recognition and classification research. This dataset is comprised of a diverse range of 3D CAD models, including Point Cloud and Mesh representations, as well as our processed 2D data in both Grey Images and RGB Images. For the convenience of research and testing, we have partitioned the complete dataset into two separate subsets, consisting of 3, 991 2D-3D pairs for training and 908 for testing, thereby facilitating more efficient and precise analysis.

3) ModelNet40 [1]: The ModelNet40 dataset is a more comprehensive version of the ModelNet10 dataset, consisting of 3D CAD models from 40 distinct categories. To enhance the effectiveness of the research, we have thoughtfully partitioned the complete dataset into two distinct subsets, 9, 843 2D-3D pairs for training and 2, 468 for testing.

4) MI3DOR [39]: MI3DOR is specifically designed to support monocular image-based 3D model retrieval, with a focus on retrieving 3D models that are based on queries made by object-centric monocular images. The 3D data in the MI3DOR dataset is provided in Mesh format, while the

TABLE I

General Statistics of the Four Datasets Used in the Experiments, Where "*/*" in the "*Instances*" Column Stands for the Number of Training/Testing Sets and "*/*" in the "*RGB Image Feature*", "*Gray Image Feature*", "*Point Cloud Feature*", "*Mesh Feature*" Column Stands for the Feature Dimension in *Pseudo Labels Annotation* and *Learning From Noisy Pseudo Labels*

Dataset	Instances	Classes	RGB Image Feature	Gray Image Feature	Point Cloud Feature	Mesh Feature
3D MNIST [38]	5000 / 1000	10	1024 / 256	1024 / 256	1024 / 256	-
ModelNet10 [1]	2468 / 908	10	1024 / 256	1024 / 256	1024 / 256	1024 / 256
ModelNet40 [1]	9840 / 3991	40	1024 / 256	1024 / 256	1024 / 256	1024 / 256
MI3DOR [39]	3848 / 3848	21	1024 / 256	1024 / 256	-	1024 / 256

Algorithm 1 Training Strategy of Proposed RoMo

```
Input: The 2D-3D cross-modal training dataset $\mathcal{D}$, $\alpha$,
       $\lambda_1$, $\lambda_2$, $\tau_1$, $\tau_2$, $\eta$, maximal epoch $N_e$, batch
       size $N_b$ and learning rate $lr$.
Output: Optimized parameters
        $\{\Theta^{\mathcal{X}}_{bb}, \Theta^{\mathcal{X}}_{mlp}, \Theta^{\mathcal{Y}}_{bb}, \Theta^{\mathcal{Y}}_{mlp}\}$.
Initialize $\Theta^{\mathcal{X}}_{bb}$ and $\Theta^{\mathcal{Y}}_{bb}$ with pre-trained parameters;
Calculate the representations for all samples
  by Equations (1) and (2);
Acquire modality-specific centers in memory banks
  $\mathcal{M}^{\mathcal{X}}$ and $\mathcal{M}^{\mathcal{Y}}$ by modality-specific clustering to
  initialize classifiers;
for 1, 2, $\cdots$, 20 do
    repeat
        Randomly select $N_b$ samples from each
          modality to build a multimodal mini-batch;
        Calculate the representations for all samples of
          the mini-batch by $f_{\mathcal{X}}(\cdot; \Theta^{\mathcal{X}}_{bb})$ and $f_{\mathcal{Y}}(\cdot; \Theta^{\mathcal{Y}}_{bb})$;
        Update network parameters $\{\Theta^{\mathcal{X}}_{bb}, \Theta^{\mathcal{Y}}_{bb}\}$ by
          minimizing $\mathcal{L}_{PLA}$ in Equation (3) with Adam;
        Update memory banks with Equation (6);
    until all samples selected;
end
Generate the pseudo labels $\mathcal{Z}$ for each samples;
Reinitialize $f_{\mathcal{X}}(\cdot; \Theta^{\mathcal{X}}_{bb})$ and $f_{\mathcal{Y}}(\cdot; \Theta^{\mathcal{Y}}_{bb})$ with pre-trained
  parameters;
for 1, 2, $\cdots$, $N_e - 20$ do
    repeat
        Randomly select $N_b$ samples from each
          modality to build a multimodal mini-batch;
        Calculate the representations for all samples of
          the mini-batch by $f_{\mathcal{X}}(\cdot; \Theta^{\mathcal{X}}_{bb})$ and $f_{\mathcal{Y}}(\cdot; \Theta^{\mathcal{Y}}_{bb})$;
        According to $\mathcal{Z}$, update network parameters
          $\{\Theta^{\mathcal{X}}_{bb}, \Theta^{\mathcal{Y}}_{bb}, \Theta^{\mathcal{X}}_{mlp}, \Theta^{\mathcal{Y}}_{mlp}\}$ by minimizing $\mathcal{L}_{LNP}$
          in Equation (4) with Adam;
    until all samples selected;
end
```

2D data is available in both RGB Images and Gray Images. For the convenience of the experiments and optimizing our analysis, we split the dataset into two separate training and testing sets, each of which consisted of $3,848$ 2D-3D pairs.

