Meta-Guided Sample Reweighting for Robust Cross-Modal Hashing Retrieval with Noisy labels

Ziang Tan[1], **Weitao An**[1], **Erkun Yang**[1]*

[1]School of Electronic Engineering, Xidian University, Xi'an, China
tanziang66@gmail.com, weitaoan900@gmail.com, erkunyang@gmail.com

Abstract

Cross-modal hashing (CMH) is an effective tool for large-scale retrieval due to its low storage cost and high efficiency. However, real-world multi-modal datasets often contain noisy annotations, which can significantly impair model performance. Many existing methods address this issue by using the small-loss criterion to select a likely clean subset of data to guide model training. Nonetheless, this clean subset is typically dominated by easy samples, and treating all samples within it equally can undermine the model's generalization ability. In this paper, we propose a novel meta-learning-based framework, named Meta-Guided Sample Reweighting for Cross-Modal Hashing Retrieval (MGSH), which integrates meta-learning into robust cross-modal hashing. To address the above issues, we design a Meta-Similarity Weighting Network (MSWN) that dynamically assigns importance weights to samples during training. By employing a bi-level optimization strategy, the meta-importance weights are used to scale the loss of training samples during the main network update, encouraging the model to focus on more challenging examples. Additionally, to further distinguish between noisy and clean samples, we incorporate adaptive-margin and meta-guided center aggregation into a robust hashing loss, both guided by the learned meta-importance weights. Extensive experiments on three widely used benchmark datasets demonstrate that MGSH consistently outperforms state-of-the-art methods, validating its effectiveness.

Code — https://github.com/supertanziang/MGSH

Introduction

Cross-modal retrieval (CMR) aims to enable retrieval across different modalities. Among various CMR techniques, cross-modal hashing (CMH) has emerged as an efficient solution for large-scale retrieval due to its compact representations and high retrieval efficiency (Deng et al. 2019b; Liang et al. 2024). While existing supervised CMH methods have advanced semantic alignment by leveraging label information (Kang et al. 2025; Zhang, Li, and Wang 2025), they typically assume that all collected labels are accurate, which is unrealistic due to inevitable noisy labels from manual or non-expert annotations(Kuznetsova et al. 2020). It is a challenging problem to mitigate the performance degradation caused by noisy labels in cross-modal retrieval.

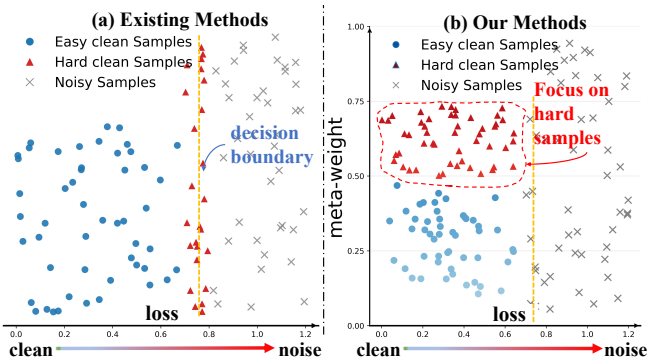


Figure 1: (a) Existing methods typically select clean samples based on a small-loss criterion, which often overlook clean but hard samples that lie near the decision boundaries. (b) In contrast, by isolating noisy samples via a meta-guided small-loss criterion, we reduce the losses of hard samples to facilitate noise separation and dynamically adjust their importance by meta-importance weights. This assigns higher weights to hard samples, enhancing model robustness.

Previous approaches to noisy cross-modal hashing mainly fall into two categories: sample selection and robust loss function design. Sample selection methods (Yang et al. 2022; Pu et al. 2025) typically adopt a small-loss criterion to explicitly identify clean samples from noisy labels, thereby mitigating the adverse effects of label noise. In parallel, robust loss function methods (Wang et al. 2024; Hu et al. 2021) aim to design noise-tolerant objectives that stabilize the training process. However, as illustrated in Figure 1, existing methods face two key challenges: (1) They treat all training samples with equal importance, which often causes the model to overfit to the majority of easy examples while neglecting hard samples located near the small-loss selection boundary within the clean subset. (2) Due to static sample selection strategies and fixed margin constraints, the selection mechanism tends to misidentify clean samples with relatively large losses as noisy ones, which ultimately degrades the model's performance.

*Corresponding author.

To address the above problems, we propose a novel method named Meta-Guided Sample Reweighting Hashing (MGSH), as illustrated in Figure 2. MGSH is a meta-learning framework with a dual-network architecture that leverages a Meta-Similarity Weighting Network (MSWN) to dynamically assign meta-importance weights to each training pairs based on their cross-modal feature representations. This sample reweighting strategy encourages the model to focus on hard instances while suppressing the influence of noisy ones. To overcome the limitations of fixed sample selection and rigid margin constraints, we further integrate adaptive margin and dynamic center aggregation mechanisms, both drived by the learned meta-importance weights. These components jointly enhance model robustness at both the instance and category levels. Overall, MGSH performs consistently well on three benchmark datasets and leads to noticeable improvements in cross-modal retrieval accuracy under noisy label conditions.

Our main contributions are summarized as follows:

- We propose a novel framework named Meta-Guided Sample Reweighting Hashing (MGSH) that leverages meta-importance weights to focus informative hard samples in clean subset, enhancing training robustness.
- We propose a robust loss that incorporates the adaptive margin and dynamic center aggregation mechanism guided by meta-importance weights. This design enhances the model's ability to distinguish noisy and clean samples, improving robustness and generalization.
- Extensive experiments on three public datasets demonstrate that MGSH consistently outperforms state-of-the-art baselines, achieving 2–5% gains in MAP under different noise levels.

Related Work

Meta Learning

Meta-learning aims to enable models to rapidly adapt to new tasks with limited supervision. Existing methods fall into three main categories: metric-based approaches (Snell, Swersky, and Zemel 2017; Allen, Xie, and Xu 2019), which learn similarity metrics over support sets; optimization-based approaches (Finn, Abbeel, and Levine 2017; Raghu et al. 2020; Mu et al. 2025), which learn initialization for fast adaptation; model-based approaches (Lee et al. 2019; Rusu et al. 2016), which employ task-specific adaptation modules. Recent efforts extend this paradigm via amortized inference (Ravi, Larochelle, and Finn 2023), benchmark standardization (Triantafillou et al. 2020), and unsupervised meta-updates (Zhang et al. 2021). Nonetheless, current methods remain constrained by sensitivity to hyperparameters, limited robustness to ambiguous tasks, and high computational cost. These limitations motivate the development of scalable and generalizable meta-learning strategies.

