Learning Cross-Modal Retrieval with Noisy Labels

Peng Hu[1,2] Xi Peng[1*] Hongyuan Zhu[2] Liangli Zhen[3] Jie Lin[2]
[1]College of Computer Science, Sichuan University, Chengdu 610065, China
[2]Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore
[3]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

Abstract

Recently, cross-modal retrieval is emerging with the help of deep multimodal learning. However, even for unimodal data, collecting large-scale well-annotated data is expensive and time-consuming, and not to mention the additional challenges from multiple modalities. Although crowd-sourcing annotation, e.g.*, Amazon's Mechanical Turk, can be utilized to mitigate the labeling cost, but leading to the unavoidable noise in labels for the non-expert annotating. To tackle the challenge, this paper presents a general Multimodal Robust Learning framework (MRL) for learning with multimodal noisy labels to mitigate noisy samples and correlate distinct modalities simultaneously. To be specific, we propose a Robust Clustering loss (RC) to make the deep networks focus on clean samples instead of noisy ones. Besides, a simple yet effective multimodal loss function, called Multimodal Contrastive loss (MC), is proposed to maximize the mutual information between different modalities, thus alleviating the interference of noisy samples and cross-modal discrepancy. Extensive experiments are conducted on four widely-used multimodal datasets to demonstrate the effectiveness of the proposed approach by comparing to 14 state-of-the-art methods.*

1. Introduction

With rapid growth of multimedia data, cross-modal retrieval becomes a compelling topic in the multimodal learning community due to its flexibility in retrieving semantically relevant samples across distinct modalities, *e.g.*, image query text [6, 16]. However, most existing methods require clean-annotated training data, which are expensive and time-consuming. Although some unsupervised multimodal learning methods can mitigate such labeling pressure, their performance is usually much worse than the supervised counterparts' [60]. To balance performance and labeling cost, semi-supervised multimodal learning methods are proposed to simultaneously utilize labeled and unlabeled data to learn common discriminative representations [61, 17]. However, semi-supervised approaches still require a certain number of clean-annotated data to reach reasonable performance.

To alleviate the high labeling cost, some non-expert sources, *e.g.*, Amazon's Mechanical Turk and the surrounding tags of collected data, can be used to annotate large-scale data, but resulting in unavoidable noise in labels [48]. Some recent unimodal studies reveal that DNNs easily overfit to noisy labels leading to poor generalization performance [59, 28]. It is challenging to learn with noisy labels. To tackle the challenge, numerous studies are conducted to explore how to robustly learn with noisy labels, such as correction methods [49, 9], MentorNet [19, 58], and Co-teaching [10]. Although they achieve promising performance in the unimodal scenario, they cannot simultaneously tackle multiple modalities, such as real-world multimedia data. Hence, it is significant and valuable to explore how to learn satisfactory representations from multimodal data with noisy labels, but which is rarely touched in previous works.

We perform an empirical study of recent cross-modal learning methods under noisy labels with results shown in Figure 2. From the figure, one can see that the networks will fast overfit to the noisy training set with a widely-used loss function cross-entropy [50, 53] in multimodal learning. Moreover, different modalities exist a large diversity in validation set since they may lay in completely different spaces with heterogeneity, making learning from noisy samples more difficult. Lastly, noisy labels can confuse the discriminative connections across distinct modalities, resulting in difficulty bridging the heterogeneous gap. Thus, it is more challenging and complex to consider both noisy labels and cross-modal discrepancy simultaneously.

To address the aforementioned problems, we propose a Multimodal Robust Learning framework (MRL) to simultaneously mitigate the influence of noisy samples and narrow the heterogeneous gap in this paper. The pipeline of the proposed method is shown in Figure 1 wherein our

*Corresponding author: Xi Peng (pengx.gm@gmail.com).

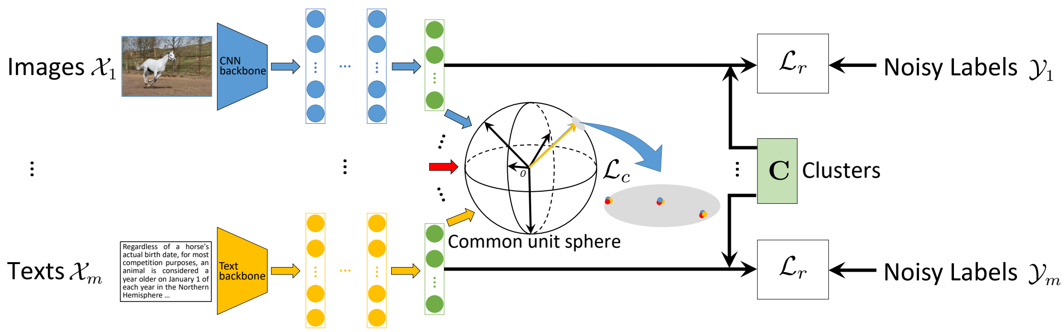
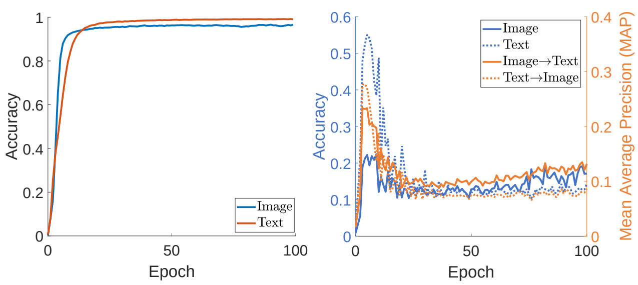


Figure 1: The pipeline of the proposed method for m modalities, *e.g.*, images $\mathcal{X}_1$ with noisy labels $\mathcal{Y}_1$, and texts $\mathcal{X}_m$ with noisy labels $\mathcal{Y}_m$. The modality-specific networks learn common representations for m different modalities. The Robust Clustering loss $\mathcal{L}_r$ is adopted to mitigate the noise in labels for learning discrimination and narrow the heterogeneous gap. The outputs of networks interact with each other to learn common representations by using instance- and pair-level contrast, *i.e.*, multimodal contrastive learning ($\mathcal{L}_c$), thus further mitigating noisy labels and cross-modal discrepancy. $\mathcal{L}_c$ tries to maximally scatter inter-modal samples while compacting intra-modal points over the common unit sphere/space.



