Semantic-Consistent Bidirectional Contrastive Hashing for Noisy Multi-Label Cross-Modal Retrieval

Likang Peng[1], Chao Su[1], Wenyuan Wu[1], Yuan Sun[2], Dezhong Peng[1,3], Xi Peng[1,3], Xu Wang[1,4*]

[1]The College of Computer Science, Sichuan University, Chengdu, China
[2]National Key Laboratory of Fundamental Algorithms and Models
for Engineering Numerical Simulation, Sichuan University, Chengdu, China
[3]Tianfu Jincheng Laboratory, Chengdu, China
[4]Centre for Frontier AI Research (CFAR), A*STAR, Singapore
likangpeng@stu.scu.edu.cn, suchao.ml@gmail.com, wuwenyuan97@gmail.com, sunyuan_work@163.com,
pengdz@scu.edu.cn, pengx.gm@gmail.com, wangxu.scu@gmail.com

Abstract

Cross-modal hashing (CMH) facilitates efficient retrieval across different modalities (e.g., image and text) by encoding data into compact binary representations. While recent methods have achieved remarkable performance, they often rely heavily on fully annotated datasets, which are costly and labor-intensive to obtain. In real-world scenarios, particularly in multi-label datasets, label noise is prevalent and severely degrades retrieval performance. Moreover, existing CMH approaches typically overlook the partial semantic overlaps inherent in multi-label data, limiting their robustness and generalization. To tackle these challenges, we propose a novel framework named Semantic-Consistent Bidirectional Contrastive Hashing (SCBCH). The framework comprises two complementary modules: (1) Cross-modal Semantic-Consistent Classification (CSCC), which leverages cross-modal semantic consistency to estimate sample reliability and reduce the impact of noisy labels; (2) Bidirectional Soft Contrastive Hashing (BSCH), which dynamically generates soft contrastive sample pairs based on multi-label semantic overlap, enabling adaptive contrastive learning between semantically similar and dissimilar samples across modalities. Extensive experiments on four widely-used cross-modal retrieval benchmarks validate the effectiveness and robustness of our method, consistently outperforming state-of-the-art approaches under noisy multi-label conditions.

Code — https://github.com/Plke/SCBCH

Introduction

The landscape of digital data has expanded from traditional text to multimodal content like images and videos, driven by internet advancements that complicate information retrieval. Cross-modal retrieval (Su et al. 2025b; Liu et al. 2024) tackles this challenge by enabling searches across diverse modalities. For scalable and efficient large-scale retrieval across modalities, cross-modal hashing (Jiang and Li 2017; Lan et al. 2025) transforms heterogeneous data into
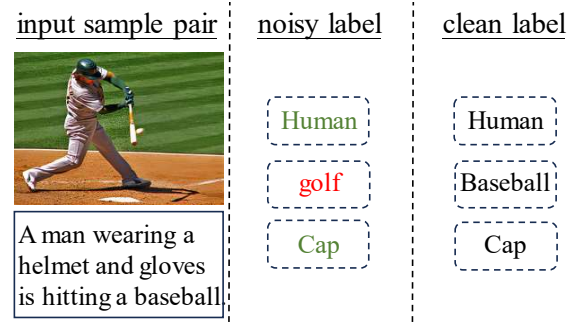

Figure 1: Cross-modal (image-text) pair with noisy multi-label annotations, where green indicates correct labels and red indicates incorrect ones.

a shared Hamming space through compact binary codes, allowing fast search with low computational overhead.

However, most existing methods rely on clean and fully annotated training data (Zhu et al. 2024; Yin et al. 2025; Wu et al. 2025). Such data are expensive to obtain in the real world, due to inherent data complexity (Frenay and Verleysen 2014) and human annotation errors (Song et al. 2022). While unsupervised methods reduce labeling costs, they often underperform compared to supervised approaches. A practical compromise is to use non-expert annotations, which inevitably introduce label noise, particularly in cross-modal multi-label scenarios.

As illustrated in Figure 1, an image-text pair is annotated with three labels, one of which is noisy: the label "golf" is incorrect due to an annotation error. While robust learning strategies such as sample selection (Song et al. 2022) and loss correction (Wang et al. 2019) have been explored in single-modal tasks, they are not directly applicable to cross-modal settings due to modality heterogeneity. To address label noise in cross-modal retrieval while maintaining scalability and retrieval efficiency, several noise-robust hash learning methods (Shu et al. 2024; Su et al. 2024; Cao et al. 2024) have been proposed. Although these methods aim to improve robustness, they often introduce new trade-

---

offs. For example, (Shu et al. 2024) mitigates noise by excluding suspected noisy samples from training, which may inadvertently discard valuable supervisory signals. Meanwhile, (Wang et al. 2024) treats any pair with at least one shared label as a positive pair, and all others as negative. This binary treatment may overlook the nuanced semantic relationships common in multi-label scenarios, potentially resulting in imbalanced pair construction and underutilization of ambiguous yet informative samples.