Cross-modal Retrieval

Cross-modal retrieval aims to learn semantically aligned representations across different modalities. Existing methods can be broadly categorized into supervised (Kang et al. 2025; Li, Long, and Yang 2025) and unsupervised approaches (Hu et al. 2022; Zhen and et al. 2019; Zhu and et al. 2019). Supervised methods leverage label information to learn more discriminative common representations. For instance, EGATH (Jin et al. 2024) employs a Transformer-based graph to integrate CLIP features with semantic priors, while InvGC (Jian and Wang 2023) enhances instance discrimination through inverse graph convolution. However, the performance of these methods heavily relies on the quality of annotations. In contrast, unsupervised methods (Deng et al. 2019a; Yang et al. 2019, 2018) rely solely on data distribution or cross-modal co-occurrence to learn shared representations without requiring explicit semantic labels. For example, UDAH (Zhu and et al. 2019) adopts adversarial training to align feature distributions across modalities. Nevertheless, the absence of reliable supervision often leads to semantic ambiguity and suboptimal retrieval performance.

Learning with Noisy Labels

Learning with noisy labels aims to prevent deep networks from overfitting corrupted labels while maintaining generalizable representations. Existing approaches can be broadly categorized into two types: (1) sample selection, which identifies clean samples or refines corrupted labels, and (2) robust optimization, which regularizes the training objective to alleviate memorization. Sample selection methods include small-loss filtering (Han et al. 2018; Li et al. 2022) and label correction through semi-supervised learning (Li, Socher, and Hoi 2020; Zhang et al. 2020). Robust optimization approaches introduce explicit regularization, where ELR (Liu et al. 2020) employs early-stage regularization to delay overfitting, and L2B (Zhou et al. 2024) leverages bootstrap consistency to improve robustness. More recent efforts incorporate contrastive learning (Yang et al. 2021) or early stopping strategies (Bai et al. 2021) to mitigate overfitting to noisy labels and improve the robustness of learned representations. In addition, unsupervised regularization has also proven effective for training under label noise (Sun et al. 2024a; Li, Xiong, and Hoi 2021).

Methodology

Problem Definition

Let $D = \{(x_i^m, \boldsymbol{y}_i)\}_{i=1}^N$ denote a multi-modal dataset containing N instances, where x_i^m denotes the i-th sample from modality $m \in \{1, 2\}$, corresponding to the image and text modalities, respectively. The label vector $\boldsymbol{y}_i \in \mathbb{R}^C$ represents the associated multi-hot vector over K semantic categories, where each entry $y_{ik} \in \{0, 1\}$ indicates whether sample i belongs to category $k \in \{1, \ldots, K\}$. In real-world scenarios, label noise is ubiquitous due to annotation errors and incomplete tagging. Accordingly, we assume the training labels $\boldsymbol{y}_i$ are potentially corrupted and contain noise.

The goal of cross-modal hashing is to learn compact binary codes that map heterogeneous data from different modalities into a shared hamming space, where semantically similar samples are assigned nearby codes, while dissimilar ones are pushed far apart. We denote the binary hash codes for modality m as $\mathbf{B}^m = \{\mathbf{b}_j^m\}_{j=1}^N$, where each

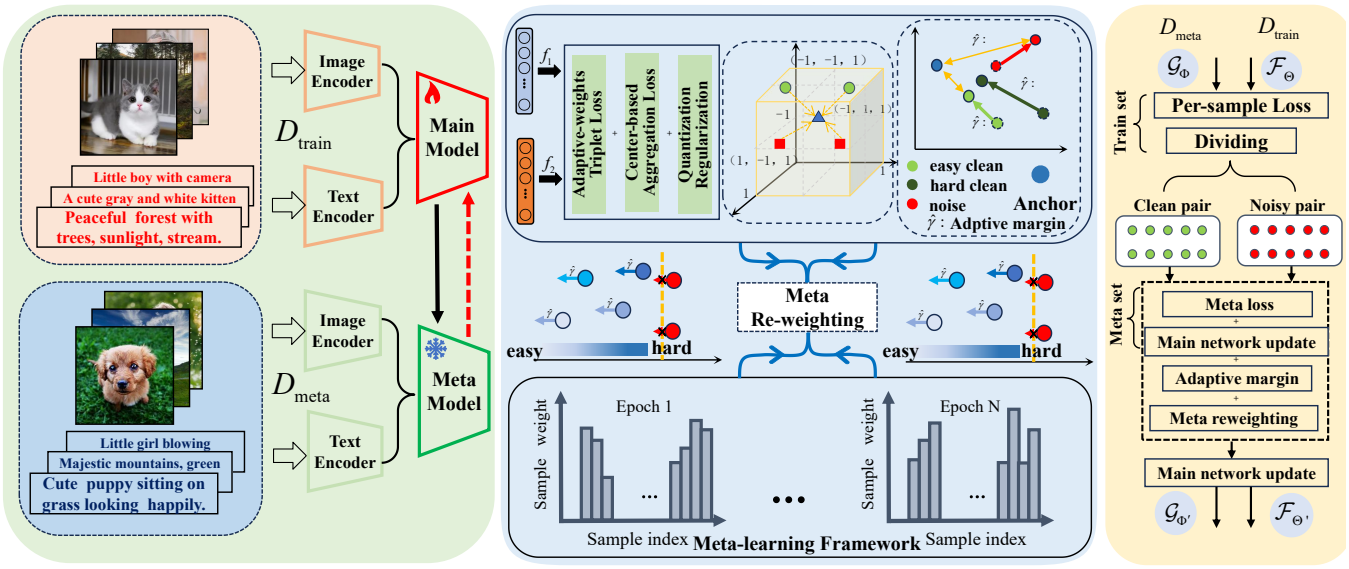


Figure 2: Overview of the MGSH framework. The framework consists of three main components: (1) Feature Extraction: A dual-stream encoder processes training and meta samples to obtain modality-specific representations. (2) Bi-level Network Architecture: Based on the meta-importance weights, the main model $\mathcal{F}_\Theta$ computes a robust hashing loss function, while the meta model $\mathcal{G}_\Phi$ reweights the samples and provides the updated weights to the main model. (3) Meta Pipeline: The meta-learning process updates the parameters of both models, enabling robust sample reweighting and adaptive margin adjustment.