(a) (b)

Figure 2: Training with Cross-Entropy loss (CE) [53] on INRIA-Websearch [23] under 0.6 symmetric noise. (a) Accuracy *vs*. epoch on the training set of INRIA-Websearch for the image modality and the text modality, respectively. (b) Accuracy/MAP *vs*. epoch on the validation set of INRIA-Websearch. Accuracy is utilized to evaluate the classification performance on the individual modality. Mean Average Precision (MAP) is adopted to evaluate the retrieval performance across different modalities, *i.e.*, image query text (Image → Text) and text query image (Text → Image). From the figure, we can see that noisy labels will make the multimodal learning overfit on the noisy training set while corrupting performance on the validation set.

method consists of multiple modality-specific networks and two novel losses: Robust Clustering (RC) and Multimodal Contrastive (MC) losses. To be specific, we present a novel common clustering loss to alleviate traditional classification loss functions (*e.g.*, cross-entropy) overfitting to noisy labels with a common classifier. From the previous studies [2, 10, 3], clean samples are easier for learning than noisy/incorrect samples, and leading to faster learning and lower loss for clean samples. This similar phenomenon can be observed in Figure 2, wherein the networks can faster learn clean samples and achieve a certain accuracy but decreasing the performance with the interference of noisy samples after further training. To distract the attention of deep networks from noisy samples to clean ones, our RC automatically weakens the influence of minor losses, which are more likely to be produced by noisy samples, for alleviating the interference of noisy labels, thus embracing more robustness. In addition to mitigating noisy samples, RC also can narrow the heterogeneous gap by projecting different modalities into a common clustering space. Besides, inspired by recent unimodal contrastive learning works [54, 4], we propose a simple yet effective multimodal loss function, termed Multimodal Contrastive loss (MC), to simultaneously maximize the inter-instance and inter-pair variances in the intra- and inter-modalities. Different from previous contrastive learning methods, our MC maximizes the mutual information between intrinsically co-occurred modalities, which can further narrow the heterogeneous gap across distinct modalities while excavating the discrimination from instance-level contrasting. Hence, our MC can further mitigate the interference of noisy samples by contrasting their inter- and intra-modal counterparts with an unsupervised manner.

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

- We propose a novel framework for cross-modal retrieval with noisy labels. It can robustly learn the common discriminative representations from noisy labels by using both supervised and unsupervised manners.
- We present a Robust Clustering loss (RC) that improves robustness and narrows the cross-modal gap on noisy samples simultaneously.

- A novel multimodal contrastive loss is proposed to maximize the inter-instance variances while minimizing the intra-instance ones by considering the inter- and intra-modality similarities.
- Extensive experiments are conducted on four widely-used multimodal datasets to demonstrate the robust performance of the proposed methods for noisy labels.

2. Related Work

This section briefly reviews some of the most related works about learning with noisy labels and multimodal learning approaches.

2.1. Learning with Noisy Labels

To learn from noisy labels, numerous approaches are proposed to alleviate the noise in labels to learn the objective information. One typical direction is to improve the learning quality by correcting the wrong labels or loss functions, called correction methods [30, 49, 9]. However, such approaches always require extra ground-truth to support their learning schemes, which is often unavailable and cost-prohibitive in real-world applications [32, 48]. To avoid false corrections, numerous works attempt to elaborately design adaptive training strategies to select true-labeled samples for learning automatically, thus embracing robustness to noisy labels, such as MentorNet [19, 58], and Co-teaching [10]. Moreover, some methods try to divide the noisy data into labeled and unlabeled data, while utilizing semi-supervised paradigms to learn from the obtained labeled and unlabeled data iteratively [56, 35, 28]. Motivated by its powerful learning ability, meta learning has been successfully applied to improve the robustness of neural networks for noise samples [42, 29, 46]. The aforementioned approaches are either dependent on complex adaptive training processes that need carefully tuning and designing or sensitive to the hyper-parameters that will cost much time for tuning [32]. Differently, another direction is to design robust loss functions to make the optimization schemes robust to noise samples [8, 52, 32]. However, most prior arts for noisy labels are specifically designed for unimodal scenarios, and it is challenging to extend them to multimodal cases.

2.2. Multimodal Learning

Multimodal learning methods target to project multiple modalities into a common space, in which cross-modal downstream tasks could be conducted on the learned common representations, such as cross-modal retrieval [45, 16]. One typical technique is to maximize the cross-modal correlation across different modalities [12, 26, 38, 63, 18]. To utilize the semantic information in class labels, some supervised multimodal methods are proposed to utilize the discrimination to learn a common discriminative space. Specifically, discriminative criteria are introduced into multimodal learning to maximize the within-class similarity while minimizing the between-class similarity [21, 13, 27, 14]. Alternatively, a common classifier is directly adopted to enforce the neural networks to learn a common discriminative space [53, 62, 16, 55, 15]. To alleviate the over-dependence on labels, some multimodal semi-supervised paradigms are proposed to leverage the labeled and unlabeled to learn common representations simultaneously [61, 17, 57]. Moreover, to clean the noise from labels, Mandal *et al*. adopted a two-step pre-processing method to obtained cleaned labels and feed to cross-modal methods [33]. However, it is much more difficult to directly learn common discrimination from noisy labels, which is rarely touched in previous studies.

3. The Proposed Method

3.1. Notations

For a clear presentation, we first give some definitions for notations in the papers. Boldface uppercase letters (*e.g.*, $\mathbf{X}$) and boldface lowercase letters (*e.g.*, $\mathbf{x}$) represent matrices and column vectors, respectively. Give a K-category multimodal dataset with noisy labels as $\mathcal{D} = \{\mathcal{M}_i\}_{i=1}^m$, where $\mathcal{M}_i = \{(\mathbf{x}_j^i, y_j^i)\}_{j=1}^N$ is the i-th modality, $\mathbf{x}_j^i \subset \mathbb{R}^{d_i}$ is the j-th sample from the i-th modality, $y_j^i \in \{1, 2, \cdots, K\}$ is the label (possibly incorrect) for $\mathbf{x}_j^i$. For the convenience of presentation, $\mathcal{D}$ can be seen as a minibatch of N instances, each of which owns m samples from distinct modalities, in the following sections. Note that, although different modalities usually co-occur to describe the same objects or instances with the pairwise property, the noisy labels of distinct modalities may be not be paired, *e.g.*, there are different annotators for the image modality and the text modality separately.

