Mutual Quantization for Cross-Modal Search with Noisy Labels

Erkun Yang[1], Dongren Yao[2], Tongliang Liu[3], Cheng Deng[1]*
[1]School of Electronic Engineering, Xidian University, Xi'an, China
[2]MEEI and Harvard Medical School, 243 Charles Street, Boston, MA, USA
[3]TML Lab, Sydney AI Centre, The University of Sydney, Australia
{erkunyang, chdeng}@gmail.com, dongren_yao@meei.harvard.edu,tongliang.liu@sydney.edu.au

Abstract

Deep cross-modal hashing has become an essential tool for supervised multimodal search. These models tend to be optimized with large, curated multimodal datasets, where most labels have been manually verified. Unfortunately, in many scenarios, such accurate labeling may not be available. In contrast, datasets with low-quality annotations may be acquired, which inevitably introduce numerous mistakes or label noise and therefore degrade the search performance. To address the challenge, we present a general robust cross-modal hashing framework to correlate distinct modalities and combat noisy labels simultaneously. More specifically, we propose a proxy-based contrastive (PC) loss to mitigate the gap between different modalities and train networks for different modalities jointly with small-loss samples that are selected with the PC loss and a mutual quantization loss. The small-loss sample selection from such joint loss can help choose confident examples to guide the model training, and the mutual quantization loss can maximize the agreement between different modalities and is beneficial to improve the effectiveness of sample selection. Experiments on three widely-used multimodal datasets show that our method significantly outperforms existing state-of-the-arts.

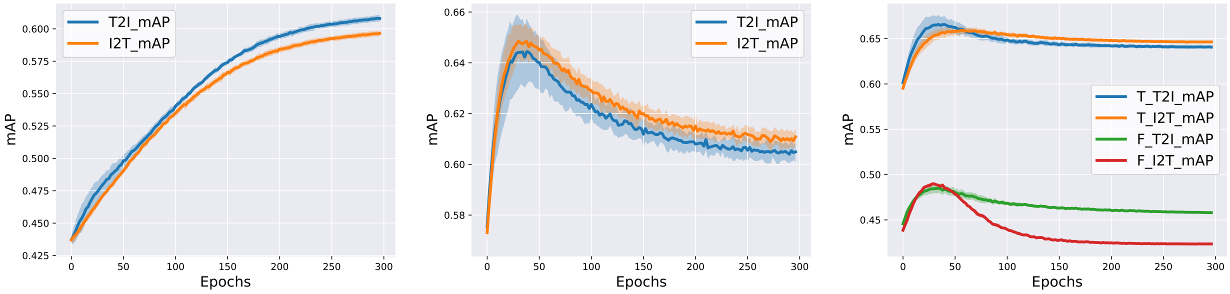
1. Introduction

By transforming high-dimensional data from multiple modalities into compact binary hash codes in a common Hamming space, cross-modal hashing offers remarkable efficiency for large-scale multi-modal data storage and search. Recently, supervised deep learning-based cross-modal hashing methods have achieved promising results and been applied to many multi-modal learning tasks [4–6, 43, 49]. These models usually rely on a large number of training instances with clean and intact labels. However, noisy labels, which are systematically corrupted, are ubiquitous and unavoidable in our daily life, such as social-network tagging [7], crowdsourcing [38], medical diagnosis [13], and financial analysis [1]. As deep networks have large model capacities, they can easily memorize and eventually overfit these noisy labels, which correspondingly degenerates the model generalization [51].

To combat the impact of noisy labels, numerous studies have been conducted, such as correction method [14], MentorNet [19], Co-teaching [15], and T-revision [40]. These methods can learn robust continuous representations for unimodal learning tasks. However, they cannot simultaneously tackle multi-modal inputs, such as real-world multimedia data. Moreover, continuous representations are inefficient for storage and computation, and it is non-trivial to binarize continuous representations. Improper binarization may introduce large quantization errors and severely degrade the model performance. Therefore, it is important to explore how to learn robust binary codes for cross-modal search with noisy labels. While this is rarely touched in previous works.

We perform an empirical study on a general deep cross-modal hashing framework trained with noisy labels. Figure 1a illustrates the mean average precision (mAP) values in different epochs for the training datasets evaluated with noisy labels. We can see that the performance continues to increase during the training stage, which suggests that the model will keep memorizing noisy labels during the whole training procedure. Figure 1b shows the performance of testing data evaluated with clean labels, which also demonstrates that the models will fast overfit to noisy labels, thus degrading the search performance. Furthermore, from Figure 1a, we can see that there exists diversity in the performance of different modalities since representations from different modalities may exist in entirely different spaces with heterogeneity, making learning from noisy data more difficult. Lastly, wrongly labeled data can confuse the discriminative connections across distinct modalities, resulting in challenges mitigating the heterogeneous gap. Therefore,

[*] Corresponding author.



(a) Train results with noisy labels

(b) Test results

(c) Train results with clean labels

Figure 1. We train a basic cross-modal hashing model with binary cross-entropy loss on MIRFlickr-25k dataset with 0.6 asymmetric noise. The mean average precision (mAP) values with different epochs are reported in the figures: (a) mAP values based on noisy labels for the training dataset; (b) mAP values for the testing dataset; (c) mAP values based on clean labels for the training dataset, where "T_T2I_mAP" and "T_I2T_mAP" are for the clean training dataset, and "F_T2I_mAP" and "F_I2T_mAP" are for the corrupted training dataset. For each line, we run five times and report the mean values. The error bar for STD in each sub-figure is highlighted as a shade.

it is more challenging and complex to simultaneously consider both noisy data and inter-modal discrepancy.