$\mathbf{b}_j^m \in \{-1, +1\}^L$ represents the binary code of the j-th sample, and L denotes the code length.

To generate hash codes, we first compute modality-specific feature representations denoted by f_j^m, which are obtained by applying a modality-specific hash function to the input $f_j^m = h_m(\mathbf{x}_j^m; \theta_m)$, where $h_m(\cdot; \theta_m)$ denotes the hash function for modality m, parameterized by θ_m. Then the final binary hash code is derived by applying a sign activation to the feature vector f_j^m:

$$\mathbf{b}_j^m = \operatorname{sign}\left(f_j^m\right). \tag{1}$$

We assume the model is trained on a noisy multi-modal training set D_{train}, while a clean test set D_{test} is available for evaluation. The main network is parameterized by $\mathcal{F}_\Theta$, and the Meta-Similarity Weighting Network (MSWN) is parameterized by $\mathcal{G}_\Phi$ as a multilayer perceptron (Shu et al. 2019). To guide the training process, we additionally employ a small clean meta-dataset $D_{\text{meta}} = \{(x_i^m, y_i)\}_{i=1}^M$, where $M \ll N$ denotes the number of meta-samples.

Meta-Guided Robust Hashing Loss Function

Meta Reweighting Assignment Triplet loss (Schroff et al. 2015) is widely used in cross-modal retrieval and is typically formulated as:

$$\mathcal{L}_{\text{triplet}} = \max\left(0,\ \gamma + d_{ij} - d_{ik}\right), \tag{2}$$

where d_{ij} and d_{ik} denote the distances between the anchor and its positive and negative samples, respectively, and γ is a fixed margin. After each training epoch, we partition the training set D_{train} into a clean subset S_c and a noisy subset S_n based on the low-loss criterion, as defined by:

$$S_c = \underset{S_c : |S_c| \geq R(t)|D_{\text{train}}|}{\arg\min} \mathcal{L}_{\text{train}}(D_{\text{train}}), \tag{3}$$

which follows the approach of (Yang et al. 2022). However, since the clean subset is usually dominated by small-loss instances, the model tends to overlook hard samples near the decision boundary, which may contain informative cues. This bias prevents the model from capturing discriminative patterns around the boundary, thereby limiting its representational capacity and degrading overall performance.

To address this issue, we propose a Meta-Similarity Weighting Network (MSWN) that estimates the semantic difficulty of each sample and assigns an adaptive importance weight. Specifically, we construct matched pairs sharing identical semantic labels and mismatched pairs created by randomly shuffling the second modality. Under the noisy label settings, hard pairs near the small-loss boundary exhibit higher losses than easy clean pairs yet retain semantic consistency (Han et al. 2018). By learning from clean matched and mismatched pairs in a meta set, the MSWN captures this semantic difficulty and emphasizes hard but informative samples during training, enabling difficulty-aware weighting for robust cross-modal representation learning.

Accordingly, given each constructed feature pair (f_i^1, f_i^2), we assign a binary label $y_i \in \{0, 1\}$, where $y_i = 0$ denotes a matched (easy) pair and $y_i = 1$ denotes a mismatched (hard) pair. We use its one-hot encoding $\mathbf{y}_i = (y_{i0}, y_{i1})$ to represent the ground-truth class over $c \in \{0, 1\}$. Given a meta-batch of m pairs, the MSWN is then trained as a binary classifier by minimizing the following cross-entropy loss:

$$\mathcal{L}_{\text{meta}} = -\frac{1}{m} \sum_{i=1}^{m} \sum_{c=0}^{1} y_{ic} \log p_\Phi\left(c \mid f_i^1, f_i^2\right), \tag{4}$$

where the class probability is obtained by applying a soft-

max to the logits produced by the MSWN:

$$p_\Phi\left(c \mid f_i^1, f_i^2\right) = \frac{\exp\left(\mathcal{G}_\Phi(f_i^1, f_i^2)_c\right)}{\sum_{k=0}^{1} \exp\left(\mathcal{G}_\Phi(f_i^1, f_i^2)_k\right)}. \quad (5)$$

Here, the MSWN $\mathcal{G}_\Phi(\cdot)$ outputs a two-dimensional logit vector corresponding to the easy and hard classes. This training objective encourages the MSWN to distinguish between easy and hard pairs, thereby capturing the latent alignment difficulty in the multi-modal feature space. The optimal parameters of the MSWN are obtained by minimizing this loss:

$$\Phi^* = \arg\min_{\Phi} \mathcal{L}_{\text{meta}}(B_{\text{meta}}; \Phi). \quad (6)$$

Then we obtain meta-importance weight $\hat{\omega}_i$ of each sample by computing the softmax probability of the hard class:

$$\hat{\omega}_i = \frac{\exp\left(\mathcal{G}_\Phi(f_i^1, f_i^2)_1\right)}{\sum_{c=0}^{1} \exp\left(\mathcal{G}_\Phi(f_i^1, f_i^2)_c\right)}, \quad (7)$$

which quantifies how likely the sample is considered difficult. To stabilize training and maintain consistent scaling, we normalize $\hat{\omega}_i$ within each mini-batch:

$$\omega_i = \frac{\max\left(\hat{\omega}_i, 0\right)}{\sum_{i=1}^{m} \max\left(\hat{\omega}_i, 0\right)}, \quad (8)$$

where $\omega_i \in (0, 1)$. In the subsequent robust loss formulation, this weight adaptively modulates each sample's contribution, allowing the model to emphasize informative hard samples while suppressing noisy ones, thereby enhancing robustness against label noise.