3.2. Multimodal Robust Learning

The goal of cross-modal retrieval is to retrieve the correlated samples across different modalities in a common representation space $\mathcal{Z}$. To project distinct modalities into $\mathcal{Z}$, existing methods attempt to learn m modality-specific functions $\{f_i : \mathcal{X}_i \mapsto \mathcal{Z}\}_{i=1}^m$ for m modalities, where f_i can be a DNN with parameters Θ_i for the i-th modality. Given a data point $\mathbf{x}_j^i$, the common normalized representation $\mathbf{z}_j^i$ can be computed by

$$\mathbf{z}_j^i = f_i(\mathbf{x}_j^i) \subset \mathbb{R}^L, \tag{1}$$

where L is the dimention of the common space.

To learn the mapping functions $\{f_i(\cdot, \Theta_i)\}_{i=1}^m$ with noisy labels, we propose a general framework that consists of a robust clustering loss and a cross-modal correlation loss. A novel robust clustering loss, called Robust Clustering loss (RC) $\mathcal{L}_r$, is proposed to robustly extract the common discrimination shared across different modalities from noisy labels. Specifically, it can simultaneously alleviate the cross-modal discrepancy and noisy samples.

To further narrow the heterogeneous gap, we propose a novel cross-modal correlation loss, termed Multimodal Contrastive loss (MC) $\mathcal{L}_c$, to excavate the instance- and pair-level discrimination to boost the performance of cross-modal retrieval. Following sections will elaborate the aforementioned loss functions with details.

3.2.1 Robust Clustering Assignment

To excavate the discrimination from noisy labels, we propose to firstly cluster multimodal data to K common class-specific clustering assignments $\mathbf{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$ for alleviating outliers, where $\mathbf{c}_k$ is the normalized clustering center vector for the k-th class. Then, for a sample $\mathbf{x}_j^i$, its probability belonging to the k-th class can be defined as:

$$p(k|\mathbf{x}_j^i) = \frac{\exp(\frac{1}{\tau_1}\mathbf{c}_k^T\mathbf{z}_j^i)}{\sum_{t=1}^{K}\exp(\frac{1}{\tau_1}\mathbf{c}_t^T\mathbf{z}_j^i)}, \tag{2}$$

where τ_1 is a temperature parameter [54, 34]. By maximizing the joint probabilities of Equation (2) with the ground-truths, the multimodal points with the same semantics could be compacted into the same cluster in the common space [37], which could be directly achieved by the standard Cross-Entropy criterion (CE) as:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{K} q(k|\mathbf{x}_j^i)\log p(k|\mathbf{x}_j^i), \tag{3}$$

where $q(k|\mathbf{x}_j^i)$ is the ground-truth probability over different labels of sample $\mathbf{x}_j^i$. In this work, we only focus on single-label case wherein each sample $\mathbf{x}_j^i$ only belongs to one class y_j^i, *i.e.*, $\mathbf{q}$ is simply a one-hot label vector defined as follows:

$$q(k|\mathbf{x}_j^i) = \begin{cases} 1 & \text{if } k = y_j^i; \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

For convenience of presentation, cross-entropy loss $\mathcal{L}_{CE}$ can be rewritten as $\mathrm{CE}(p) = \mathrm{CE}(p, y) = -log(p)$ for a sample $\mathbf{x}_j^i$ with its ground-truth y, where $p = p(y_j^i|\mathbf{x}_j^i)$ and $y = y_j^i$. Then, the CE loss could be plotted as the blue curve in Figure 3. From the figure, one can see that CE tends to focus on optimizing harder samples since they producing larger loss values, which has been proven to work well in clean-annotated labels. However, for noisy labels, recent studies [2, 10, 3] reveal that clean/correct samples are easier for learning than noisy/incorrect ones, and leading to faster learning for clean ones as shown in Figure 2. Like CE, conventional supervised loss functions usually put more attentions on harder samples that possibly are noisy ones in noisy-label cases [32, 28], thus leading to worse performance on noisy datasets.

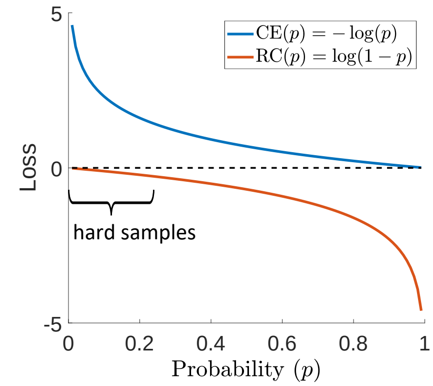


Figure 3: Comparison between the standard Cross-Entropy criterion (CE) and the proposed Robust Clustering loss (RC). The proposed RC is to reduce the relative loss of CE for hard samples that usually are with noisy/incorrect labels [2, 10, 3]. Thus, our RC can put more focus on clean samples instead of noisy ones and mitigate noise interference.

To further alleviate the influence of noisy samples, we propose to reshape the loss function to down-weight difficult samples and up-weight easy ones and thus focus training on clean samples instead of noisy ones. Different from CE loss, we minimize the log-likelihood of negative samples instead of the negative log-likelihood of the positive one. Our RC can be formulated as:

$$\mathcal{L}_r = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{N}\log(1 - p(y_j^i|\mathbf{x}_j^i)). \tag{5}$$

Note that, the target of our RC is opposite to the fully supervised losses, such as Focal Loss [31], which mainly focuses on hard samples from a clean dataset with no noisy labels.

The learning process of the RC function is visualized in Figure 3 by comparing it with the standard cross-entropy loss. From the figure, one can note that our RC loss can remarkably down-weight the loss values of noisy samples while up-weight the values of clean ones. That is to say, the clean samples will produce much larger loss values than noisy ones, and dominate the gradient into the correct direction, thus embracing superior performance as evidenced in our experimental results.