To address the above problems, we first provide a more in-depth analysis of the impact of noisy labels on deep cross-modal search models. Then we propose our method to combat the effect of noisy labels. Specifically, according to the ground-truth labels, we split the noisy training dataset into clean dataset that are accurately labeled and corrupted dataset that are wrongly labeled. Then, we show the mAP values for these two datasets based on correct labels with different training epochs, respectively. The results are shown in Figure 1c. As been revealed by previous studies on image classifications, there exists a memorization effect for deep neural networks (DNNs): DNNs tend first to memorize and fit majority (clean) patterns and then overfit minority (noisy) patterns. From Figure 1b and Figure 1c, we can also obtain the following critical findings for cross-modal search tasks: (1) the performance evaluated with correct labels for both clean cross-modal training data and corrupted cross-modal training data will first increase and then decrease, which show that, during the early learning stage, the cross-modal search model can fit clean data and also generalize well to corrupted data; (2) the performance of test data first increase and then decrease, suggesting that the learning from clean data can dominate the early learning stage and then be overwhelmed by the overfitting to noisy labels.

Based on the above analysis, we propose a robust cross-modal hashing framework called Cross-Modal Mutual Quantization (CMMQ) to combat the impact of noisy labels and narrow the heterogeneous gap simultaneously. Firstly, to correlate different modalities, we design to generate proxy codes for each class based on the Hadamard matrix [34] and adopt a proxy-based contrastive (PC) loss to push examples from different modalities to the corresponding shared proxy codes. Secondly, to excavate the discriminative information from noisy labels, we exploit the memorization effect of deep cross-modal models and preferentially select examples with small PC losses to train the network confidently. Thirdly, different models are unlikely to agree on noisy examples. Hence we adopt a mutual quantization loss to maximize the agreement of networks from different modalities, which can further improve the effectiveness of sample selection. The overall learning framework is illustrated in Figure 2.

Before delving into details, we clearly emphasize our contributions as follows.

- We propose a proxy-based contrastive (PC) loss, which can push examples from different modalities to their corresponding shared proxy codes and narrow the heterogeneous gap effectively.
- By preferentially selecting examples with small losses, our method can effectively exploit the memorization effect of deep cross-modal networks and combat the impact of noisy labels.
- A mutual quantization loss is proposed to maximize the agreement of models from different modalities, thus can improve the quality of the predicted codes.
- Experiments on three cross-modal benchmark datasets clearly demonstrate that the method can outperform many state-of-the-art approaches in various settings.

The rest of the paper is organized as follows. In Section 2, we briefly review some closely related works. In Section 3, we propose our cross-modal mutual quantization paradigm. Section 4 shows the experimental results. Finally, concluding remarks are provided in Section 5.

2. Related Work

2.1. Deep Cross-Modal Hamming Search

Different from the traditional "learn to hash" for Cross-Modal Hamming Search (CMHS) [25,35]. Deep CMHS [4, 9, 10, 20, 42, 45, 46] utilizes deep networks to transform multi-modal inputs into binary codes in Hamming space. Due to the high model capacity and the joint processing of feature extraction and binary quantization, deep CMHS usually achieves better search performance. To utilize the semantic information in supervisory labels, some supervised deep CMHS methods are proposed to learn a common discriminative Hamming space for multi-modal data. These methods are usually learned with *classification-based losses* (e.g., cross-entropy loss) [41] or *distance-based losses* (e.g., pair-wise or triplet losses) [9, 20]. To alleviate the demand of large, correctly labeled data, some semi-supervised learning methods [48,52,54,55] are proposed to exploit both of labeled and unlabeled data. There is also one method NrDCMH [36] tries to address the problem by learning with noisy label data. Specifically, NrDCMH first detects the noisy training examples based on the margin between the feature similarity and the label similarity and then reweights data pairs based on the similarity margins. However, this work can only handle the scenario, where only the semantic labels that examples do not belong to can be flipped. In real-world applications, this assumption may be too strong. Moreover, it is much harder to address the problems that all the data labels can be corrupted and flipped. While this is rarely touched in previous works.

2.2. Noisy Label Learning

Learning with noisy data has been well studied [2, 17, 23, 27, 37, 47]. Existing methods for learning with noisy labels can be mainly categorized into model-based and model-free methods. The first type models the relationship between clean labels and noisy labels by estimating the noise transition matrix that denotes the probabilities of clean labels flipping into noisy labels [30, 39]. With a perfectly estimated noisy transition matrix [30], these methods can guarantee the classifier learned from the noisy data to be consistent with the optimal classifier (i.e., the minimizer of the clean risk) [27]. However, current methods are usually fragile to estimate the noise transition matrix for heavy noisy data and are also hard to handle a large number of classes [15]. The second strand usually employs heuristics to reduce the side-effect of noisy labels. For example, many state-of-the-art approaches in this category are specifically designed to, e.g., select reliable examples [15, 50], reweight examples [19, 31], correct labels [21, 28], employ side information [33], and (implicitly) add regularization [14]. Although the differences between the learned classifiers from those methods and the optimal classifiers for clean data are not guaranteed to vanish, those methods are reported to work empirically well. However, all existing noisy labels methods are specifically designed for unimodal learning with continuous features, and it is a challenge to extend them to multi-modal learning with binary representations.

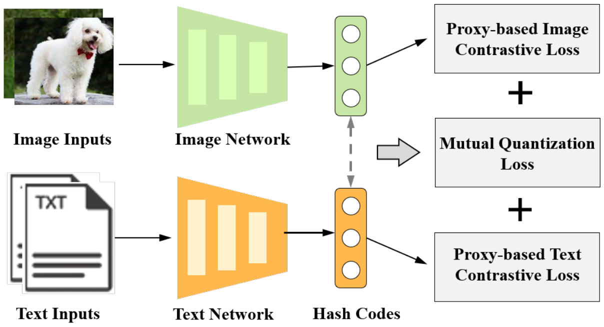


Figure 2. Framework of the proposed CMMQ.