Adaptive Margin Triplet Loss In standard triplet loss, a fixed margin is applied uniformly to both clean and noisy samples, which often inflates the loss of clean yet hard examples. Consequently, these informative hard samples are prone to being misidentified as noisy under the small-loss selection, thereby degrading overall performance. To mitigate this issue, we propose the adaptive-margin triplet loss:

$$\mathcal{L}_w = \max\left(0,\ \hat{\gamma}_i + d_{ij} - d_{ik}\right), \quad (9)$$

where the adaptive margin $\hat{\gamma}_i$ is computed as:

$$\hat{\gamma}_i = \frac{\gamma_i}{1 + \left(\frac{1}{\omega_i} - 1\right)^{-\tau}}, \quad (10)$$

with $\tau > 0$ being a hyperparameter. Since larger meta-importance weights ω_i indicate harder samples, the self-adaptive margin mechanism assigns them smaller margins from the equation, reducing their triplet loss. This helps preserve informative hard samples by reducing their loss, making them less likely to be mistakenly excluded as noisy under the small-loss criterion.

Dynamic Semantic Center Update To reduce intra-class variations in multi-modal representations, we introduce dynamic semantic centers $\mathbf{c}_k \in \mathbb{R}^L$ for each class $k \in \{1, \ldots, K\}$. These centers are updated as the weighted average of all sample representations belonging to class k. Specifically, given a meta-importance weight $\omega_i \in (0, 1)$ for each sample, the center $\mathbf{c}_k$ is computed as:

$$\mathbf{c}_k = \frac{\sum_{i \in C_k} \omega_i \cdot (f_i^1 + f_i^2)}{|C_k|}, \quad (11)$$

where $|C_k|$ denotes the set of samples assigned to class k. The use of ω_i ensures that hard samples contribute more to the center update, thereby promoting a more robust and discriminative representation.

To evaluate the semantic alignment between a sample and its corresponding class center, we define a confidence score v_i^m for each modality $m \in \{1, 2\}$ as:

$$v_i^m = \sum_{k=1}^{K} y_{ik} \frac{\exp\left(f_i^{m\top} \mathbf{c}_k / t\right)}{\sum_{j=1}^{K} \exp\left(f_i^{m\top} \mathbf{c}_j / t\right)}, \quad (12)$$

where t is a temperature parameter that controls the sharpness of the softmax distribution, and y_{ik} is the one-hot label indicating whether sample i belongs to class k.

Inspired by the self-paced learning paradigm (Pu et al. 2025; Sun et al. 2024b), we define the center aggregation loss using an r-generalized cross-entropy formulation:

$$\mathcal{L}_c = \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{2} \left[(1 - r) \frac{1 - (v_i^m)^r}{r} + r\left(1 - v_i^m\right) \right], \quad (13)$$

where $r \in (0, 1]$ is a hyperparameter controlling the trade-off between sensitivity and robustness.

Overall Optimization To further regularize the learning process and minimize quantization error during hash code generation, we introduce a quantization loss over the continuous hash representations $f_{j,k}^m \in \mathbb{R}$:

$$\mathcal{L}_q = \frac{1}{L} \sum_{k=1}^{L} \left(\left|f_{j,k}^m\right| - 1\right)^2, \quad (14)$$

where L is the hash code length. This term encourages each hash value to approach binary states (± 1), ensuring more stable and discriminative code generation.Thus, the final robust training loss that is used to guide the separation of noisy and clean samples is formulated as:

$$\mathcal{L}_{\text{train}} = \mathcal{L}_w + \lambda_1 \mathcal{L}_c + \lambda_2 \mathcal{L}_q, \quad (15)$$

where λ_1 and λ_2 are loss weighting hyperparameters. Through joint optimization, the model achieves a better balance between robustness and category semantic alignment.

Bi-level Optimization Process Solving the bi-level optimization problem in our MGSH framework is computationally expensive. To reduce complexity, we approximate the inner-level update by performing a single-step gradient descent on the main network parameters Θ, followed by an alternating update of the MSWN parameters Φ within each training iteration. The overall training process is summarized in Algorithm 1.

To simulate the impact of sample weights on the model's generalization capability, we first perform a temporary gradient update to obtain temporarily updated parameters Θ' :

$$\Theta' = \Theta - \alpha \nabla_\Theta \mathcal{L}_{\text{train}}(B_{\text{train}}; \Theta), \quad (16)$$

where α is the learning rate of the main network. Based on the temporarily updated main network, we then compute the hash presentation (f_i^1, f_i^2) and update the parameters of the meta-network as follows:

$$\Phi' = \Phi - \eta \nabla_{\Phi} \mathcal{L}_{\text{meta}}, \tag{17}$$

where η is the learning rate of the meta network. Finally, the main network is updated by minimizing the meta-weighted overall training loss, where each sample loss is scaled by its corresponding meta-importance weight ω_i:

$$\Theta_{t+1} = \Theta_t - \alpha \nabla_{\Theta} \left(\sum_{i=1}^{n} \omega_i \mathcal{L}_{\text{train}}(B_{\text{train}}; \Theta) \right). \tag{18}$$

Algorithm 1: Meta-Guided Sample Reweighting for Robust Cross-Modal Hashing Retrieval

Require: Training data D_{train}, meta data D_{meta}, initial parameters Θ_0, Φ_0, initial margin γ, hyperparameters λ_1, λ_2, r, τ, maximal epoch number T.

Initialize network $\mathcal{F}_{\Theta}$ and $\mathcal{G}_{\Phi}$; set $\Theta \leftarrow \Theta_0$, $\Phi \leftarrow \Phi_0$.
Warm up the network on all training data using $\mathcal{L}_{\text{train}}$.
for epoch $t = 0$ **to** $T - 1$ **do**
  Compute per-sample loss via Eq. 15.
  Separate noisy and clean samples via Eq. 3.
  Partition D_{train} into clean/noisy sets (S_c, S_n).
  for $j = 1$ **to** num_steps **do**
    Sample mini-batch $B_{\text{train}} \sim S_c$.
    Sample meta-batch $B_{\text{meta}} \sim D_{\text{meta}}$.
    Update main temporary parameter via Eq. 16.
    Update meta network parameter Φ via Eq. 6.
    Compute meta-importance weight $\hat{\omega}_i$ by Eq. 7.
    Normalized meta-importance weight ω_i by Eq. 8.
    Compute adaptive margin $\hat{\gamma}$ via Eq. 10.
    Update final main network parameters via Eq. 18.
  end for
end for

Ensure: Final parameters Θ_T, Φ_T.