3.2.2 Multimodal Contrastive Learning

In the unimodal scenario, contrastive learning has achieved promising performance in unsupervised learning applications [54, 4]. Unlike supervised methods, contrastive learning approaches apply instance-level classification to learn the discriminative representations. Specifically, to achieve that, it maximizes the agreement between different augmented views from the same sample via a contrastive loss [54, 4]. Instead of augmentation, multimodal data intrinsically consist of multiple modalities that can be naturally utilized to maximize their mutual information. Inspired by such recent contrastive learning works [54, 4], we present to learn common representations by maximizing agreement between different modalities in the common space.

First, we define the probability of a sample $\mathbf{x}_j^i$ belonging to the j-th instance in m modalities as:

$$P(j|\mathbf{x}_j^i) = \frac{\sum_{l=1}^{m} \exp\left(\frac{1}{\tau_2}\left(\mathbf{z}_j^l\right)^T \mathbf{z}_j^i\right)}{\sum_{l=1}^{m}\sum_{t=1}^{N} \exp\left(\frac{1}{\tau_2}\left(\mathbf{z}_t^l\right)^T \mathbf{z}_j^i\right)}, \tag{6}$$

where τ_2 is a temperature parameter [54, 34]. $P(j|\mathbf{x}_j^i)$ also could be seen as the probability of a multimodal cluster centered by $\mathbf{x}_j^i$ being correctly recognized as the j-th cluster.

To eliminate the cross-modal discrepancy and excavate the instance-level discrimination, we make the multimodal samples from the same instance (*e.g.*, $\{\mathbf{x}_j^k\}_{k=1}^m$ for the j-th instance) compact while the samples from distinct instances (*e.g.*, $\{\mathbf{x}_l^k\}_{l\neq j}$ for the j-th instance) scattered, *i.e.*, maximizing the probabilities. Then, the learning objective MC could be formulated as maximizing a joint probability $\prod_{i=1}^{m}\prod_{j=1}^{N} P(j|\mathbf{x}_j^i)$, which also can be equivalent to minimize the following negative log-likelihood [54] as:

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{N} \log\left(P(j|\mathbf{x}_j^i)\right). \tag{7}$$

By minimizing Equation (7), the multimodal networks are enforced to compact the positive samples (considered as the relevant cross-modal pairs, *e.g.*, $\{\mathbf{x}_j^k\}_{k=1}^m$ for $\mathbf{x}_j^i$) while scattering the negative samples (considered as irrelevant instances, *e.g.*, $\{\mathbf{x}_l^k\}_{l\neq j}$ for $\mathbf{x}_j^i$) on the common unit sphere/space as shown in Figure 1. The outputs of multimodal networks contrast with each other to encapsulate the common discrimination by using the instance- and pair-level comparison, thus further mitigating noisy labels and cross-modal discrepancy.

3.2.3 Optimization

The final loss function can be formulated as:

$$\mathcal{L} = \beta\mathcal{L}_r + (1-\beta)\mathcal{L}_c. \tag{8}$$

By minimizing the joint loss function, our MRL can be iteratively optimized in a batch-by-batch manner using a stochastic gradient descent optimization algorithm, like Adam [22]. Algorithm 1 summarizes the optimization process of our MRL.

Algorithm 1 Main optimization process of our MRL.

Input: The training multimodal data $\mathcal{D} = \{\mathcal{M}_i\}_{i=1}^m$, the dimensionality of common representations L, batch size N_b, maximal epoch number N_e, balance parameter β, temperature parameters τ_1 and τ_2, and learning rate α.

for $1, 2, \cdots, N_e$ **do**
  repeat
    Randomly select N_b samples from each modality to construct a multimodal mini-batch $\{\mathcal{X}_i, \mathcal{Y}_i\}_{i=1}^m$.
    Calculate the representations for all samples of the mini-batch by using their corresponding modality-specific mapping functions $\{f_i(\cdot, \Theta_i)\}_{i=1}^m$, according to Equation (1).
    Normalize the clusters $\mathbf{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$.
    Compute RC and MC according to Equations (5) and (7) on the mini-batch, respectively.
    Update the network parameters $\{\Theta_i\}_{i=1}^m$ and clusters $\mathbf{C}$ by minimizing $\mathcal{L}$ in Equation (8) with descending their stochastic gradient:
    $\Theta_i = \Theta_i - \alpha(\beta\frac{\partial\mathcal{L}_r}{\partial\Theta_i} + (1-\beta)\frac{\partial\mathcal{L}_c}{\partial\Theta_i}), i = 1, \cdots, m$
    $\mathbf{C} = \mathbf{C} - \alpha\beta\frac{\partial\mathcal{L}_r}{\partial\mathbf{C}}$
  until all samples selected
end for

Output: Optimized network parameters $\{\Theta_i\}_{i=1}^m$.

4. Experiments

To evaluate our MRL, we conduct extensive comparison experiments on four widely-used multimodal datasets, *i.e.*, Wikipedia [41], INRIA-Websearch [23], NUS-WIDE [5], and XMediaNet [40].

4.1. Implementation Details

In this work, we employ ADAM [22] as our optimizer to train our MRL. For all datasets, we use a maximal epoch number of 100 (N_e in Algorithm 1). The learning rate (α in Algorithm 1) is initialized with 0.0001. The temperature parameters (τ_1 and τ_2 in Algorithm 1) are set as 1. The batch size is set as 50, 200, 500 and 500 for Wikipedia [41], INRIA-Websearch [23], NUS-WIDE [5], and XMediaNet [40], respectively. For Wikipedia, NUS-WIDE and XMediaNet, we adopt the pretrained VGG-19 [47] on ImageNet as the CNN backbone for images, and the pretrained Doc2Vec model[1] [25] as the text backbone for texts. On INRIA-Websearch, the pretrained AlexNet [24] on ImageNet and LDA are used as the backbones for images and texts, respectively. Three fully-connected (FC) layers are stacked on the backbones to learn the common representations for all modalities. Each FC

[1]https://github.com/jhlau/doc2vec.

layer follows a ReLU layer except the last layer. The numbers of FC hidden units are respectively $4,096$, $4,096$ and L, where L is the dimensionality of common representations. For a fair comparison with baselines, all backbones are frozen in the training stage. Our MRL is implemented on the PyTorch framework [36].