3. The Proposed Approach

3.1. Preliminary

For cross-modal search with m modalities, we denote a multi-modal dataset with N examples as $\mathcal{D} = \{\mathcal{M}_i\}_{i=1}^{m}$, where $\mathcal{M}_i = \{\boldsymbol{x}_j^i, \boldsymbol{y}_j^i\}_{j=1}^{N}$ is the i-th modality, $\boldsymbol{x}_j^i \in \mathcal{X}_i$ is the j-th example from the i-th modality, and $\boldsymbol{y}_j^i$ is the corresponding label. In real-world applications, the clean label $\boldsymbol{y}_j^i$ may be randomly corrupted to noisy label $\tilde{\boldsymbol{y}}_j^i$ before being observed. Therefore we assume that, during training, we can only access to a noisy multi-modal training dataset $\tilde{\mathcal{D}}_{tr} = \{\tilde{\mathcal{M}}_i\}_{i=1}^{m}$ with $\tilde{\mathcal{M}}_i = \{\boldsymbol{x}_j^i, \tilde{\boldsymbol{y}}_j^i\}_{j=1}^{N}$. While, to accurately evaluate the performance of the proposed method, we assume that there exists a clean test dataset $\mathcal{D}_{te}$.

Definition 1 (Robust Cross-Modal Hamming Search)**.** *To achieve effective cross-modal search, multi-modal inputs are usually transformed into a common Hamming space $\mathcal{B}$ via different hash functions $\{h_i : \mathcal{X}_i \rightarrow \mathcal{B}\}$. Here, h_i is the hash function for the i-th modality, which can be instantiated with a DNN model parameterized with Θ_i. Then, given an input data $\boldsymbol{x}_j^i$, we can obtain the corresponding binary codes with*

$$\boldsymbol{b}_j^i = h_i(\boldsymbol{x}_j^i, \Theta_i). \tag{1}$$

For robust cross-modal Hamming search, we want to learn a family of hash functions $\{h_i, i = 1, ..., m\}$ with the noisy multi-modal training dataset $\tilde{\mathcal{D}}_{tr}$, so that the hash codes obtained from Eq. (1) can perform well on the clean test dataset $\mathcal{D}_{te}$.

In the following, we first propose a proxy-based contrastive learning method to maximize the correlation between different modalities. Then, based on this framework, we elaborate on the strategy to select confident examples and also the cross-modal mutual regularization to counteract the impact of noisy labels.

3.2. Proxy-based Contrastive Quantization

To enable effective cross-modal search, inspired by the proxy learning techniques [29] and center representation learning approaches [11], we first generate a set of shared proxy codes $\boldsymbol{O} = \{\boldsymbol{o}_1, ..., \boldsymbol{o}_N\}$, where $\boldsymbol{o}_i \in \{0,1\}^K$ is the proxy code with length K for the i-th example. Then, we maximize the similarities between the generated hash codes from different modalities and the corresponding shared proxy codes to minimize the cross-modal semantic gap. Specifically, for an example $\boldsymbol{x}_j^i$, its probability belonging to the j-th proxy codes can be estimated by

$$P(\boldsymbol{o}_j|\boldsymbol{x}_j^i) = \delta(S(\boldsymbol{b}_j^i, \boldsymbol{o}_j)), \tag{2}$$

where $S(\boldsymbol{b}_j^i, \boldsymbol{o}_j)$ measures the similarity between the example $\boldsymbol{b}_j^i$ and the proxy code $\boldsymbol{o}_j$. Since both $\boldsymbol{b}_j^i$ and $\boldsymbol{o}_j$ are binary vectors, we set $S(\cdot,\cdot)$ as the negative Binary Cross Entropy (BCE): $S(\boldsymbol{b}_j^i, \boldsymbol{o}_j) = -\frac{1}{K}\sum_{k\in K}(\boldsymbol{o}_{j,k}\log \boldsymbol{b}_{j,k}^i + (1-\boldsymbol{o}_{j,k})\log \boldsymbol{b}_{j,k}^i)$. While $\delta(\cdot)$ is a function that can transform $S(\boldsymbol{b}_j^i, \boldsymbol{o}_j)$ into probabilities, here we simply set $\delta(x) = \exp(-\beta x)$. Then, we can formulate the Proxy-based Contrastive (PC) Loss as

$$\mathcal{L}_p = \sum_{j=1}^{N}\sum_{i=1}^{m}[\log(\delta(S(\boldsymbol{b}_j^i, \boldsymbol{o}_j))) + \alpha R(\boldsymbol{b}_j^i)], \tag{3}$$

where $R(\boldsymbol{b}_j^i)$ is a regularization term to reduce the quantization error of the learned binary codes, and α is a hyper-parameter to control the two parts. Following [20], we can set $R(\boldsymbol{b}_j^i) = \sum_{k=1}^{K}(|b_{j,k}^i| - 1)$.

From Eq. (3), we can see that by minimizing the PC loss, the similarity between examples and their corresponding proxy codes will be maximized. As introduced in the following, the proxy codes are constructed based on data labels. Therefore, hash codes of examples in the same classes from different modalities will share common proxy codes, and minimizing Eq. (3) can explicitly improve intra-modal discriminability and enhance cross-modal correlation simultaneously.

Proxy Codes Generation. The proxy code serves as a shared optimization target for binary codes from different modalities. Here, we exploit the Hadamard matrix to construct it. Specifically, we first build a $K \times K$ Hadamard matrix with Sylvester's construction [34] as

$$\boldsymbol{H}_K = \begin{bmatrix} \boldsymbol{H}_{2^{n-1}}, & \boldsymbol{H}_{2^{n-1}} \\ \boldsymbol{H}_{2^{n-1}}, & -\boldsymbol{H}_{2^{n-1}} \end{bmatrix} = \boldsymbol{H}_2 \otimes \boldsymbol{H}_{2^{n-1}}, \tag{4}$$

where $\otimes$ indicates the Hadamard product, and $K = 2^n$. The two initial Hadamard matrix are $\boldsymbol{H}_1 = [1]$ and $\boldsymbol{H}_2 = \begin{bmatrix} 1, & 1 \\ 1, & -1 \end{bmatrix}$. The Hadamard matrix has the following two nice properties: (1) it is a binary matrix with elements being either -1 or $+1$; (2) the rows of a Hadamard matrix are mutually orthogonal, which means that the Hamming distance between any two row vectors equals to $K/2$. Note that we want to assign different proxy codes for different categories, therefore, if $c \leq K$, we directly choose each row to be a proxy code. While if $K < C \leq 2K$, we use a combination of two Hadamard matrices $\boldsymbol{H}_{2K} = [\boldsymbol{H}_k, -\boldsymbol{H}_K]^\top$ to construct the proxy codes. For single-label data, we assign one proxy code for each category. While for multi-label data, we first assign one proxy code for each class and then decide the proxy code for multi-label data by majority voting.