B. Benchmark Methods and Evaluation Metric

In the experimental section, we will assess the validity of the proposed approach against several state-of-the-art benchmarks on T^{2}UCR. To provide meaningful comparisons, we will evaluate our approach alongside three shallow methods (i.e. CCA [40], KCCA [11], MCCA [41]) and eleven deep learning methods (i.e. DCCA [23], DCCAE [42], RevGard [43], MEDA [44], DJSRH [45], JDSH [46], DGCPN [47], UCCH [24], SCL [48], and PT-FUCH [49]. Our evaluation approach will be based on assessing the mean average precision of all class retrieved results (mAP@All). To validate the robustness of RCLL in dealing with pseudo label noise, we also compared our proposed RCLL with other commonly-used robust loss functions, such as CE [51], MAE [32], AUE [52], NCE+AGCE [52].

C. Experimental Settings

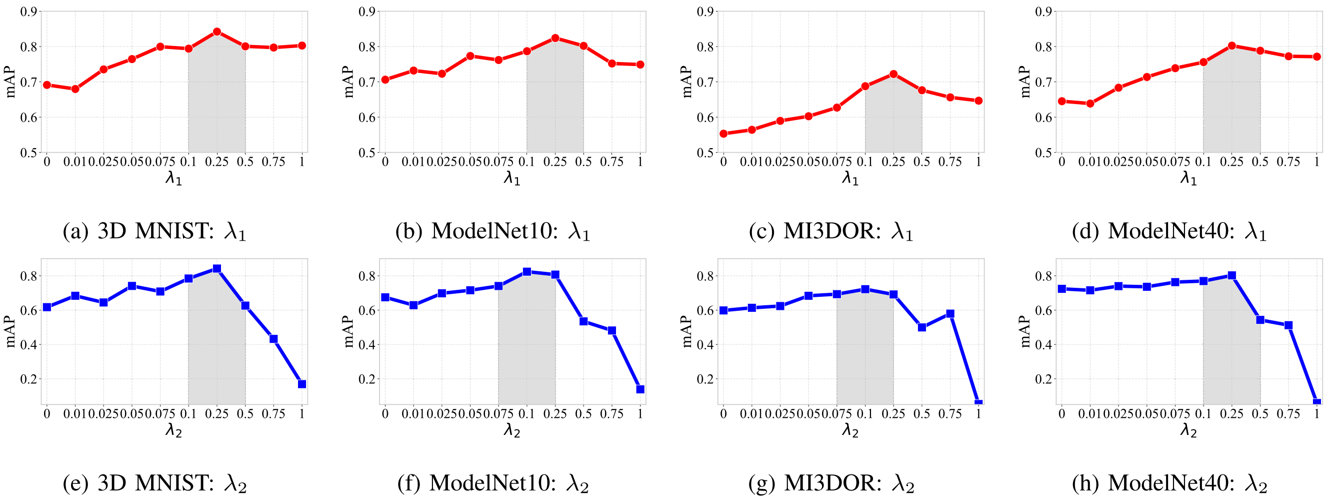
To comprehensively compare the performance of our method with the benchmarks, we established the following eight tasks, i.e. RP (RGB Images to Point Cloud), RM (RGB Images to Mesh), GP (Gray Images to Point Cloud), GM (Gray Images to Mesh), PR (Point Cloud to RGB Images), PG (Point Cloud to Grayscale Images), MR (Mesh to RGB Images), and MG (Mesh to Gray Images). Our RoMo is trained on Nvidia GeForce RTX 3090 GPUs with Pytorch. All comparison experiments were conducted on Nvidia GeForce RTX 3090 and Tesla V100 GPUs. The code for RoMo is available at https://github.com/LYXRhythm/RoMo.

In the training process, a pre-trained ResNet18 is utilized to initialize the projector of RGB Images and Gray Images. Pre-trained DGCNN [19] and MeshNet [20] are adopted as the projector of Point Cloud and Mesh, respectively. We reload pre-trained parameters in both warm-up and robust learning steps for backbones to reduce error accumulation. Meanwhile, we set the learning rate $lr = 2 \times 10^{-4}$ and train the models with Adam. Maximal epochs N_e is set as 70, and batch size N_b is set as 50. As suggested by [54], momentum η is set to 0.9 to promote smooth optimization. The parameter analyses of α, λ_1, and λ_2 are shown in the subsequent subsections, i.e., Figs. 4 and 5.

D. Comparison With the State-of-the-Art Methods

We comprehensively analyze the hopeful and promising performance of our proposed RoMo across four benchmark 2D-3D multimodal datasets. The experimental results are presented on Tables II to V. Based on the observations, we can draw the following viewpoints.