4.2. Experimental Setup

In our experiments, we compare the proposed method with 14 state-of-the-art methods that include four unsupervised methods (*i.e.*, MCCA [43], PLS [44], DCCA [1], and DCCAE [51]) and ten supervised ones (*i.e.*, GMA [45], MvDA [20], MvDA-VC [21], GSS-SL [61], ACMR [50], deep-SM [53], FGCrossNet [11], SDML [16], DSCMR [62], and SMLN [17]). For a fair comparison, all baselines utilize the same features extracted from the corresponding backbones as our MRL. For all methods, we report their results on the testing set when they achieve the best performance on the validation set. To evaluate the performance of these methods, we perform $m(m-1)$ distinct cross-modal retrieval tasks on each dataset. Without loss of generality, all experiments are conducted on bimodal datasets to evaluate two cross-modal tasks: using an image query to retrieve the related text samples (Image $\rightarrow$ Text), and using a text query to retrieve the relevant image points (Text $\rightarrow$ Image). We adopt Mean Average Precision (MAP), which is the mean value of Average Precision (AP) scores for each query, as the evaluation metric to measure the accuracy scores of retrieval results. MAP is extensively adopted to measure the retrieval performance since it simultaneously evaluates retrieval precision and ranking of returned results. Note that, we compute MAP scores on all retrieval results in the experiments. Furthermore, to comprehensively evaluate the robustness of the methods, we set the label noise to be symmetric, and the noise rates to 0.2, 0.4, 0.6 and 0.8 in the experiments.

4.3. Datasets

Without loss of generality, we adopt four widely-used image-text datasets to evaluate the cross-modal performance in the paper. In this section, we briefly introduce them as follows:

- **Wikipedia [41]** contains 2,866 image-text pairs that belong to 10 classes. Following [7], we divide the dataset into 3 subsets: 2,173, 231 and 462 pairs for training, validation and testing sets, respectively.
- **INRIA-Websearch [23]** consists of $71,478$ images and $71,478$ text descriptions (sentences or tags). In our experiments, we use the subset of INRIA-Websearch provided by the authors of [53]. In this subset, $14,698$ samples of 100 largest classes are selected from the original set. We randomly divide the dataset into three subsets: $9,000$, $1,332$ and $4,366$ image-text pairs for training, validation and testing sets, respectively.
- **NUS-WIDE [5]** consists of about $270,000$ images distributed over 81 categories. In the experiments, we use a subset of NUS-WIDE provided by [39], wherein each sample belongs to only one of ten classes. Besides, we randomly split the dataset into three subsets, *i.e.*, $42,941$, $5,000$ and $23,661$ image-text pairs for training, validation and testing sets, respectively.
- **XMediaNet [40]** is a large-scale multimodal dataset with 200 non-overlap categories. In this paper, only images and texts are selected to conduct experiments. Following [40], we randomly divide the dataset into three subsets: $32,000$, $4,000$ and $4,000$ pairs for training, validation and testing sets, respectively.

4.4. Comparison with the State-of-the-Art

We apply cross-modal retrieval on four datasets to evaluate the performance of our MRL and the baselines. The experimental results in terms of MAP scores are reported in Tables 1 and 2 for four datasets, respectively. As shown in these tables, our MRL is superior to the baselines on the four datasets. From the experimental results, we can draw the following observations:

- Some existing multimodal methods (*e.g.*, GMA, SDML, DSCMR, and SMLN) have certain anti-interference ability to noisy labels since their supervised and unsupervised components like our MRL, thus indicating that this framework has more robustness to the noisy labels.
- Noisy labels remarkably influence the performance of supervised multimodal methods. With the noise rate increasing in labels, their accuracies will decrease fast. On the contrary, unsupervised methods have no such issues.
- The number of classes affects the anti-interference performance of supervised methods to noisy labels. Neural networks have a powerful fitting ability, which makes them easier to overfit the harder task (more classes) on noisy labels than shallow methods, thus leading to worse performance. The shallow supervised methods are more robust to noisy labels for more classes than deep supervised approaches.
- Most supervised multimodal approaches are superior to the unsupervised ones in lower noise rate, which evidences that labeled data are important for cross-modal retrieval even though noisy labels contained. Another related observation also can be obtained, *i.e.*, the more pure labels, the better performance obtained.
- Our MRL is superior to both the unsupervised and supervised approaches in the cross-modal retrieval tasks.

Table 1: Performance comparison in terms of MAP scores under the symmetric noise rates of 0.2, 0.4, 0.6 and 0.8 on the Wikipedia and INRIA-Websearch datasets. The highest MAP score is shown in **bold**.