3.3. Confident Example Selection

Optimizing with the PC loss in Eq. (3) can mitigate the semantic gap between different modalities. However, as Figure 1b makes clear, when trained with noisy data, the model will finally overfit noisy labels resulting in sub-optimal performance. Also, from Figure 1c, we can observe that the learning from clean data will dominate the optimization at the early learning phase. This is consistent with [26], which reveals that, for the cross-entropy loss, its gradient is well correlated with the correct direction at the early learning stage. In other words, the loss of clean data will be minimized during the early learning phase.

With the above understanding, we intuitively consider examples with small losses as confident examples (i.e., examples with high probabilities to be clean) and apply the "small-loss" criterion to select confident examples. To be specific, we conduct the small-loss selection as

$$\tilde{D}_n = \mathop{\arg\min}_{D_n^{'}:|D_n^{'}|\geq R(t)|D_n|} \mathcal{L}_a(D_n^{'}), \tag{5}$$

where D_n indicates the mini-batch data, $R(t)$ controls the percentage of selected small-loss examples out of the mini-batch, and $\mathcal{L}_a$ is the adopted overall loss function. As demonstrated in Figure 1a, deep networks can learn clean and easy patterns in the initial epochs even with the existence of noisy labels. So they can filter out noisy examples using the loss values at the beginning of training. Yet, the problem is that when the number of epochs becomes large, the model will eventually overfit noisy labels. To rectify this problem, we want to keep more examples in the mini-batch at the start, i.e., $R(t)$ is large. Then, we gradually increase the drop rate, i.e., $R(t)$ becomes smaller, so that we can keep clean examples and drop the noisy ones before our model memorize them. The detailed setting of $R(t)$ can be found in Section 4.

Cross-Modal Mutual Quantization. To select small-loss examples with Eq. (5), we consider the loss in different modalities independently and use the sum of the losses.

While, from the view of agreement maximization principles [3, 32], different models would agree on labels of most clean examples and are unlikely to agree on corrupted labels. Based on this observation, we propose a mutual quantization loss to maximize the agreement between different modalities. Specifically, we adopt the Jensen-Shannon (JS) Divergence. To simplify the implementation, we use the symmetric Kullback-Leibler (KL) Divergence to surrogate this term

$$L_m = \sum_{i,j=1,i\neq j}^{m} D_{KL}(\boldsymbol{B}_i|\boldsymbol{B}_j) + D_{KL}(\boldsymbol{B}_j|\boldsymbol{B}_i), \qquad (6)$$

where the KL Divergence $D_{KL}(\boldsymbol{B}_i|\boldsymbol{B}_j)$ measures the multi-modal agreement for hash codes $\boldsymbol{B}_i$ and $\boldsymbol{B}_j$

$$D_{KL}(\boldsymbol{B}_i|\boldsymbol{B}_j) = \sum_{n=1}^{N}\sum_{k=1}^{K} \boldsymbol{b}^i_{n,k} \log(\frac{\boldsymbol{b}^i_{n,k}}{\boldsymbol{b}^j_{n,k}}). \qquad (7)$$

By matching the predictions from different modalities, the mutual quantization loss in Eq. (6) could help our algorithm to select examples with clean labels since an example with small mutual quantization loss means that networks from different modalities reach an agreement on its predictions. Furthermore, the regularization from networks in other modalities can also help the model to find a much wider minimum, which is expected to improve the generalization performance [53].

3.4. Optimization

Combining the proxy-based contrastive loss in Eq. (3) and the mutual quantization loss in Eq. (6) together, we can set the overall loss function as

$$\mathcal{L}_a = \mathcal{L}_p + \lambda \mathcal{L}_m. \qquad (8)$$

After selecting small-loss examples as in Eq. (5), we can calculate the loss on these examples for further optimization:

$$\mathcal{L} = \frac{1}{|\tilde{D}_n|} \sum_{\boldsymbol{x}^i_j \in \tilde{D}_n} \mathcal{L}_a(\boldsymbol{x}^i_j). \qquad (9)$$

The learning procedure is summarized in Algorithm 1.

4. Experiments

4.1. Datasets

We adopt three cross-modal benchmark datasets to evaluate the proposed method. Table 1 summarizes the main statistics of these three datasets, and the detailed information is provided below.

MIRFlickr-25K [18] contains 25,000 images collected from the Flickr website. Each image is assigned with a related text description that is represented as a 1386-demensional BoW vector. Following the previous method [20], totally 20,015 image-text pairs with 24 most frequent labels are used in our experiment.

```
Algorithm 1: Cross-Modal Mutual Quantization
Input: Noisy training image set $\tilde{D}_{tr}$, Networks
  $h_i, i = 1, ..., m$, code length $K$, learning rate $\eta$,
  epoch $T_{max}$, and iteration $I_{max}$.
Learning the proxy codes with Eq. (4);
Assign the proxy code $\boldsymbol{o}_j$ for each instance with
  majority voting;
for epoch $t = 1$ to $T_{max}$ do
    Fetch mini-batch $D_n$ from $\tilde{D}_{tr}$;
    $\boldsymbol{b}^i_j = h_i(\boldsymbol{x}^i_j, \Theta_i), \forall \boldsymbol{x}^i_j \in D_n$;
    Calculate the loss $\mathcal{L}_a$;
    Obtain small-loss sets $\tilde{D}_n$ by Eq. (5) from $\tilde{D}_n$ ;
    Obtain loss $\mathcal{L}$ by Eq. (9) on $\tilde{D}_n$;
    Update the network with backpropagation;
end
Output: Network parameters $\Theta_i, i = 1, ..., m$.
```

Table 1. Statistics of three datasets used in our experiments.