Experiments

Datasets

To validate the effectiveness of MGSH, we conduct extensive experiments on three widely used benchmark datasets: MS-COCO (Lin et al. 2014), NUS-WIDE (Chua et al. 2009), and MIRFlickr-25K (Huiskes and Lew 2008). The detail of these datasets is described in Supplementary Material.

Experimental Settings

To evaluate the performance of MGSH, we conduct experiments on two standard cross-modal retrieval tasks: I2T (retrieving texts from image queries) and T2I (retrieving images from text queries). Following previous works, we use Mean Average Precision (MAP) as the evaluation metric, which is widely used in cross-modal retrieval.

To comprehensively examine the robustness of our method, we introduce symmetric label noise at different noise rates, specifically 0.2, 0.5, and 0.8. The hash code lengths are set to 16, 32, 64, and 128 bits. All experiments are conducted on a single NVIDIA GeForce RTX A6000 GPU with 48GB of memory. Additional implementation details are provided in the sup due to space limitations.

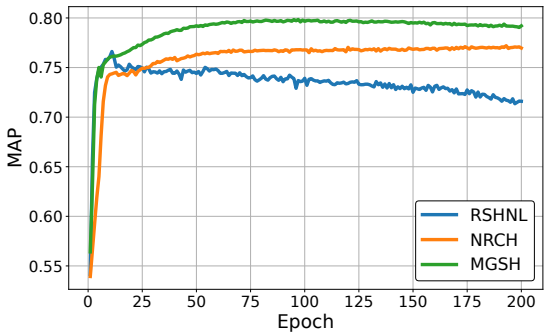


Figure 3: MAP scores versus epochs on MIRFlickr-25K dataset with 128-bit hash codes under 0.5 noise rate.

Comparison with State-of-the-Art Methods

To demonstrate the superiority of our proposed method, we compare MGSH with several existing cross-modal hashing methods. Specifically, we select a representative set of baseline methods, including classical supervised methods such as DCMH (Jiang and Li 2017), CMHH (Cao et al. 2018), CMMQ (Yang et al. 2022), NRCH (Wang et al. 2024), and RSHNL (Pu et al. 2025), as well as classical unsupervised methods such as DGCPN (Yu et al. 2021), DJSRH (Su, Zhong, and Zhang 2019), and UCCH (Hu et al. 2022). Among these, CMMQ, NRCH, and RSHNL are specifically designed to address cross-modal retrieval under noisy label scenarios. We pay particular attention to comparing MGSH with NRCH and RSHNL, which represent the best-performing recent cross-modal hashing methods under noisy supervision, as illustrated in Figure 3.

For fair comparison, we re-train all baseline methods under the same hardware environment. The model with the highest MAP on the validation set is selected for evaluation on the test set. All methods use the same training and testing splits and are implemented under identical hardware settings. In the experimental tables, the highest MAP value is highlighted in bold, while the second highest is underlined. In addition, we plot Precision-Recall (PR) curves for different datasets, as shown in Figure 4.

From the experimental results, we observe the following:

- As the noise rate increases, the MAP performance of all supervised cross-modal hashing methods degrades. In contrast, unsupervised methods (UCCH, DJSRH, DGCPN) remain stable because they do not depend on noisy labels. However, their lack of label supervision also limits their achievable retrieval performance.
- Among all baseline methods, increasing the hash code length consistently leads to higher MAP scores. Regarding noise separation, both NRCH and RSHNL achieve comparable performance to our method, but our method consistently outperforms them.

Dataset	Method	20%				50%				80%			
		16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
MS-COCO	DJSRH (ICCV'19)	0.482	0.520	0.542	0.568	0.482	0.520	0.542	0.568	0.482	0.520	0.542	0.568
	DGCPN (AAAI'21)	0.589	0.609	0.620	0.626	0.589	0.609	0.620	0.627	0.589	0.609	0.620	0.627
	UCCH (TPAMI'23)	0.570	0.580	0.593	0.621	0.570	0.580	0.593	0.621	0.570	0.580	0.593	0.621
	DCMH (CVPR'17)	0.545	0.579	0.583	0.597	0.480	0.470	0.474	0.460	0.396	0.387	0.358	0.353
	CMHH (ECCV'18)	0.643	0.648	0.650	0.652	0.603	0.608	0.612	0.611	0.601	0.576	0.569	0.587
	CMMQ (CVPR'22)	0.614	0.638	0.642	0.642	0.595	0.622	0.626	0.633	0.600	0.616	0.626	0.634
	NRCH (ACMMM'24)	0.637	0.678	0.686	0.690	0.651	0.667	0.681	0.693	0.630	0.655	0.666	0.687
	RSHNL (AAAI'25)	0.555	0.594	0.629	0.632	0.560	0.594	0.608	0.630	0.569	0.590	0.602	0.641
	MGSH (ours)	**0.678**	**0.708**	**0.714**	**0.724**	**0.666**	**0.689**	**0.714**	**0.721**	**0.661**	**0.673**	**0.700**	**0.702**
	Δ	+3.5%	+3.0%	+2.8%	+3.4%	+1.5%	+2.2%	+3.3%	+2.8%	+3.1%	+1.8%	+3.4%	+1.5%
MIRFlickr25K	DJSRH (ICCV'19)	0.613	0.621	0.631	0.639	0.613	0.621	0.631	0.639	0.613	0.621	0.631	0.639
	DGCPN (AAAI'21)	0.688	0.693	0.704	0.710	0.688	0.693	0.704	0.710	0.688	0.693	0.704	0.710
	UCCH (TPAMI'23)	0.689	0.711	0.712	0.716	0.689	0.711	0.712	0.716	0.689	0.711	0.712	0.716
	DCMH (CVPR'17)	0.703	0.701	0.703	0.698	0.651	0.645	0.639	0.630	0.629	0.622	0.617	0.616
	CMHH (ECCV'18)	0.692	0.695	0.702	0.700	0.653	0.638	0.639	0.646	0.607	0.599	0.602	0.608
	CMMQ (CVPR'22)	0.726	0.730	0.736	0.738	0.698	0.718	0.720	0.724	0.694	0.712	0.716	0.720
	NRCH (ACMMM'24)	0.746	0.760	0.762	0.760	0.740	0.758	0.765	0.760	0.731	0.755	0.757	0.762
	RSHNL (AAAI'25)	0.683	0.710	0.717	0.716	0.696	0.700	0.717	0.714	0.685	0.714	0.719	0.716
	MGSH (ours)	**0.758**	**0.775**	**0.790**	**0.795**	**0.754**	**0.774**	**0.784**	**0.790**	**0.752**	**0.776**	**0.780**	**0.784**
	Δ	+1.2%	+1.5%	+2.8%	+3.5%	+1.4%	+1.6%	+1.9%	+3.0%	+2.1%	+2.1%	+2.3%	+2.2%
NUS-WIDE	DJSRH (ICCV'19)	0.417	0.451	0.465	0.492	0.417	0.451	0.465	0.492	0.417	0.451	0.465	0.492
	DGCPN (AAAI'21)	0.570	0.592	0.611	0.623	0.570	0.592	0.611	0.623	0.570	0.592	0.611	0.623
	UCCH (TPAMI'23)	0.575	0.597	0.623	0.636	0.575	0.597	0.623	0.636	0.575	0.597	0.623	0.636
	DCMH (CVPR'17)	0.494	0.496	0.494	0.479	0.451	0.448	0.443	0.431	0.408	0.402	0.397	0.393
	CMHH (ECCV'18)	0.572	0.575	0.576	0.580	0.563	0.570	0.566	0.568	0.493	0.495	0.503	0.505
	CMMQ (CVPR'22)	0.633	0.637	0.647	0.654	0.586	0.598	0.608	0.623	0.538	0.565	0.579	0.585
	NRCH (ACMMM'24)	0.621	0.630	0.632	0.641	0.553	0.569	0.587	0.602	0.523	0.512	0.557	0.557
	RSHNL (AAAI'25)	0.604	0.606	0.641	0.650	0.597	0.602	0.616	0.618	0.543	0.580	0.598	0.589
	MGSH (ours)	**0.683**	**0.692**	**0.700**	**0.705**	**0.625**	**0.658**	**0.682**	**0.676**	**0.582**	**0.614**	**0.625**	**0.639**
	Δ	+5.0%	+5.5%	+5.3%	+5.1%	+2.8%	+5.6%	+6.6%	+4.0%	+0.7%	+1.7%	+0.2%	+0.3%