- The proposed framework exhibits promising performance compared to benchmarks. Our RoMo outputs better than the best competitor by 0.048, 0.077, 0.053, 0.018 for



(a) 3D MNIST: λ_1 (b) ModelNet10: λ_1 (c) MI3DOR: λ_1 (d) ModelNet40: λ_1

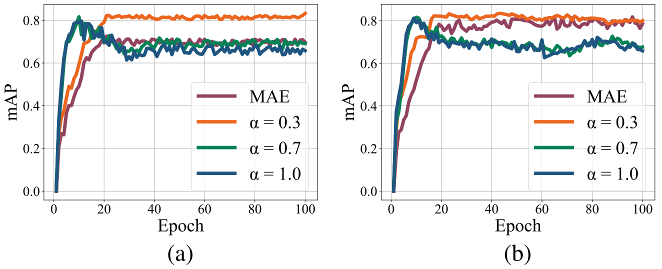
(e) 3D MNIST: λ_2 (f) ModelNet10: λ_2 (g) MI3DOR: λ_2 (h) ModelNet40: λ_2

Fig. 4. Retrieval average performance of RoMo with different values of trade-off parameters on λ_1 and λ_2. The **gray shaded area** indicates the recommended parameter range suggested by the authors for further fine-tuning.

TABLE II

PERFORMANCE COMPARISON IN TERMS OF mAP SCORES ON 3D MNIST AND ModelNet10 DATASET, INCLUDING 14 STATE-OF-THE-ART BENCHMARK METHODS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, WHILE THE SECOND HIGHEST PERFORMANCE IS UNDERLINED

Method	3D MNIST				ModelNet10							
	2D → 3D		3D → 2D		2D → 3D				3D → 2D			
	RP	GP	PR	PG	RP	RM	GP	GM	PR	PG	MR	MG
CCA [40]	0.394	0.383	0.415	0.386	0.625	0.629	0.604	0.613	0.627	0.624	0.618	0.640
KCCA [11]	0.470	0.481	0.457	0.459	0.566	0.539	0.570	0.541	0.659	0.661	0.658	0.657
MCCA [41]	0.437	0.415	0.427	0.420	0.645	0.658	0.663	0.654	0.639	0.641	0.633	0.630
DCCA [23]	0.605	0.590	0.593	0.587	0.684	0.681	0.682	0.647	0.678	0.689	0.689	0.683
DCCAE [42]	0.602	0.593	0.625	0.613	0.703	0.688	0.692	0.686	0.693	0.684	0.705	0.678
RevGard [43]	0.608	0.613	0.648	0.620	0.659	0.676	0.675	0.673	0.664	0.661	0.658	0.678
MEDA [44]	0.666	0.669	0.653	0.675	0.711	0.677	0.740	0.680	0.716	0.699	0.708	0.704
DJSRH [45]	0.623	0.625	0.622	0.614	0.548	0.578	0.541	0.574	0.548	0.543	0.554	0.545
JDSH [46]	0.745	0.743	0.747	0.761	0.757	0.754	0.732	0.755	0.761	0.766	0.776	0.777
DGCPN [47]	0.791	0.787	0.786	0.771	0.765	0.751	0.793	0.757	0.759	0.761	0.769	0.753
UCCH [24]	0.798	0.784	0.781	0.808	0.771	0.769	0.758	0.758	0.770	0.776	0.771	0.779
SCL [48]	0.702	0.701	0.735	0.701	0.735	0.735	0.736	0.749	0.739	0.762	0.799	0.754
PT-FUCH [49]	0.724	0.731	0.745	0.732	0.765	0.750	0.770	0.766	0.762	0.763	0.761	0.756
CFRH [50]	0.683	0.736	0.691	0.739	0.692	0.714	0.693	0.741	0.689	0.703	0.699	0.693
RoMo	**0.849**	**0.843**	**0.836**	**0.839**	**0.829**	**0.812**	**0.848**	**0.811**	**0.832**	**0.830**	**0.813**	**0.813**



(a) (b)

Fig. 5. The influence of α to robustness on ModelNet10 dataset. The epoch count begins at the Learning from Noisy Pseudo Labels stage. (a) The average mAP values of 2D-3D retrieval, including RP, RM, GP, and GM. (b) The average mAP values of 3D-2D retrieval, including PR, PG, MR, and MG.

average on four datasets, respectively. This indicates that achieving discriminative mining of semantic information and setting it as the guidance are of great significance for the better understanding of T^{2}UCR task.

- During the training process, we observed a synergistic improvement in the results of $2D \rightarrow 3D$ and $3D \rightarrow 2D$. This indicates that obtaining accurate feature mapping models is a foundation for achieving effective cross-modal learning.
- Due to their limited fitting ability, shallow methods often fail to achieve good performances when processing complex 2D-3D data. The deep learning method can meet the requirements of capturing intricate and nonlinear feature expressions in T^{2}UCR problems.
- We attribute the beneficial effects of RoMo over others to the following main reasons: 1) discriminative mining of unsupervised cross-modal data achieved through self-supervised learning; 2) robust learning from imperfectly labeled data; and 3) consideration of cross-modal consistency throughout all processes, thereby mitigating the issue of excessive cross-modal noise correspondence.