Dataset	Train	Test	Database
MIRFlickr-25k	10,000	2,000	18,015
NUS-WIDE	10,500	2,100	178,321
MS-COCO	10,000	5,000	117,218

NUS-WIDE [8] is a public web image dataset containing $269,648$ web images with 81 ground-truth annotated concepts. The associated text is represented as a $1,000$-dimensional Bow vector. After pruning the data that has no label or text information, we got 190,421 image-text pairs with 10 most frequent labels as our benchmark.

MS-COCO [24] has about 120,000 images, and each image is associated with a text, which is represented as a $2,000$-dimension BoW vector. Each image-text data pair is annotated with at least one of the 80 categories.

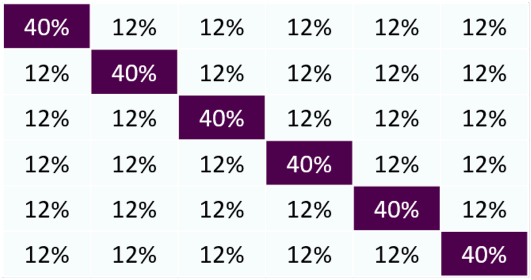
4.2. Experimental Setup

Baselines. We adopt four popular supervised deep cross-modal hashing methods, including DCMH [20], PRDH [44], SSAH [22], and CMHH [4] as the baselines. DCMH and PRDH are both based on pairwise labels. SSAH employs a network to learn representations from labels. CMHH considers to up-weight difficult pairs to improve the performance. The code for DCMH, PRDH, SSAH are kindly provided by the authors and we implement CMHH by ourselves.

Implementation. For a fair comparison, we use the ResNet18 [16] as the image modality backbone and a three-layer multi-layer perceptron (MLP) as the text modality

Table 2. mAP results of five methods on three cross-modal datasets, with best results shown in boldface.

Task	Method	MIRFlickr-25K			NUS-WIDE			MS-COCO		
		32 bits	64 bits	128 bits	32 bits	64 bits	128 bits	32 bits	64 bits	128 bits
I2T	DCMH	0.660	0.668	0.635	0.582	0.583	0.576	0.533	0.542	0.505
	PRDH	0.660	0.672	0.663	0.586	0.578	0.564	0.530	0.534	0.554
	SSAH	0.658	0.665	0.631	0.575	0.577	0.581	0.536	0.532	0.545
	CMHH	0.652	0.637	0.648	0.576	0.573	0.574	0.539	0.543	0.541
	CMMQ(Ours)	**0.737**	**0.742**	**0.757**	**0.601**	**0.606**	**0.607**	**0.542**	**0.556**	**0.565**
T2I	DCMH	0.707	0.709	0.693	0.582	0.585	0.577	0.524	0.523	0.534
	PRDH	0.671	0.696	0.692	0.577	0.579	0.592	0.529	0.534	0.522
	SSAH	0.662	0.653	0.678	0.566	0.570	0.576	0.502	0.529	0.525
	CMHH	0.704	0.683	0.658	0.571	0.577	0.585	0.527	0.536	0.514
	CMMQ(Ours)	**0.723**	**0.725**	**0.722**	**0.600**	**0.604**	**0.603**	**0.531**	**0.549**	**0.541**



40%	12%	12%	12%	12%	12%
12%	40%	12%	12%	12%	12%
12%	12%	40%	12%	12%	12%
12%	12%	12%	40%	12%	12%
12%	12%	12%	12%	40%	12%
12%	12%	12%	12%	12%	40%

Figure 3. The transition matrix of asymmetric noise with 0.6 noise rate (using 6 classes as an example).

backbone for our method and the four baselines. For the image modality, we replace the last fully-connected layer with a new fully-connected layer with K hidden units followed by a $\tanh(\cdot)$ activation function. We initialize this new layer randomly and initialize all the preceding layers with the model pre-trained on ImageNet [12]. The MLP for text modality is trained from scratch with random initialization. We employ the RMSprop optimizer with the factor of weight decay as 10^{-5}. For all the three datasets, we set the min-batch size as 128, the learning rate as 10^{-5}, and set $\alpha = 0.0001$, $\beta = 1$, $\lambda = 0.7$. To comprehensively evaluate the robustness of the methods, we set the label noise to be symmetric [15] and the noise rate to 0.6 in the experiments. The transition matrix between the noisy labels and ground-truth labels is illustrated in Figure 3.

Evaluation settings. Without loss of generality, all experiments are conducted on bi-modal datasets to evaluate two cross-modal tasks: using an image query to retrieve the related text samples (I2T), and using a text query to retrieve the relevant image points (T2I). Four evaluation metrics are used to evaluate the search performance, including mean of average precision (mAP), TopN-precision, recall@k, and precision-recall curves. The first three metrics are based on Hamming ranking, which ranks data points based on their Hamming distances to the query. While the precision-recall metric is based on the hash lookup protocol. Specifically, **mAP** is one of the most widely-used criteria for evaluating retrieval accuracy since it can simultaneously evaluate retrieval precision and ranking of returned results. Given a query and a list of R ranked retrieval results, the average precision (AP) for this query can be computed. mAP is defined as the average of APs for all queries. For all the three datasets, we set R as $5,000$. **TopN-precision** is defined as the average ratio of similar instances among the top N retrieved instances for all queries in terms of Hamming distance. In the experiments, N is set to 100. **Recall@k** counts the percentage of true neighbors from the top k retrieved instances among all the ground-truth. In the experiments, k is set to 100. Subjects are treated as similar if they share at least one common semantic label, otherwise, they are considered dissimilar.