Table 1: Average MAP scores of I2T and T2I tasks under 20%, 50%, and 80% noise rates (16/32/64/128 bit) on the MS-COCO, MIRFlickr25K, and NUS-WIDE datasets. The highest and second-highest scores are shown in bold and underlined, respectively. "Δ" is the relative gain of MGSH over the second-highest (%) .

- For the PR curves, our method outperforms most baseline methods across the majority of settings on all three benchmark datasets. In terms of MAP, our MGSH achieves performance gains of 2.7%, 3.6%, and 2.2% in average on the respective datasets. The superior PR curves observed in most cases further demonstrate the effectiveness and robustness of our proposed method.

Ablation Study

To validate the effectiveness of each proposed component, we conduct ablation studies on the MS-COCO dataset using 128-bit hash codes under different noise rates (Table 2). Specifically, we analyze the impact of three components: (1) warm-up phase, (2) the meta-pipeline for meta-importance weight ω_i, (3) the adaptive margin $\hat{\gamma}$.

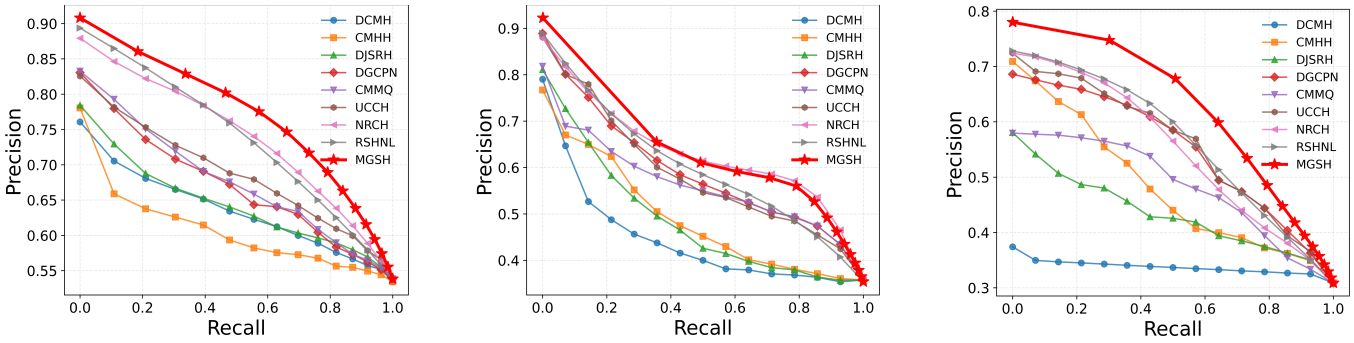
Meta-pipeline: Removing the meta-learning pipeline causes the largest performance drop (e.g., I2T: 0.728→0.683, T2I: 0.720→0.698 at 20% noise), showing that meta-weighted sample reweighting is crucial for suppressing noise and retaining informative examples.

Warmup: Disabling the warm-up phase also degrades performance under low-noise settings (e.g., I2T: 0.728→0.710 at 20% noise), indicating its role in stabilizing early training and supporting reliable meta-optimization.

Configuration			Image → Text (MAP)		
Warmup	Meta-pipeline	$\hat{\gamma}$	0.2	0.5	0.8
✓		✓	0.683	0.690	0.686
✓	✓		0.718	0.699	0.693
	✓	✓	0.710	0.705	0.689
✓	✓	✓	**0.728**	**0.724**	**0.702**
Configuration			**Text → Image (MAP)**		
Warmup	Meta-pipeline	$\hat{\gamma}$	0.2	0.5	0.8
✓		✓	0.698	0.696	0.687
✓	✓		0.716	0.706	0.691
	✓	✓	0.717	0.704	0.695
✓	✓	✓	**0.720**	**0.718**	**0.703**

Table 2: Ablation study on MS-COCO dataset.