Method	Wikipedia								INRIA-Websearch							
	Image $\rightarrow$ Text				Text $\rightarrow$ Image				Image $\rightarrow$ Text				Text $\rightarrow$ 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
MCCA [43]	0.202	0.202	0.202	0.202	0.189	0.189	0.189	0.189	0.275	0.275	0.275	0.275	0.277	0.277	0.277	0.277
PLS [44]	0.337	0.337	0.337	0.337	0.320	0.320	0.320	0.320	0.387	0.387	0.387	0.387	0.398	0.398	0.398	0.398
DCCA [1]	0.281	0.281	0.281	0.281	0.260	0.260	0.260	0.260	0.188	0.188	0.188	0.188	0.182	0.182	0.182	0.182
DCCAE [51]	0.308	0.308	0.308	0.308	0.286	0.286	0.286	0.286	0.167	0.167	0.167	0.167	0.164	0.164	0.164	0.164
GMA [45]	0.200	0.178	0.153	0.139	0.189	0.160	0.141	0.136	0.425	0.372	0.303	0.245	0.437	0.378	0.315	0.251
MvDA [20]	0.379	0.285	0.217	0.144	0.350	0.270	0.207	0.142	0.286	0.269	0.234	0.186	0.285	0.265	0.233	0.185
MvDA-VC [21]	0.389	0.330	0.256	0.162	0.355	0.304	0.241	0.153	0.288	0.272	0.241	0.192	0.286	0.268	0.238	0.190
GSS-SL [61]	0.444	0.390	0.309	0.174	0.398	0.353	0.287	0.169	0.487	0.424	0.272	0.075	0.510	0.451	0.307	0.085
ACMR [50]	0.276	0.231	0.198	0.135	0.285	0.194	0.183	0.138	0.175	0.096	0.055	0.023	0.157	0.114	0.048	0.021
deep-SM [53]	0.441	0.387	0.293	0.178	0.392	0.364	0.248	0.177	0.495	0.422	0.238	0.046	0.509	0.421	0.258	0.063
FGCrossNet [11]	0.403	0.322	0.233	0.156	0.358	0.284	0.205	0.147	0.278	0.192	0.105	0.027	0.261	0.189	0.096	0.025
SDML [16]	0.464	0.406	0.299	0.170	0.448	0.398	0.311	0.184	0.506	0.419	0.283	0.024	0.512	0.412	0.241	0.066
DSCMR [62]	0.426	0.331	0.226	0.142	0.390	0.300	0.212	0.140	0.500	0.413	0.225	0.055	0.536	0.464	0.237	0.052
SMLN [17]	0.449	0.365	0.275	0.251	0.403	0.319	0.246	0.237	0.331	0.291	0.262	0.214	0.391	0.349	0.292	0.254
Ours	**0.514**	**0.491**	**0.464**	**0.435**	**0.461**	**0.453**	**0.421**	**0.400**	**0.559**	**0.543**	**0.512**	**0.417**	**0.587**	**0.571**	**0.533**	**0.424**

Table 2: Performance comparison in terms of MAP scores under the symmetric noise rates of 0.2, 0.4, 0.6 and 0.8 on the NUS-WIDE and XMediaNet datasets. The highest MAP score is shown in **bold**.

Method	NUS-WIDE								XMediaNet							
	Image $\rightarrow$ Text				Text $\rightarrow$ Image				Image $\rightarrow$ Text				Text $\rightarrow$ 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
MCCA [43]	0.523	0.523	0.523	0.523	0.539	0.539	0.539	0.539	0.233	0.233	0.233	0.233	0.249	0.249	0.249	0.249
PLS [44]	0.498	0.498	0.498	0.498	0.517	0.517	0.517	0.517	0.276	0.276	0.276	0.276	0.266	0.266	0.266	0.266
DCCA [1]	0.527	0.527	0.527	0.527	0.537	0.537	0.537	0.537	0.152	0.152	0.152	0.152	0.162	0.162	0.162	0.162
DCCAE [51]	0.529	0.529	0.529	0.529	0.538	0.538	0.538	0.538	0.149	0.149	0.149	0.149	0.159	0.159	0.159	0.159
GMA [45]	0.545	0.515	0.488	0.469	0.547	0.517	0.491	0.475	0.400	0.380	0.344	0.276	0.376	0.364	0.336	0.277
MvDA [20]	0.590	0.551	0.568	0.471	0.609	0.585	0.596	0.498	0.329	0.318	0.301	0.256	0.324	0.314	0.296	0.254
MvDA-VC [21]	0.531	0.491	0.512	0.421	0.567	0.525	0.550	0.434	0.331	0.319	0.306	0.274	0.322	0.310	0.296	0.265
GSS-SL [61]	0.639	0.639	0.631	0.567	0.659	0.658	0.650	0.592	0.431	0.381	0.256	0.044	0.417	0.361	0.221	0.031
ACMR [50]	0.530	0.433	0.318	0.269	0.547	0.476	0.304	0.241	0.181	0.069	0.018	0.010	0.191	0.043	0.012	0.009
deep-SM [53]	0.693	0.680	0.673	0.628	0.690	0.681	0.669	0.629	0.557	0.314	0.276	0.062	0.495	0.344	0.021	0.014
FGCrossNet [11]	0.661	0.641	0.638	0.594	0.669	0.669	0.636	0.596	0.372	0.280	0.147	0.053	0.375	0.281	0.160	0.052
SDML [16]	0.694	0.677	0.633	0.389	0.693	0.681	0.644	0.416	0.534	0.420	0.216	0.009	0.563	0.445	0.237	0.011
DSCMR [62]	0.665	0.661	0.653	0.509	0.667	0.665	0.655	0.505	0.461	0.224	0.040	0.008	0.477	0.224	0.028	0.010
SMLN [17]	0.676	0.651	0.646	0.525	0.685	0.650	0.639	0.520	0.520	0.445	0.070	0.070	0.514	0.300	0.303	0.226
Ours	**0.696**	**0.690**	**0.686**	**0.669**	**0.697**	**0.695**	**0.688**	**0.673**	**0.625**	**0.581**	**0.384**	**0.334**	**0.623**	**0.587**	**0.408**	**0.359**

In particular, the proposed method outperforms the best baselines by more than 6.5% with 80% noise. It demonstrates that the proposed framework is more robust to the noisy labels and could give a guide for future multimodal learning with noisy labels.

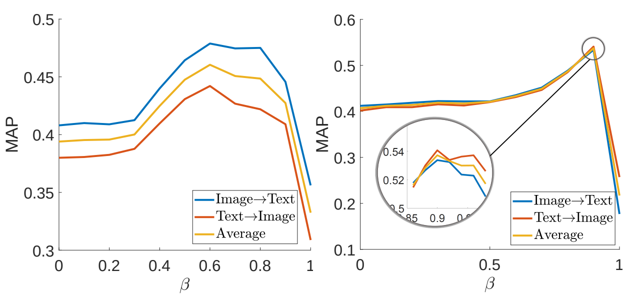
4.5. Ablation Study

In this section, we evaluate the performance of the proposed components (*i.e.*, $\mathcal{L}_r$ and $\mathcal{L}_c$) for cross-modal retrieval. To extensively investigate the contributions of each component, we compare our MRL with its three counterparts on the Wikipedia dataset, which are the cross-entropy (CE) baseline [53] and two variations of our MRL: MRL with $\mathcal{L}_c$ only and MRL with $\mathcal{L}_r$ only. All the compared methods are train with the same settings as our MRL for a fair comparison. The experimental results are shown in Table 3. From the results, one can see that the performance of MRL without $\mathcal{L}_c$ or $\mathcal{L}_r$ are worse than our full MRL on Wikipedia, which indicates that both of the two components contribute to cross-modal retrieval in our framework. By comparing to CE, we can see that our $\mathcal{L}_r$ can achieve much more robust performance, which indicates that the proposed $\mathcal{L}_r$ can alleviate the interference of noisy labels.