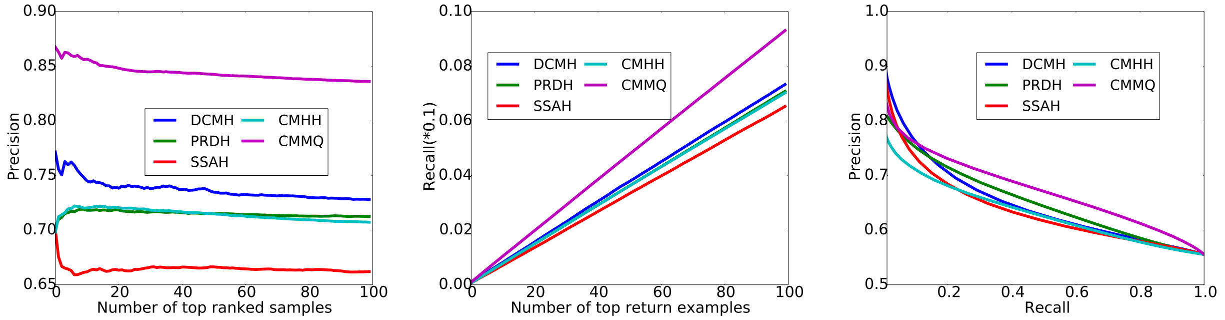
Selection setting. Following [15], we assume that the noise rate ρ is known and set the ratio of small-loss samples $R(t)$ as: $R(t) = 1 - \min\{\frac{t}{T_k}\rho, \rho\}$, where T_k is set as 10 for all the datasets. In practice, if the noisy rate ρ is not known in advance, it can be inferred using validation sets [27].

4.3. Results and Analysis

4.3.1 Results of Hamming Ranking

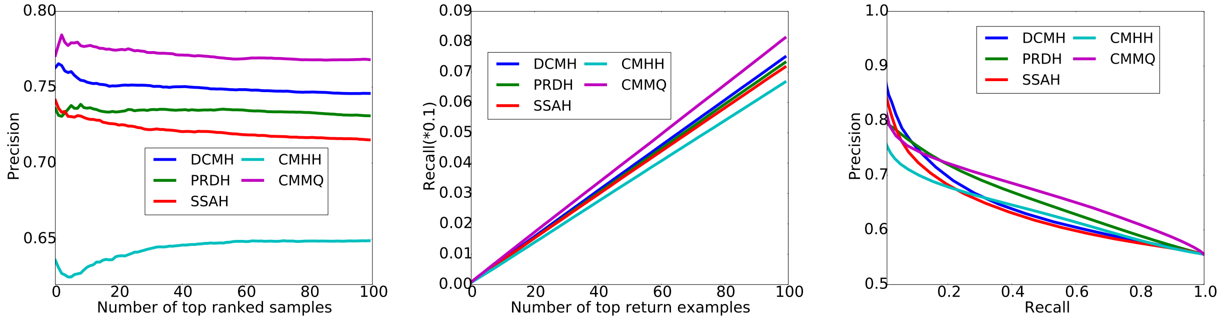
We first present mAP values for all methods on the three datasets using different lengths of hash code (*i.e.*, K) to provide a global evaluation. Then we select MIRFlickr-25K dataset and report the topN-precision and recall@k curves with $K = 64$ for a comprehensive contrastive study.

The mAP results for all methods on the three datasets are reported in Table 2. From the results, we can see that CMMH, which is more advancing in the clean label setting, cannot outperform DCMH for noisy data. This may be attributed to that CMMH pays more attention to hard examples that are more likely to be corrupted data when training with noisy labels. The results also verify that simultaneously dealing with cross-modal retrieval



(a) TopN-precision (b) Recall@k (c) Precision-recall

Figure 4. Precision-recall, topN-precision, and recall@k curves on MIRFlickr-25k with the hash code length of $K = 64$ for "I2T" task.



(a) TopN-precision (b) Recall@k (c) Precision-recall

Figure 5. Precision-recall, topN-precision, and recall@k curves on MIRFlickr-25k with the hash code length of $K = 64$ for "T2I" task.

and noisy labels is a nontrivial task. Moreover, from Table 2, we can also obtain that CMMQ usually outperforms other competing methods by large margins. For instance, compared to CMHH (the state-of-the-art deep cross-modal hashing method), CMMQ can obtain an absolute increase of $4.78\%/2.7\%$ in average mAP on the three datasets for the two retrieval tasks, respectively. The results clearly validate the superiority of the CMMQ method over previous approaches for cross-modal search with noisy labels.

The topN-precision and the recall@k curves for the "I2T" task with $K = 64$ achieved by different methods on MIRFlickr-25K dataset are shown in Figures 4a and 4b, and the corresponding curves for the "T2I" task are presented in Figures 5a and 5b. From those figures, we can obtain that the proposed CMMQ generally achieves higher precisions and recalls, which is consistent with the mAP evaluation. For cross-modal nearest neighbor search, users usually focus on the top returned results. Thus, the relevance of top returned instances with the query is more important. From the topN-precision and recall@k curves, one can see that CMMQ outperforms the other methods by a large margin when the number of returned instances is small (*e.g.*, < 40), which further demonstrates that CMMQ can better combat noisy labels and learn hash codes with high quality.

4.3.2 Results of Hash Lookup

Given a query object, the precision and recall values for the returned objects within any Hamming radius can be computed. By investigating these values with every Hamming radius from 0 to K, we can draw the precision-recall curves. Figures 4c and 5c report the precision-recall results of different methods on MIRFlickr-25k with $K = 64$ for "I2T" and "T2I" tasks, respectively. From these figures, one can again observe that CMMQ consistently achieves the best performance. From Table 2, Figure 4, and Figure 5, one can see that CMMQ generally obtains superior performance in terms of both Hamming ranking metrics (*i.e.*, mAP, topN-precision, and recall@k) and hash lookup metric (*i.e.*, precision-recall), demonstrating the utility of our method in learning binary codes for cross-modal search with noisy labels.