Adaptive margin $\hat{\gamma}$: Excluding the adaptive margin leads to consistent drops across noise levels (e.g., I2T: 0.724→0.699 at 50% noise), validating its effectiveness in preventing over-penalization of hard but clean samples.



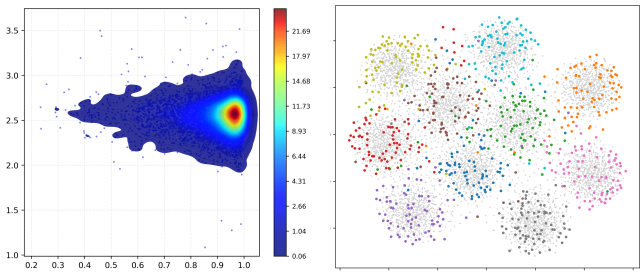
(a) MIRFlickr-25K (b) MS-COCO (c) NUS-WIDE

Figure 4: PR curves on three datasets with 128-bit hash codes under 50% label noise, averaged of I2T and T2I retrieval tasks.

Parameter	Default	Range	MAP
Margin (γ)	0.5	$[0.4, 0.6, 0.7]$	**0.788**$[0.786, 0.788, 0.785]$
λ_1	0.5	$[0.3, 0.4, 0.6]$	**0.788**$[0.782, 0.784, 0.785]$
λ_2	0.5	$[0.3, 0.4, 0.6]$	**0.788**$[0.781, 0.783, 0.786]$
r	0.7	$[0.5, 0.6, 0.8]$	**0.788**$[0.786, 0.785, 0.783]$

Meta-Dataset Size	Image $\longrightarrow$ Text	Text $\longrightarrow$ Image
1%	0.800 ± 0.003	0.777 ± 0.003
2%	0.798 ± 0.002	0.776 ± 0.003
5%	0.800 ± 0.005	0.777 ± 0.004
10%	0.799 ± 0.004	0.776 ± 0.005

Table 3: Sensitivity analysis of key hyperparameters (top) and effect of meta-dataset size (bottom).



(a) Meta Weights vs Loss (b) Visualize hard examples

Figure 5: Visualizing hard examples under NUS-WIDE dataset with 50% noise rate.

Sensitivity Analysis

In this section, we conduct a sensitivity analysis of the key hyperparameters in the proposed MGSH framework. Specifically, we focus on three main hyperparameters: the initial margin γ, the loss weighting hyperparameters λ_1 and λ_2, the weight factor r in dynamic center aggregation loss. All experiments in this analysis are conducted on the MIRFlickr-25K dataset under symmetric noise with a noise rate of 0.5 and a hash code length of 128 bits.

Table 3 presents a sensitivity analysis demonstrating that the default hyperparameter settings consistently achieve optimal MAP scores, highlighting the robustness of adaptive margin mechanism(γ) and dynamic center aggregation (r) approach. Additionally, varying the meta-dataset size between 1% and 10% results in negligible performance fluctuations, underscoring the efficiency and scalability of the proposed MGSH framework in noisy cross-modal hashing.

Analyzing Independence of MGSH

We visualize the relationship between the meta-importance weights of our proposed MGSH framework and the classification loss, as shown in Figure 5a. It can be observed that the meta-importance weights and the loss function are not directly correlated, indicating that the meta-learned weights provide an independent method for identifying hard samples without relying on the loss value.

In Figure 5b, we further visualize the hard samples identified by the MGSH framework on the NUS-WIDE dataset using the t-SNE method, with the ten most frequent classes shown in the figure. The gray circles represent simple samples that are concentrated in the feature space, while the colorful circles correspond to the hard examples selected by MGSH. It can be clearly observed that these hard samples mostly lie at the category boundaries, forming a ring-like structure. This demonstrates the core mechanism of MGSH: it effectively captures samples located near the decision boundaries that contain rich discriminative information and are inherently difficult to distinguish, thereby enhancing the model's discriminative ability and robustness.

Conclusion

In this paper, we propose MGSH, a meta-learning based framework for addressing noisy labels in cross-modal hashing retrieval. MGSH leverages a Meta-Similarity Weighting Network (MSWN) to dynamically assign reliability-aware sample weights through bi-level optimization, adaptively controlling the contribution of each training sample. It further incorporates an adaptive-margin mechanism and meta-guided center aggregation into a robust loss formulation, enhancing the discrimination of noisy and hard samples. Extensive experiments on three benchmark datasets demonstrate that MGSH consistently outperforms existing methods under various noise levels in cross-modal retrieval.

Acknowledgments

Our work was supported in part by the National Natural Science Foundation of China (62202365) and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

Allen, J.; Xie, L.; and Xu, Y. 2019. Few-shot Learning with Graph Neural Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 3348–3357.

Bai, Y.; Yang, E.; Han, B.; Yang, Y.; Li, J.; Mao, Y.; Niu, G.; and Liu, T. 2021. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34: 24392–24403.

Cao, Y.; Liu, B.; Long, M.; and Wang, J. 2018. Cross-modal hamming hashing. In *Proceedings of the European conference on computer vision (ECCV)*, 202–218.

Chua, T.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, 1–9. ACM.

Deng, C.; Yang, E.; Liu, T.; Li, J.; Liu, W.; and Tao, D. 2019a. Unsupervised semantic-preserving adversarial hashing for image search. *IEEE transactions on image processing*, 28(8): 4032–4044.

Deng, C.; Yang, E.; Liu, T.; and Tao, D. 2019b. Two-stream deep hashing with class-specific centers for supervised image search. *IEEE Transactions on Neural Networks and Learning Systems*, 31(6): 2189–2201.

Finn, C.; Abbeel, P.; and Levine, S. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 1126–1135.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I. W.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, volume 31.

Hu, P.; Peng, X.; Zhu, H.; Zhen, L.; and Lin, J. 2021. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5403–5413.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.-P.; and Peng, X. 2022. Unsupervised contrastive cross-modal hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3877–3889.

Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 39–43.

Jian, X.; and Wang, Y. 2023. InvGC: Robust Cross-Modal Retrieval by Inverse Graph Convolution. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 836–865.

Jiang, Q.-Y.; and Li, W.-J. 2017. Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3232–3240.