To overcome these limitations, we propose a novel framework called Semantic-Consistent Bidirectional Contrastive Hashing (SCBCH), which jointly leverages Cross-modal Semantic-Consistent Classification (CSCC) and Bidirectional Soft Contrastive Hashing (BSCH) objective to improve robustness under noisy supervision. Specifically, in CSCC, for each anchor sample from both modalities, we select its most similar neighbors in the feature space. The label similarities with these neighbors serve as soft weights to compute a weighted classification loss, enabling more reliable supervision under noisy labels. Meanwhile, BSCH leverages partial label overlap in multi-label settings to construct soft contrastive pairs, instead of relying on hard positives or hard negatives. This allows for a more nuanced assessment of semantic similarity between samples. The resulting dynamic pairing strategy enables the formation of more reliable sample pairs, effectively mitigating the impact of noisy annotations and enhancing the model's robustness. The main contributions of the proposed SCBCH framework are summarized as follows:

- We study an underexplored problem of learning cross-modal hash codes under noisy multi-label supervision, and propose Semantic-Consistent Bidirectional Contrastive Hashing (SCBCH), a framework designed to improve retrieval performance under label noise.

- A Cross-modal Semantic-Consistent Classification (CSCC) module is introduced to exploit semantic consensus from neighbors, adaptively reweighting all samples instead of filtering, thereby improving training stability and robustness in the presence of label noise.

- We propose a novel Bidirectional Soft Contrastive Hashing (BSCH), a contrastive learning paradigm tailored for multi-label scenarios that dynamically constructs soft pairs based on label overlap to capture fine-grained semantics and enhance robustness.

- Extensive experiments conducted on four widely used benchmark datasets demonstrate the effectiveness of our approach. Compared with 12 state-of-the-art baselines, SCBCH achieves consistently superior robustness under various levels of label noise.

Related Work

Noisy Label Learning

To address the challenge of learning with noisy labels, a variety of strategies have been proposed, which can be broadly categorized into two main approaches: robust learning and clean sample selection. For robust learning (Wang et al.

2025), since the conventional cross-entropy loss is sensitive to label noise, several studies have introduced noise-tolerant loss functions (Zhang and Sabuncu 2018) and regularization techniques (Liu et al. 2020; Wei et al. 2020) to enhance model robustness. For example, the symmetric cross-entropy loss (Wang et al. 2019) incorporates a noise-resistant objective that facilitates faster convergence and improved performance. However, empirical studies have shown that robust learning methods often face a trade-off between noise tolerance and predictive accuracy (Wang et al. 2022). In contrast, sample selection methods aim to identify clean instances from noisy datasets based on predefined criteria. For instance, MentorNet (Yu et al. 2019) introduces an auxiliary network to filter out noisy samples, while DivideMix (Li, Socher, and Hoi 2020) employs a Gaussian Mixture Model to dynamically divide training data into clean and noisy subsets. Nevertheless, these methods inevitably overlook hard examples with high loss that may still be informative, and they are less effective at bridging the modality gap in cross-modal tasks, as they are primarily designed for single-modality settings.

Cross-modal Hashing

Numerous cross-modal hashing methods have emerged, mainly categorized into unsupervised and supervised approaches. Unsupervised methods (Zhang et al. 2021; Zhu et al. 2022) learn a shared Hamming space by leveraging data structure without labels. For example, UCCH (Hu et al. 2023) incorporates contrastive learning with a triplet ranking loss to reduce modality gaps. However, the absence of supervision often limits their performance. In contrast, supervised methods (Wang et al. 2021; Wang and Peng 2022; Su et al. 2025a) assume that all training labels are clean and accurate, and generally optimize classification or similarity-preserving objectives, such as pairwise losses (Jiang and Li 2017) and triplet losses (Deng et al. 2018). This assumption, however, rarely holds in real-world multi-label scenarios where label noise is common. To tackle noisy supervision, robust methods like CMMQ (Yang et al. 2022) and NRCH (Wang et al. 2024) have been proposed. CMMQ introduces a proxy-based contrastive loss that prioritizes low-loss samples, while NRCH filters out unreliable samples using hard thresholds. However, such filtering risks discarding informative data, and most contrastive learning methods overlook semantic overlap in multi-label scenarios. These limitations highlight the necessity for more dynamic and noise-tolerant strategies for robust cross-modal hashing.