To further illustrate the advancement, additional experimental results, such as feature visualizations and qualitative results, can be found in the APPENDIX.

TABLE III

PERFORMANCE COMPARISON IN TERMS OF mAP SCORES ON MI3DOR AND ModelNet40 DATASET, INCLUDING 14 STATE-OF-THE-ART BENCHMARK METHODS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, WHILE THE SECOND HIGHEST PERFORMANCE IS UNDERLINED

Method	MI3DOR				ModelNet40							
	2D → 3D		3D → 2D		2D → 3D				3D → 2D			
	RM	GM	MR	MG	RP	RM	GP	GM	PR	PG	MR	MG
CCA [40]	0.264	0.283	0.283	0.283	0.532	0.551	0.513	0.524	0.524	0.520	0.505	0.485
KCCA [11]	0.325	0.331	0.331	0.322	0.514	0.502	0.504	0.507	0.511	0.519	0.486	0.457
MCCA [41]	0.315	0.325	0.326	0.331	0.541	0.534	0.551	0.550	0.533	0.544	0.541	0.535
DCCA [23]	0.518	0.480	0.523	0.537	0.584	0.603	0.576	0.584	0.569	0.556	0.551	0.541
DCCAE [42]	0.532	0.553	0.543	0.543	0.593	0.577	0.588	0.587	0.572	0.576	0.572	0.567
RevGard [43]	0.541	0.563	0.531	0.570	0.703	0.701	0.726	0.733	0.691	0.670	0.706	0.700
MEDA [44]	0.583	0.609	0.582	0.595	0.721	0.720	0.716	0.712	0.710	0.713	0.704	0.712
DJSRH [45]	0.546	0.535	0.532	0.554	0.665	0.662	0.667	0.670	0.667	0.653	0.675	0.674
JDSH [46]	0.611	0.573	0.573	0.591	0.732	0.735	0.749	0.739	0.744	0.738	0.747	0.755
DGCPN [47]	0.612	0.587	0.587	0.591	0.705	0.707	0.694	0.698	0.699	0.694	0.704	0.687
UCCH [24]	0.667	0.674	0.677	0.658	0.755	0.768	0.761	0.778	0.739	0.759	0.763	0.786
SCL [48]	0.624	0.631	0.645	0.632	0.765	0.750	0.770	0.766	0.762	0.763	0.761	0.756
PT-FUCH [49]	0.664	0.678	0.692	0.688	0.744	0.768	0.792	0.783	0.744	0.757	0.772	0.746
CFRH [50]	0.616	0.601	0.583	0.613	0.733	0.734	0.742	0.751	0.733	0.729	0.736	0.728
RoMo	**0.744**	**0.719**	**0.721**	**0.705**	**0.810**	**0.816**	**0.810**	**0.785**	**0.801**	**0.811**	**0.780**	**0.801**

TABLE IV

PERFORMANCE COMPARISON IN TERMS OF mAP SCORES ON 3D MNIST AND ModelNet10 DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, WHILE THE SECOND HIGHEST PERFORMANCE IS UNDERLINED

Method	3D MNIST				ModelNet10							
	2D → 3D		3D → 2D		2D → 3D				3D → 2D			
	RP	GP	PR	PG	RP	RM	GP	GM	PR	PG	MR	MG
CE [51]	0.777	0.794	0.781	0.755	0.770	0.768	0.784	0.766	0.763	0.775	0.785	0.779
MAE [32]	0.797	0.805	0.807	0.790	0.792	0.798	0.813	0.794	0.789	0.801	0.793	0.790
GCE [53]	0.831	0.832	0.827	0.823	0.826	0.806	0.833	0.803	0.818	0.809	0.801	0.802
AUE [52]	0.819	0.810	0.803	0.817	0.791	0.791	0.772	0.787	0.788	0.799	0.809	0.783
NCE+AGCE [52]	0.829	0.812	0.825	0.823	0.801	0.809	0.836	0.804	0.789	0.812	**0.813**	0.794
RoMo	**0.849**	**0.843**	**0.836**	**0.839**	**0.829**	**0.812**	**0.848**	**0.811**	**0.832**	**0.830**	**0.813**	**0.813**

TABLE V

PERFORMANCE COMPARISON IN TERMS OF mAP SCORES ON MI3DOR AND ModelNet40 DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, WHILE THE SECOND HIGHEST PERFORMANCE IS UNDERLINED

Method	MI3DOR				ModelNet40							
	2D → 3D		3D → 2D		2D → 3D				3D → 2D			
	RM	GM	MR	MG	RP	RM	GP	GM	PR	PG	MR	MG
CE [51]	0.709	0.691	0.697	0.679	0.774	0.766	0.778	0.754	0.772	0.797	0.753	0.771
MAE [32]	0.717	0.698	0.693	0.688	0.795	0.788	0.773	0.775	0.788	0.781	0.773	0.781
GCE [53]	0.733	0.707	0.717	0.701	0.807	0.788	0.798	0.774	0.793	0.807	0.769	0.791
AUE [52]	0.705	0.685	0.718	0.693	0.791	0.793	0.807	0.762	0.781	0.787	0.779	0.783
NCE+AGCE [52]	0.729	0.705	0.712	0.704	0.806	0.793	0.796	0.781	0.786	0.798	0.774	0.792
RoMo	**0.744**	**0.719**	**0.721**	**0.705**	**0.810**	**0.816**	**0.810**	**0.785**	**0.801**	**0.811**	**0.780**	**0.801**

E. Ablation Studies

In the ablation study, we investigate each component of our method. The results are shown in Tables VI and VII.