Jin, H.; Zhang, Y.; Shi, L.; Zhang, S.; Kou, F.; Yang, J.; Zhu, C.; and Luo, J. 2024. An End-to-End Graph Attention Network Hashing for Cross-Modal Retrieval. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Kang, X.; Liu, X.; Zhang, X.; Xue, W.; Nie, X.; and Yin, Y. 2025. Semi-Supervised Online Cross-Modal Hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; et al. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7): 1956–1981.

Lee, K.; Maji, S.; Ravichandran, A.; and Soatto, S. 2019. Meta-Learning with Differentiable Convex Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10657–10665.

Li, J.; Socher, R.; and Hoi, S. C. 2020. DivideMix: Learning with noisy labels as semi-supervised learning. In *ICLR*.

Li, J.; Xiong, C.; and Hoi, S. 2021. Learning from Noisy Data with Robust Representation Learning. In *ICCV*.

Li, N. K.; Rizve, M. N.; Rahnavard, N.; Mian, A.; and Shah, M. 2022. UNICON: Combating Label Noise Through Uniform Selection and Contrastive Learning. In *CVPR*, 9676–9686.

Li, Y.; Long, J.; and Yang, Z. 2025. Asymmetric Cross-Modal Hashing Based on Formal Concept Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.

Liang, X.; Yang, E.; Yang, Y.; and Deng, C. 2024. Multi-relational deep hashing for cross-modal search. *IEEE Transactions on Image Processing*.

Lin, T.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *Lecture Notes in Computer Science, vol. 8693, Computer Vision – ECCV 2014*, 740–755. Springer.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. In *NeurIPS*, volume 33, 20331–20342.

Mu, C.; Qu, Y.; Yan, J.; Yang, E.; and Deng, C. 2025. Meta-Learning Dynamic Center Distance: Hard Sample Mining for Learning with Noisy Labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 415–425.

Pu, R.; Sun, Y.; Qin, Y.; Ren, Z.; Song, X.; Zheng, H.; and Peng, D. 2025. Robust Self-Paced Hashing for Cross-Modal Retrieval with Noisy Labels. *arXiv preprint arXiv:2501.01699*.

Raghu, A.; Raghu, M.; Bengio, S.; and Vinyals, O. 2020. Rapid learning or feature reuse? Towards understanding the effectiveness of MAML. In *International Conference on Learning Representations (ICLR)*.

Ravi, S.; Larochelle, H.; and Finn, C. 2023. Amortized Meta-Learning with Structured Latent Spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Rusu, A. A.; Nica, E.; Rabinowitz, N.; Parisotto, E.; He, H.; Zhang, L.; Fracchia, F.; Vezhnevets, A.; Kumaran, D.; and Hinton, G. E. 2016. Progressive Neural Networks. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS)*, 1397–1405.

Schroff, F.; Kalenichenko; Dmitry; Philbin; and James. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

Shu, J.; Xie, Q.; Yi, L.; Zhao, Q.; Zhou, S.; Xu, Z.; and Meng, D. 2019. Meta-Weight-Net: Learning an explicit mapping for sample weighting. In *NeurIPS*, volume 32.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 4077–4087.

Su, S.; Zhong, Z.; and Zhang, C. 2019. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3027–3035.

Sun, X.; Zhang, S.; Liu, J.; Li, Y.; and Wang, J. 2024a. Prediction Consistency Regularization for Learning with Noisy Labels. *Entropy*, 26(4): 308.

Sun, Y.; Dai, J.; Ren, Z.; Chen, Y.; Peng, D.; and Hu, P. 2024b. Dual Self-Paced Cross-Modal Hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15184–15192.

Triantafillou, E.; Zhu, T.; Dumoulin, V.; Lamblin, P.; Evci, U.; Xu, K.; Goroshin, R.; Gelada, C.; Swersky, K.; Tarlow, D.; et al. 2020. Meta-Dataset: A Dataset of Datasets for Learning to Learn from Few Examples. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Wang, L.; Qin, Y.; Sun, Y.; Peng, D.; Peng, X.; and Hu, P. 2024. Robust Contrastive Cross-modal Hashing with Noisy Labels. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5752–5760.

Yang, E.; Deng, C.; Liu, T.; Liu, W.; and Tao, D. 2018. Semantic structure-based unsupervised deep hashing. In *Proceedings of the 27th international joint conference on artificial intelligence*, 1064–1070.

Yang, E.; Liu, T.; Deng, C.; Liu, W.; and Tao, D. 2019. Distillhash: Unsupervised deep hashing by distilling data pairs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2946–2955.

Yang, E.; Yao, D.; Liu, T.; and Deng, C. 2022. Mutual quantization for cross-modal search with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7551–7560.

Yang, Y.; Sun, Z.; Zhang, C.; Shen, F.; Wu, Q.; Zhang, J.; and Tang, Z. 2021. Jo-SRC: A Contrastive Approach for Combating Noisy Labels. In *CVPR*, 5192–5201.

Yu, J.; Zhou, H.; Zhan, Y.; and Tao, D. 2021. Deep graph-neighbor coherence preserving network for unsupervised cross-modal hashing. In *Proceedings of the AAAI conference on artificial intelligence*, 5, 4626–4634.

Zhang, X.; Li, Q.; and Wang, L. 2025. Vision-guided Text Mining for Unsupervised Cross-Modal Hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39.

Zhang, Y.; Vinyals, O.; Eslami, S. M. A.; and Heess, N. 2021. Meta-Learning Update Rules for Unsupervised Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14164–14173.

Zhang, Z.; Zhang, H.; Arik, S. O.; Lee, H.; and Pfister, T. 2020. Distilling effective supervision from severe label noise. In *CVPR*, 9294–9303.

Zhen, L.; and et al. 2019. Joint Cross-Modal Correlation and Semantic Preserving Hashing for Unsupervised Cross-Modal Retrieval. *IEEE Transactions on Multimedia*.

Zhou, Y.; Li, X.; Liu, F.; Wei, Q.; Chen, X.; Yu, L.; Xie, C.; Lungren, M. P.; and Xing, L. 2024. L2B: Learning to Bootstrap Robust Models for Combating Label Noise. In *CVPR*.

Zhu, Y.; and et al. 2019. Unsupervised Deep Adversarial Hashing for Cross-Modal Retrieval. In *ACM MM*.