Table 3: Comparison between our MRL (full version) and its three counterparts (CE and two variations of MRL) under the symmetric noise rates of 0.2, 0.4, 0.6 and 0.8 on the Wikipedia dataset. The highest score is shown in **bold**.

Method	Image → Text			
	0.2	0.4	0.6	0.8
CE	0.441	0.387	0.293	0.178
MRL (with $\mathcal{L}_r$ only)	0.482	0.434	0.363	0.239
MRL (with $\mathcal{L}_c$ only)	0.412	0.412	0.412	0.412
Full MRL	**0.514**	**0.491**	**0.464**	**0.435**
	Text → Image			
CE	0.392	0.364	0.248	0.177
MRL (with $\mathcal{L}_r$ only)	0.429	0.389	0.320	0.202
MRL (with $\mathcal{L}_c$ only)	0.383	0.382	0.383	0.383
Full MRL	**0.461**	**0.453**	**0.421**	**0.400**

4.6. Parameter Analysis

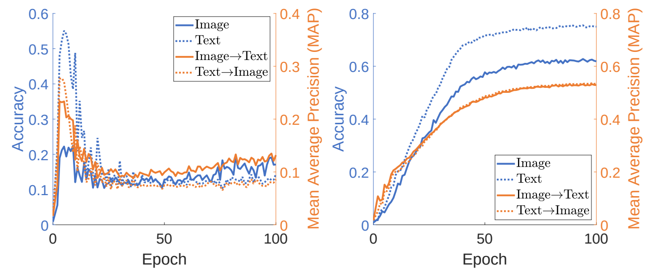


(a) Wikipedia (b) INRIA-Websearch

Figure 4: Cross-modal retrieval performance of our MRL in terms of MAP scores versus different values of β on the validation sets of the Wikipedia and INRIA-Websearch datasets, respectively. The noise rate is 0.6.

To evaluate the impact of the trade-off hyper-parameter β, we plot the cross-modal retrieval accuracy versus β on the validation sets of Wikipedia and INRIA-Websearch in Figure 4. From the figure, we can see that both Robust Clustering loss ($\mathcal{L}_r$) and Multimodal Contrastive loss ($\mathcal{L}_r$) contribute to excavating the discrimination from the multimodal data, which is consist with our ablation study. However, the contributions of each component are distinct for different datasets, which may be caused by the difficulty level of the datasets (*e.g.*, the more classes there are, the more difficult it will be). Furthermore, the sensitivity of β is also different on distinct datasets. To be specific, our method can obtain stable performance in a relatively larger range (*i.e.*, $0.5 \sim 0.8$) on Wikipedia, but smaller range (*i.e.*, $0.85 \sim 0.95$) on INRIA-Websearch.

4.7. Robustness Analysis



(a) Cross-entropy (b) Ours

Figure 5: Validation accuracies and MAP scores of the proposed method versus cross-entropy on INRIA-Websearch under 0.6 symmetric noise.

To visually investigate the robustness improvement, we plot accuracies/MAP scores versus epochs on the validation set of INRIA-Websearch for cross-entropy [53] and the proposed method in Figure 5. From the results, we can see that the proposed method can achieve much more stable performance on the validation set, which indicates that our method has alleviated the interference of noisy labels, and embracing more robust performance.

5. Conclusion

In this paper, we proposed a novel robust multimodal framework for learning from noisy labels, termed Multimodal Robust Learning (MRL), to project different modalities into a latent common space. Our MRL consists of multiple modality-specific networks, a multimodal robust clustering loss $\mathcal{L}_r$, and a multimodal contrastive loss $\mathcal{L}_c$. The robust clustering loss $\mathcal{L}_r$ aims to mitigate the interference of noisy labels and the cross-modal discrepancy. The multimodal contrastive loss $\mathcal{L}_c$ is conducted to narrow the heterogeneous gap between different modalities while excavating the apparent discrimination. Comprehensive experiments are conducted on four widely-used datasets. The experimental results demonstrated the effectiveness of the proposed method. Specifically, our MRL is superior to 14 state-of-the-art multimodal methods on noise settings. At the same time, we revealed that the existing cross-modal retrieval approaches are vulnerable to noisy labels.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2020YFB1406702; in part by the Fundamental Research Funds for the Central Universities under Grant YJ201949; in part by NFSC under Grants 61836006, U19A2078, U19A2081, 61625204; and A*STAR under its AME Programmatic Funds (Project No.A1892b0026 and Project No.A19E3b0099).

References

[1] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013. 6, 7

[2] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017. 2, 4

[3] Pengfei Chen, Benben Liao, Guangyong Chen, and Shengyu Zhang. Understanding and utilizing deep neural networks trained with noisy labels. 97:1062–1070, 2019. 2, 4

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, 2020. 2, 4, 5

[5] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009. 5, 6

[6] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Transactions on Image Processing*, 27(8):3893–3903, 2018. 1

[7] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 7–16. ACM, 2014. 6

[8] Aritra Ghosh, Himanshu Kumar, and PS Sastry. Robust loss functions under label noise for deep neural networks. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 3

[9] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama. Masking: A new perspective of noisy supervision. In *Advances in Neural Information Processing Systems*, pages 5836–5846, 2018. 1, 3

[10] 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. In *Advances in neural information processing systems*, pages 8527–8537, 2018. 1, 2, 3, 4

[11] Xiangteng He, Yuxin Peng, and Liu Xie. A new benchmark and approach for fine-grained cross-media retrieval. In *Proceedings of the 2019 ACM on Multimedia Conference*, 2019. 6, 7