- Without $\mathcal{L}_{ssm}$ in $\mathcal{L}_{PLA}$, RoMo fails to accurately capture semantic information, leading to the problem of semantic loss and low discriminative in generated pseudo labels.
- Without $\mathcal{L}_{mlm}$ in $\mathcal{L}_{PLA}$, RoMo fails to bridge the cross-modal gap, resulting in inconsistent semantic expression between different modalities, thus affecting subsequent performance.
- Without $\mathcal{L}_{rcll}$ in $\mathcal{L}_{LNP}$, RoMo lacks clear semantic guidance, making it difficult to effectively uncover cross-modal semantic knowledge in an unsupervised manner from the multimodal data.
- Without $\mathcal{L}_{mlm}$ in $\mathcal{L}_{LNP}$, RoMo loses the ability to address cross-modal heterogeneity, resulting in awfully poor performance in 2D-3D cross-modal retrieval tasks.

Furthermore, to assess the efficacy of memory banks in instance-level discrimination, we delve deeper into their impact through ablation experiments. However, attempting

TABLE VI

Ablation Study in Terms of mAP Scores on 3D MNIST and ModelNet10 Dataset. The Number in (∗) Indicates the Number of Warm-up Epochs. The Best Results Are Highlighted in **Bold**

$\mathcal{L}_{PLA}$		$\mathcal{L}_{LNP}$		3D MNIST				ModelNet10							
				2D → 3D		3D → 2D		2D → 3D				3D → 2D			
$\mathcal{L}_{ssm}$	$\mathcal{L}_{mlm}$	$\mathcal{L}_{rcll}$	$\mathcal{L}_{mlm}$	RP	GP	PR	PG	RP	RM	GP	GM	PR	PG	MR	MG
✓		✓		0.107	0.105	0.099	0.105	0.104	0.104	0.120	0.101	0.121	0.127	0.120	0.118
✓			✓	0.775	0.768	0.779	0.771	0.712	0.707	0.721	0.733	0.723	0.740	0.739	0.730
	✓	✓		0.128	0.106	0.105	0.119	0.124	0.129	0.102	0.117	0.130	0.106	0.130	0.123
	✓		✓	0.767	0.773	0.798	0.752	0.794	0.756	0.794	0.791	0.796	0.793	0.802	0.791
	✓	✓	✓	0.776	0.766	0.772	0.762	0.797	0.781	0.747	0.770	0.722	0.774	0.747	0.801
✓		✓	✓	0.815	0.802	0.809	0.811	0.808	0.793	0.802	0.782	0.817	0.770	0.789	0.808
✓	✓		✓	0.787	0.742	0.774	0.726	0.826	0.811	0.819	0.807	0.816	0.791	0.804	0.805
✓	✓	✓		0.122	0.117	0.121	0.114	0.132	0.125	0.113	0.143	0.118	0.115	0.129	0.117
✓	✓	✓	✓	**0.849**	**0.843**	**0.836**	**0.839**	**0.829**	**0.812**	**0.848**	**0.811**	**0.832**	**0.830**	**0.813**	**0.813**

TABLE VII

Ablation Study in Terms of mAP Scores on MI3DOR and ModelNet40 Dataset. The Number in (∗) Indicates the Number of Warm-up Epochs. The Best Results Are Highlighted in **Bold**

$\mathcal{L}_{PLA}$		$\mathcal{L}_{LNP}$		MI3DOR				ModelNet40							
				2D → 3D		3D → 2D		2D → 3D				3D → 2D			
$\mathcal{L}_{ssm}$	$\mathcal{L}_{mlm}$	$\mathcal{L}_{rcll}$	$\mathcal{L}_{mlm}$	RM	GM	MR	MG	RP	RM	GP	GM	PR	PG	MR	MG
✓		✓		0.064	0.059	0.058	0.049	0.047	0.039	0.034	0.049	0.037	0.041	0.032	0.048
✓			✓	0.607	0.626	0.611	0.628	0.717	0.708	0.710	0.727	0.726	0.726	0.713	0.721
	✓	✓		0.053	0.053	0.063	0.062	0.014	0.053	0.036	0.030	0.036	0.039	0.039	0.051
	✓		✓	0.611	0.656	0.610	0.655	0.736	0.727	0.727	0.733	0.720	0.732	0.732	0.723
	✓	✓	✓	0.617	0.695	0.643	0.657	0.764	0.772	0.762	0.775	0.742	0.733	0.767	0.775
✓		✓	✓	0.719	0.689	0.669	0.693	0.789	0.776	0.769	0.751	0.731	0.711	0.718	0.775
✓	✓		✓	0.661	0.659	0.656	0.677	0.715	0.704	0.757	0.733	0.743	0.732	0.695	0.713
✓	✓	✓		0.055	0.053	0.052	0.049	0.031	0.046	0.043	0.035	0.051	0.044	0.051	0.033
✓	✓	✓	✓	**0.744**	**0.719**	**0.721**	**0.705**	**0.810**	**0.816**	**0.810**	**0.785**	**0.801**	**0.811**	**0.780**	**0.801**