4.3.3 The Effect of Sample Selection

To analyze the effect of the small-loss sample selection for cross-modal search, we split the training data into clean data and corrupted data as in Section 1. Then we illustrate the histogram of the loss values of MIRFlickr-25k with 64 hash code length in Figure 6. From the results we can observe that most of the clean data have relatively small losses,

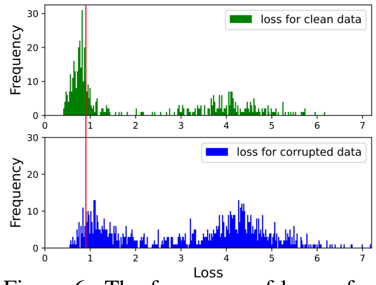


Figure 6. The frequency of losses for clean data and corrupted data.

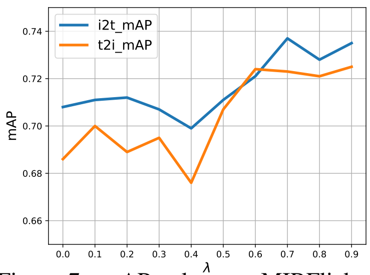


Figure 7. mAP values on MIRFlickr-25k for the two search tasks with different λ.

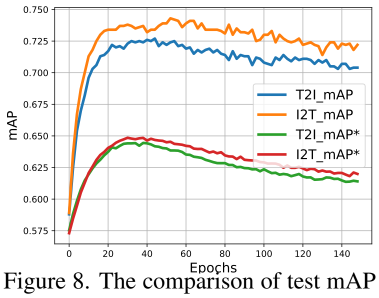


Figure 8. The comparison of test mAP between the BCE loss and our method.

while the losses for corrupted data are generally large. This figure clearly demonstrates that, by selecting examples with small losses, we can effectively construct a much cleaner subset, thus alleviating the impact of noisy labels.

4.4. Parameter Sensitivity Analysis

In this subsection, we study the parameter sensitivity for λ. We train our model on MIRFlickr-25k dataset with 32 hash code length and vary the value of λ from 0 to 1. The results are presented in Figure 7. We can see that the mAP values for both of the "I2T" and "T2I" tasks first increase and then stay at high values when the λ is larger than 0.6. The results indicate that, in practice, we can select λ from 0.6 to 0.9. In this paper, we set λ as 0.7.

By setting λ as 0, our model degenerates to the model without the mutual quantization loss. Therefore, by comparing the results with $\lambda = 0$ and other points in Figure 7, we can also verify the role of the proposed mutual quantization loss in our model. Specifically, we can observe that the search performance for both of the two tasks can be clearly improved by adding the mutual quantization loss, which demonstrates the effectiveness of the proposed mutual quantization loss. Besides, the results also verify that, by maximizing the agreement between outputs from different modalities, the model is likely to predict correct labels, thus improving the final search performance.

4.5. Robust Analysis

To visually investigate the robustness improvement, we plot mAP scores versus epochs on the test data for MIRFlickr-25k with the binary cross-entropy (BCE) and the proposed method in Figure 8, where "T2I_mAP*" and "I2T_mAP*" are for the BCE loss, and "T2I_mAP" and "I2T_mAP" are for our method. From the results, we can see that, although the BCE loss can improve the test performance at the early learning stage, noisy labels already have a severe impact on the cross-modal search model. By selecting confident examples with small losses, our method can combat noisy labels and greatly improve the model performance even at the early learning stage. Moreover, our method can continue to improve the search results when the performance of the model with BCE loss starts to decrease. Overall, compared with the original BCE loss, our proposed method can achieve much superior results, which indicates that our method can alleviate the interference of noisy labels and embrace more robust performance.

5. Conclusions

This work presents a cross-modal mutual quantization (CMMQ) method to simultaneously narrow the modality gap and combat noisy labels. To mitigate the discrepancies between different modalities, we first devise a proxy-based contrastive (PC) loss to pull the generated hash representations from different modalities close to the shared proxy codes. Moreover, to ameliorate the impact of noisy labels, we propose selecting small-loss examples that are considered as more confident examples and also design a mutual quantization loss to further improve the effectiveness of the sample selection. Finally, by optimizing with the selected cleaner subset, our method can significantly alleviate the impact of noisy labels. Experiments on three benchmark cross-modal datasets demonstrate that the proposed CMMQ outperforms several state-of-the-arts.

Broader Impacts. The proposed method predicts content based on learned statistics of selected training data points and as such will reflect biases in those data.

6. Acknowledgements

Our work was supported in part by the National Key R&D Program of China under Grant 2017YFE0104100, in part by the National Natural Science Foundation of China under Grant 62132016, Grant 62171343, and Grant 62071361, in part by Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03, in part by the Fundamental Research Funds for the Central Universities ZDRC2102, in part by Guangdong Basic and Applied Basic Research Foundation (2021A1515110026), in part by Natural Science Basic Research Program of Shaanxi (Program No.2022JQ-608), and in part by Australian Research Council Projects DE-190101473 and DP-220102121.

References

[1] Yacine Aït-Sahalia, Jianqing Fan, and Dacheng Xiu. High-frequency covariance estimates with noisy and asynchronous financial data. *Journal of the American Statistical Association*, 105(492):1504–1517, 2010. 1

[2] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *NeurIPS*, 34, 2021. 3

[3] Avrim Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *ACCLT*, 1998. 5

[4] Yue Cao, Bin Liu, Mingsheng Long, and Jianmin Wang. Cross-modal hamming hashing. In *ECCV*, pages 202–218, 2018. 1, 3, 5

[5] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Qiang Yang. Transitive hashing network for heterogeneous multimedia retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017. 1

[6] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S Yu. Hashnet: Deep learning to hash by continuation. In *ICCV*, pages 5608–5617, 2017. 1

[7] Youngchul Cha and Junghoo Cho. Social-network analysis using topic models. In *SIGIR*, pages 565–574, 2012. 1

[8] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *ICMR*, pages 1–9, 2009. 5

[9] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Trans. Image Process.*, 27(8):3893–3903, 2018. 3

[10] Cheng Deng, Erkun Yang, Tongliang Liu, Jie Li, Wei Liu, and Dacheng Tao. Unsupervised semantic-preserving adversarial hashing for image search. *IEEE Trans. Image Process.*, 28(8):4032–4044, 2019. 3