[12] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 3

[13] Peng Hu, Dezhong Peng, Yongsheng Sang, and Yong Xiang. Multi-view linear discriminant analysis network. *IEEE Transactions on Image Processing*, 28(11):5352–5365, 2019. 3

[14] Peng Hu, Dezhong Peng, Xu Wang, and Yong Xiang. Multimodal adversarial network for cross-modal retrieval. *Knowledge-Based Systems*, 180:38–50, 2019. 3

[15] Peng Hu, Xi Peng, Hongyuan Zhu, Jie Lin, Liangli Zhen, and Dezhong Peng. Joint versus independent multiview hashing for cross-view retrieval. *IEEE Transactions on Cybernetics*, 2020. 3

[16] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu. Scalable deep multimodal learning for cross-modal retrieval. In *Proceedings of the 42Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 635–644, New York, NY, USA, 2019. ACM. 1, 3, 6, 7

[17] Peng Hu, Hongyuan Zhu, Xi Peng, and Jie Lin. Semi-supervised multi-modal learning with balanced spectral decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 99–106, 2020. 1, 3, 6, 7

[18] Zhenyu Huang, Peng Hu, Joey Tianyi Zhou, Jiancheng Lv, and Xi Peng. Partially view-aligned clustering. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[19] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*, pages 2304–2313, 2018. 1, 3

[20] Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. Multi-view discriminant analysis. In *European Conference on Computer Vision*, pages 808–821, 2012. 6, 7

[21] 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, 2016. 3, 6, 7

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. 5

[23] Josip Krapac, Moray Allan, Jakob Verbeek, and Frédéric Juried. Improving web image search results using query-relative classifiers. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1094–1101. IEEE, 2010. 2, 5, 6

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 5

[25] Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. In *Workshop on Representation Learning for NLP*, pages 78–86. Association for Computational Linguistics, 2016. 5

[26] Chao Li, Cheng Deng, Lei Wang, De Xie, and Xianglong Liu. Coupled cyclegan: Unsupervised hashing network for cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 176–183, 2019. 3

[27] Chao Li, Shangqian Gao, Cheng Deng, De Xie, and Wei Liu. Cross-modal learning with adversarial samples. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 3

[28] Junnan Li, Richard Socher, and Steven CH Hoi. DivideMix: Learning with noisy labels as semi-supervised

learning. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 3, 4

[29] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Learning to learn from noisy labeled data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5051–5059, 2019. 3

[30] 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

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[32] 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*, 2020. 3, 4

[33] Devraj Mandal and Soma Biswas. Cross-modal retrieval with noisy labels. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2326–2330. IEEE, 2020. 3

[34] Julien Mairal Priya Goyal Piotr Bojanowski Armand Joulin Mathilde Caron, Ishan Misra. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, 2020. 4, 5

[35] Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. 6

[37] Xi Peng, Jiashi Feng, Shijie Xiao, Wei-Yun Yau, Joey Tianyi Zhou, and Songfan Yang. Structured autoencoders for subspace clustering. *IEEE Transactions on Image Processing*, 27(10):5076–5086, 2018. 4

[38] Xi Peng, Zhenyu Huang, Jiancheng Lv, Hongyuan Zhu, and Joey Tianyi Zhou. COMIC: Multi-view clustering without parameter selection. In *International Conference on Machine Learning*, pages 5092–5101. PMLR, 2019. 3

[39] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Transactions on Multimedia*, 20(2):405–420, Feb 2018. 6

[40] Yuxin Peng, Jinwei Qi, and Yuxin Yuan. Modality-specific cross-modal similarity measurement with recurrent attention network. *IEEE Transactions on Image Processing*, 27(11):5585–5599, 2018. 5, 6

[41] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on Multimedia*, pages 251–260. ACM, 2010. 5, 6

[42] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, 2018. 3

[43] Jan Rupnik and John Shawe-Taylor. Multi-view canonical correlation analysis. In *Conference on Data Mining and Data Warehouses (SiKDD 2010)*, pages 1–4, 2010. 6, 7

[44] Abhishek Sharma and David W Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 593–600. IEEE, 2011. 6, 7

[45] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W Jacobs. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2160–2167. IEEE, 2012. 3, 6, 7

[46] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, pages 1919–1930, 2019. 3

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[48] Hwanjun Song, Minseok Kim, Dongmin Park, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020. 1, 3

[49] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 839–847, 2017. 1, 3

[50] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 154–162. ACM, 2017. 1, 6, 7

[51] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015. 6, 7

[52] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 322–330, 2019. 3

[53] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. Cross-modal retrieval with cnn visual features: A new baseline. *IEEE Transactions on Cybernetics*, 47(2):449–460, Feb 2017. 1, 2, 3, 6, 7, 8

[54] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733–3742, 2018. 2, 4, 5

[55] De Xie, Cheng Deng, Chao Li, Xianglong Liu, and Dacheng Tao. Multi-task consistency-preserving adversarial hashing

for cross-modal retrieval. *IEEE Transactions on Image Processing*, 29:3626–3637, 2020. 3

[56] Yan Yan, Zhongwen Xu, Ivor W Tsang, Guodong Long, and Yi Yang. Robust semi-supervised learning through label aggregation. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016. 3

[57] En Yu, Jiande Sun, Jing Li, Xiaojun Chang, Xian-Hua Han, and Alexander G Hauptmann. Adaptive semi-supervised feature selection for cross-modal retrieval. *IEEE Transactions on Multimedia*, 21(5):1276–1288, 2018. 3

[58] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, 2019. 1, 3

[59] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations (ICLR)*, 2017. 1

[60] Changqing Zhang, Zongbo Han, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. CPM-Nets: Cross partial multi-view networks. In *Advances in Neural Information Processing Systems*, pages 559–569, 2019. 1

[61] Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, and Qi Tian. Generalized semi-supervised and structured subspace learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 20(1):128–141, 2018. 1, 3, 6, 7

[62] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10394–10403, 2019. 3, 6, 7

[63] Tao Zhou, Huazhu Fu, Chen Gong, Jianbing Shen, Ling Shao, and Fatih Porikli. Multi-mutual consistency induced transfer subspace learning for human motion segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10277–10286, 2020. 3