TABLE VIII

Ablation Study of Memory Banks in Terms of Average mAP Scores on 3D MNIST, ModelNet10, MI3DOR, and ModelNet40 Datasets. The Best Results Are Highlighted in **Bold**

$\mathcal{M}^{\mathcal{X}}$	$\mathcal{M}^{\mathcal{Y}}$	3D MNIST	ModelNet10	MI3DOR	ModelNet40
		0.764	0.762	0.620	0.741
✓		0.788	0.794	0.667	0.752
	✓	0.769	0.782	0.662	0.749
✓	✓	**0.842**	**0.824**	**0.722**	**0.802**

TABLE IX

Ablation Study of Training Time (in Seconds) and Memory Consumption (in Mb) With Memory Banks (Abbreviated as MB) on Various Datasets

	3D MNIST	ModelNet10	MI3DOR	ModelNet40
w/o MB(Time(s))	7.3×10^{-4}	6.3×10^{-4}	6.7×10^{-4}	7.1×10^{-4}
MB(Time(s))	1.3×10^{-1}	9.0×10^{-2}	2.6×10^{-1}	1.1×10^{-1}
MB(Memory(Mb))	19.531	9.641	38.437	15.031

to remove the memory banks directly would lead to the loss of proper optimization references, thereby introducing convergence difficulties. To trackle this problem, we substitute the reference with features from the previous epoch when calculating the loss. Here, in the first epoch, we employ the same clustering method as the proposed approach to acquire initial features. These steps are executed on both the 2D and 3D modalities. It is clear from Table VIII that the memory banks yield superior performance. We believe that memory banks are continuously iterated and updated based on the clustering results in the optimization process, rendering them less vulnerable to random factors and preventing pattern collapse. The features in the memory banks can gradually approach the class centroid and serve as different views of the sample, which makes the model encapsulate discrimination into the representation in a self-supervised learning manner.

Last but not least, it is worth noting that the utilization of the memory bank has minimal effects on computational efficiency. This is primarily attributed to the insignificant update time of the memory bank when compared to the overall duration required for training the multimodal model. As corroborated by the findings presented in Table IX, the update time typically falls within the level of milliseconds. It is also not particularly large in terms of memory consumption, thus not consume too much video card resources in the training process.

F. Parameter Sensitivity Analysis

In this section, we investigate the impact of the trade-off parameters λ_1 and λ_2 on the retrieval performance. To achieve this, we keep the other hyper-parameters constant and focus on tuning λ_1 and λ_2 through a grid search respectively. After the grid search, we fix λ_2 searching for the optimal solution of λ_1, and then fix λ_1 searching for the optimal solution of λ_2, as shown in Fig. 4. The search involves exploring values within the range of $[0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1]$. It is easy to observe that λ_1 and λ_2 yield better results when they are around 0.25. Additionally, as the value of λ_2 approaches 1, it significantly impacts the performance. This indirectly confirms the significance of addressing the 2D-3D heterogeneity gap in the T^{2}UCR task.

G. Robustness Analysis

To illustrte the influence of α on robustness, we conducted a series of experiments and shown in Fig. 5, where MAE corresponds to the case where $\alpha = 0$. From these figures, one can draw the following conclusions: 1) Smaller values of α (indicating consistent and smooth gradients for all samples) are associated with better robustness against interference from noisy labels. 2) Conversely, larger values of α (indicating gradients with a degree of discrimination for clean and noisy samples) are more prone to overfitting false supervision.

Moreover, the mAP curves demonstrate the impact of α on balancing between underfitting and overfitting across the experimental range. Based on experimental studies, α is recommended to be set as $[0.3, 0.25, 0.35, 0.3]$ for 3D MNIST, ModelNet10, MI3DOR, and ModelNet40, respectively.

V. Conclusion and Future Works

In this paper, we present a novel unsupervised 2D-3D retrieval framework RoMo. Our RoMo first utilizes an SSM to enhance the ability of model to incorporate discrimination through self-supervised learning. Then, RDL is employed to extract discrimination from the learned imperfect predictions. Moreover, MLM is employed to reduce cross-modal discrepancies and encourage SSM and RDL to produce common representations. Finally, the proposed RoMo is superior to 14 state-of-the-art unsupervised multimodal learning methods within T^{2}UCR. In future works, we plan to explore further the retrieval frameworks in the open world T^{2}UCR and provide valuable analysis for various practical retrieval applications.