[11] Cheng Deng, Erkun Yang, Tongliang Liu, and Dacheng Tao. Two-stream deep hashing with class-specific centers for supervised image search. *IEEE transactions on neural networks and learning systems*, 31(6):2189–2201, 2019. 4

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[13] Yair Dgani, Hayit Greenspan, and Jacob Goldberger. Training a neural network based on unreliable human annotation of medical images. In *ISBI*, pages 39–42, 2018. 1

[14] Bo Han, Jiangchao Yao, Gang Niu, Mingyuan Zhou, Ivor W Tsang, Ya Zhang, and Masashi Sugiyama. Masking: a new perspective of noisy supervision. In *NeurIPS*, pages 5841–5851, 2018. 1, 3

[15] 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 *NeurIPS*, pages 8527–8537, 2018. 1, 3, 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5

[17] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *CVPR*, pages 5403–5413, 2021. 3

[18] Mark J Huiskes and Michael S Lew. The mir flickr retrieval evaluation. In *ICMR*, pages 39–43, 2008. 5

[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 *ICML*, pages 2304–2313, 2018. 1, 3

[20] Qingyuan Jiang and Wujun Li. Deep cross-modal hashing. In *CVPR*, pages 3232–3240, 2017. 3, 4, 5

[21] Jan Kremer, Fei Sha, and Christian Igel. Robust active label correction. In *AISTATS*, pages 308–316, 2018. 3

[22] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *CVPR*, pages 4242–4251, 2018. 5

[23] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *CVPR*, 2022. 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 5

[25] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Semantics-preserving hashing for cross-view retrieval. In *CVPR*, pages 3864–3872, 2015. 3

[26] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, pages 20331–20342, 2020. 4

[27] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Trans. Pattern Anal. Mach. Intell*, 38(3):447–461, 2016. 3, 6

[28] Xingjun Ma, Yisen Wang, Michael E Houle, Shuo Zhou, Sarah Erfani, Shutao Xia, Sudanthi Wijewickrema, and James Bailey. Dimensionality-driven learning with noisy labels. In *ICML*, pages 3355–3364, 2018. 3

[29] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *ICCV*, pages 360–368, 2017. 4

[30] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *CVPR*, pages 2233–2241, 2017. 3

[31] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, pages 4334–4343, 2018. 3

[32] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. A co-regularized approach to semi-supervised learning with multiple views. *ICML workshop on learning with multiple views*, 2005. 5

[33] Arash Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *NeurIPS*, 2017. 3

[34] Jennifer Seberry Wallis. On the existence of hadamard matrices. *Journal of Combinatorial Theory, Series A*, 21(2):188–195, 1976. 2, 4

[35] Di Wang, Xinbo Gao, Xiumei Wang, Lihuo He, and Bo Yuan. Multimodal discriminative binary embedding for large-scale cross-modal retrieval. *IEEE Trans. Image Process.*, 25(10):4540–4554, 2016. 3

[36] Runmin Wang, Guoxian Yu, Hong Zhang, Maozu Guo, Lizhen Cui, and Xiangliang Zhang. Noise-robust deep cross-modal hashing. *Information Sciences*, 581:136–154, 2021. 3

[37] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, pages 322–330, 2019. 3

[38] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. *NeurIPS*, 23:2424–2432, 2010. 1

[39] Songhua Wu, Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Nannan Wang, Haifeng Liu, and Gang Niu. Class2simi: A noise reduction perspective on learning with noisy labels. In *ICML*, pages 11285–11295, 2021. 3

[40] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, pages 6835–6846, 2019. 1

[41] Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. Discriminant cross-modal hashing. In *ICMR*, pages 305–308, 2016. 3

[42] Erkun Yang, Cheng Deng, Chao Li, Wei Liu, Jie Li, and Dacheng Tao. Shared predictive cross-modal deep quantization. *IEEE Trans. Neural Netw. Learn. Syst.*, 2018. 3

[43] Erkun Yang, Cheng Deng, Tongliang Liu, Wei Liu, and Dacheng Tao. Semantic structure-based unsupervised deep hashing. In *IJCAI*, pages 1064–1070, 2018. 1

[44] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, pages 1618–1625, 2017. 5

[45] Erkun Yang, Mingxia Liu, Dongren Yao, Bing Cao, Chunfeng Lian, Pew-Thian Yap, and Dinggang Shen. Deep bayesian hashing with center prior for multi-modal neuroimage retrieval. *IEEE IEEE Trans. Med. Imaging*, 40(2):503–513, 2020. 3

[46] Erkun Yang, Tongliang Liu, Cheng Deng, Wei Liu, and Dacheng Tao. Distillhash: Unsupervised deep hashing by distilling data pairs. In *CVPR*, pages 2946–2955, 2019. 3

[47] Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent label-noise transition matrix using dnns. *arXiv preprint arXiv:2105.13001*, 2021. 3

[48] Xu Yang, Cheng Deng, Tongliang Liu, and Dacheng Tao. Heterogeneous graph attention network for unsupervised multiple-target domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell*, 2020. 3

[49] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *CVPR*, pages 4066–4075, 2019. 1

[50] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, pages 7164–7173, 2019. 3

[51] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 1

[52] Shifeng Zhang, Jianmin Li, and Bo Zhang. Pairwise teacher-student network for semi-supervised hashing. In *CVPR workshops*, pages 0–0, 2019. 3

[53] Ying Zhang, Tao Xiang, Timothy M Hospedales, and Huchuan Lu. Deep mutual learning. In *CVPR*, pages 4320–4328, 2018. 5

[54] Mingkai Zheng, Fei Wang, Shan You, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Weakly supervised contrastive learning. In *ICCV*, pages 10042–10051, 2021. 3

[55] Mingkai Zheng, Shan You, Fei Wang, Chen Qian, Changshui Zhang, Xiaogang Wang, and Chang Xu. Ressl: Relational self-supervised learning with weak augmentation. *NeurIPS*, 34, 2021. 3