References

[1] Z. Wu et al., "3D ShapeNets: A deep representation for volumetric shapes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1912–1920.

[2] W. Li et al., "SHREC 2019-monocular image based 3D model retrieval," in *Proc. 12th Eurograph. Workshop 3D Object Retr.*, 2019, pp. 1–8.

[3] N. C. Mithun, K. Sikka, H.-P. Chiu, S. Samarasekera, and R. Kumar, "RGB2LiDAR: Towards solving large-scale cross-modal visual localization," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 934–954.

[4] Y. Chen, F. Liu, and K. Pei, "Cross-modal matching CNN for autonomous driving sensor data monitoring," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 3110–3119.

[5] Y. Feng, H. Zhu, D. Peng, X. Peng, and P. Hu, "RONO: Robust discriminative learning with noisy labels for 2D-3D cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 11610–11619.

[6] M.-X. Lin et al., "Single image 3D shape retrieval via cross-modal instance and category contrastive learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11405–11415.

[7] S. Yu, H. Han, S. Shan, and X. Chen, "CMOS-GAN: Semi-supervised generative adversarial model for cross-modality face image synthesis," *IEEE Trans. Image Process.*, vol. 32, pp. 144–158, 2023.

[8] Y. Sun, X. Wang, D. Peng, Z. Ren, and X. Shen, "Hierarchical hashing learning for image set classification," *IEEE Trans. Image Process.*, vol. 32, pp. 1732–1744, 2023.

[9] X. Wang, P. Hu, L. Zhen, and D. Peng, "DRSL: Deep relational similarity learning for cross-modal retrieval," *Inf. Sci.*, vol. 546, pp. 298–311, Feb. 2021.

[10] P. Hu, H. Zhu, X. Peng, and J. Lin, "Semi-supervised multi-modal learning with balanced spectral decomposition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 99–106.

[11] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.

[12] M. Jaritz, T.-H. Vu, R. de Charette, E. Wirbel, and P. Pérez, "xMUDA: Cross-modal unsupervised domain adaptation for 3D semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12605–12614.

[13] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2075–2082.

[14] J. Zhang, Y. Peng, and M. Yuan, "Unsupervised generative adversarial cross-modal hashing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, 2018, pp. 1–8.

[15] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 132–149.

[16] J. Xie, Y. Xu, Z. Zheng, S.-C. Zhu, and Y. N. Wu, "Generative PointNet: Deep energy-based learning on unordered point sets for 3D generation, reconstruction and classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14976–14985.

[17] S. Zheng and M. Castellani, "Primitive shape recognition from real-life scenes using the PointNet deep neural network," *Int. J. Adv. Manuf. Technol.*, vol. 124, no. 9, pp. 3067–3082, 2023.

[18] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 652–660.

[19] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph CNN for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, pp. 1–12, 2019.

[20] Y. Feng, Y. Feng, H. You, X. Zhao, and Y. Gao, "MeshNet: Mesh neural network for 3D shape representation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8279–8286.

[21] C. Deng, X. Xu, H. Wang, M. Yang, and D. Tao, "Progressive cross-modal semantic network for zero-shot sketch-based image retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 8892–8902, 2020.

[22] H. Liu, R. Ji, Y. Wu, F. Huang, and B. Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7380–7388.

[23] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[24] P. Hu, H. Zhu, J. Lin, D. Peng, Y.-P. Zhao, and X. Peng, "Unsupervised contrastive cross-modal hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3877–3889, Mar. 2023.

[25] W. Chen, Y. Liu, E. M. Bakker, and M. S. Lew, "Integrating information theory and adversarial learning for cross-modal retrieval," *Pattern Recognit.*, vol. 117, Sep. 2021, Art. no. 107983.

[26] N. Nigam, T. Dutta, and H. P. Gupta, "Impact of noisy labels in learning techniques: A survey," in *Advances in Data and Information Sciences: Proceedings of ICDIS 2019*. Singapore: Springer, 2020, pp. 403–411.

[27] B. Han et al., "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[28] J. Li, R. Socher, and S. C. H. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," in *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–14.

[29] J. Zhang, X. Wu, and V. S. Sheng, "Learning from crowdsourced labeled data: A survey," *Artif. Intell. Rev.*, vol. 46, pp. 543–576, Jul. 2016.

[30] A. J. Bekker and J. Goldberger, "Training deep neural-networks based on unreliable labels," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Feb. 2016, pp. 2682–2686.

[31] H. Sun, C. Guo, Q. Wei, Z. Han, and Y. Yin, "Learning to rectify for robust learning with noisy labels," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108467.

[32] A. Ghosh, H. Kumar, and P. S. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 1–7.

[33] D. Arpit et al., "A closer look at memorization in deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 233–242.

[34] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3733–3742.

[35] A. Singh, "CLDA: Contrastive learning for semi-supervised domain adaptation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 5089–5101.

[36] Y.-H. Hubert Tsai, Y. Wu, R. Salakhutdinov, and L.-P. Morency, "Self-supervised learning from a multi-view perspective," 2020, *arXiv:2006.05576*.

[37] P. Hu, X. Peng, H. Zhu, L. Zhen, and J. Lin, "Learning cross-modal retrieval with noisy labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5403–5413.

[38] X. Xu, A. Dehghani, D. Corrigan, S. Caulfield, and D. Moloney, "Convolutional neural network for 3D object recognition using volumetric representation," in *Proc. 1st Int. Workshop Sens., Process. Learn. Intell. Mach. (SPLINE)*, Jul. 2016, pp. 1–5.

[39] D. Song, W.-Z. Nie, W.-H. Li, M. Kankanhalli, and A.-A. Liu, "Monocular image-based 3-D model retrieval: A benchmark," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 8114–8127, Aug. 2022.

[40] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics: Methodology and Distribution*. New York, NY, USA: Springer, 1992, pp. 162–190.

[41] J. Rupnik and J. Shawe-Taylor, "Multi-view canonical correlation analysis," in *Proc. Conf. Data Mining Data Warehouses*, 2010, pp. 1–4.

[42] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[43] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.

[44] J. Wang, W. Feng, Y. Chen, H. Yu, M. Huang, and P. S. Yu, "Visual domain adaptation with manifold embedded distribution alignment," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 402–410.

[45] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3027–3035.

[46] S. Liu, S. Qian, Y. Guan, J. Zhan, and L. Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1379–1388.

[47] J. Yu, H. Zhou, Y. Zhan, and D. Tao, "Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 4626–4634.

[48] Y. Liu, J. Wu, L. Qu, T. Gan, J. Yin, and L. Nie, "Self-supervised correlation learning for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 25, pp. 2851–2863, 2022.

[49] J. Li, F. Li, L. Zhu, H. Cui, and J. Li, "Prototype-guided knowledge transfer for federated unsupervised cross-modal hashing," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 1013–1022.

[50] L. Mingyong, L. Yewen, G. Mingyuan, and M. Longfei, "CLIP-based fusion-modal reconstructing hashing for large-scale unsupervised cross-modal retrieval," *Int. J. Multimedia Inf. Retr.*, vol. 12, no. 1, p. 2, Jun. 2023.

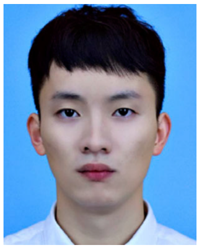
[51] Y. Wei et al., "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.

[52] X. Zhou, X. Liu, J. Jiang, X. Gao, and X. Ji, "Asymmetric loss functions for learning with noisy labels," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 12846–12856.

[53] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[54] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9729–9738.

Yongxiang Li received the master's degree in electronics and communication engineering from the College of Electronic Information, Sichuan University, Chengdu, China, in 2022. He is currently pursuing the Ph.D. degree in software engineering with the College of Computer Science, Sichuan University. His research interests include image processing and multimodal learning.



Yang Qin (Graduate Student Member, IEEE) received the bachelor's degree in software engineering from Sichuan University, Chengdu, China, in July 2021, where he is currently pursuing the Ph.D. degree with the College of Computer Science. His research interests include multi-modal learning and learning with noisy correspondence.

Yuan Sun is currently pursuing the Ph.D. degree with the College of Computer, Sichuan University, Chengdu, China. He has published more than ten papers in highly regarded journals and conferences, such as IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON MULTIMEDIA, AAAI, IJCAI, and ACM MM. His research interests include image set classification, cross-modal retrieval, and multi-view learning.

Dezhong Peng (Senior Member, IEEE) received the B.Sc. degree in applied mathematics and the M.Sc. and Ph.D. degrees in computer software and theory from the University of Electronic Science and Technology of China, Chengdu, China, in 1998, 2001, and 2006, respectively. From 2001 to 2007, he was with the University of Electronic Science and Technology of China as an Assistant Lecturer and a Lecturer. He was a Post-Doctoral Research Fellow with the School of Engineering, Deakin University, Geelong, VIC, Australia, from 2007 to 2009. He is currently a Professor with the College of Computer Science, Sichuan University, Chengdu, China. His research interests include neural networks and signal processing.

Xi Peng (Senior Member, IEEE) is currently the Cheung Kong Distinguished Professor with the College of Computer Science, Sichuan University. His current research interests include machine learning, multi-media analysis, and AI4Science. In these areas, he has co-authored around 100 articles in *Nature Communications*, JMLR, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, ICML, and NeurIPS.

Peng Hu received the Ph.D. degree in computer science and technology from Sichuan University, China, in 2019. He is currently an Associate Research Professor with the College of Computer Science, Sichuan University. His research interests include multi-view learning, cross-modal retrieval, and network compression. In these areas, he has authored more than 40 papers in top-tier conferences and journals.