Method

Problem Definition

To improve clarity, we first introduce the basic definitions and notations related to cross-modal hashing with noisy labels. Suppose the multimodal training dataset consists of N examples, denoted as $\mathcal{D} = \left\{ \{\mathbf{x}_i^m\}_{m=1}^2, \mathbf{y}_i \right\}_{i=1}^N$, where $\mathbf{x}_i^m$ represents the m-th modality of the i-th sample. Specifically, $m = 1$ corresponds to the image modality and $m = 2$ de-
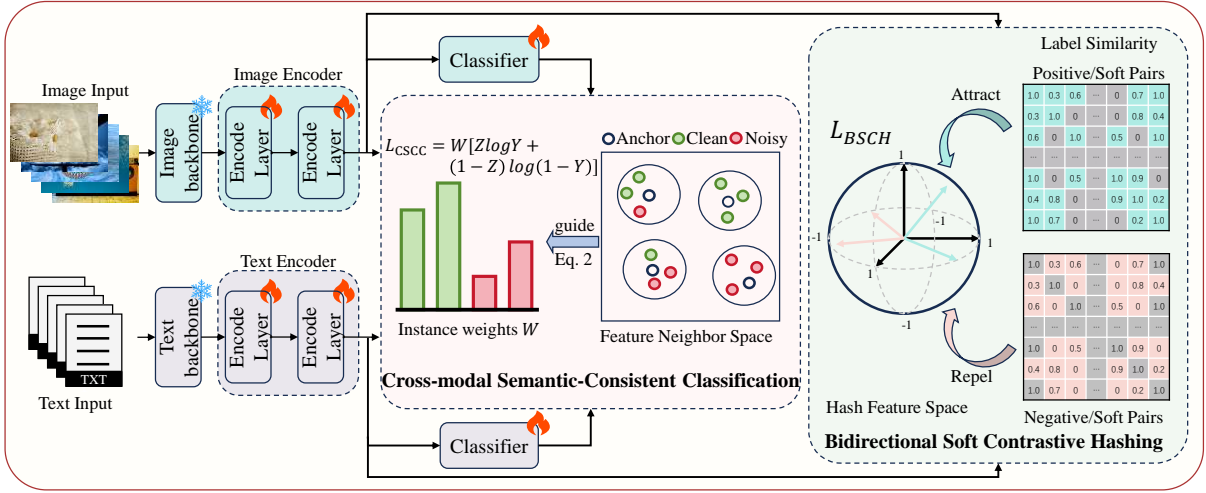
Figure 2: The SCBCH framework for cross-modal hashing under noisy multi-label supervision consists of CSCC and BSCH. CSCC enhances label reliability by leveraging cross-modal neighbor consistency to adaptively reweight samples, while BSCH employs bidirectional contrastive learning to construct reliable soft pairs based on multi-label overlap, explicitly attracting similar pairs and repelling dissimilar ones to improve robustness against noise.

notes the text modality. The label vector $\mathbf{y}_i \in \mathbb{R}^C$ is potentially noisy, and C denotes the total number of categories.

Cross-modal hashing learns modality-specific hash functions to embed heterogeneous data into a shared Hamming space. Let f^m denote the hash function for modality m, where $m \in \{1, 2\}$. During training, the continuous hash representation $\mathbf{h}_i^m \in \mathbb{R}^L$ is computed as $\mathbf{h}_i^m = \tanh(f^m(\mathbf{x}_i^m))$, where L is the length of the hash codes. The final binary hash codes $\mathbf{b}_i^m \in \{-1, +1\}^L$ are obtained by applying the sign function, i.e., $\mathbf{b}_i^m = \text{sign}(\mathbf{h}_i^m)$.

In addition, a modality-specific linear classifier followed by a sigmoid activation, denoted as g^m, is used to produce a category-wise probability distribution. The predicted probability for sample i in modality m is computed as $\mathbf{z}_i^m = g^m(\mathbf{h}_i^m)$, where $\mathbf{z}_i^m \in [0, 1]^C$.

Cross-modal Semantic-Consistent Classification

To fully utilize samples while learning from noisy sample interactions, we exploit cross-modal semantic consistency as a prior to assess label reliability: samples with semantically aligned cross-modal neighbors likely possess accurate labels. Based on this, we propose the cross-modal semantic-consistent classification (CSCC) module, which adaptively reweights samples by measuring their semantic proximity to neighbors. Firstly, we construct a soft label $\mathbf{p}_i$ for each anchor sample by aggregating the labels of its most similar cross-modal neighbors:

$$\mathbf{p}_i = \sum_{k \in \mathcal{N}_i} \mathbf{y}_k \cdot \left(\frac{1}{2} \sum_{m=1}^{2} \frac{s_{ik}^m}{\sum_{j \in \mathcal{N}_i} s_{ij}^m} \right) \quad , \qquad (1)$$

where s_{ik}^m denotes the cosine similarity between the i-th sample in modality m and its k-th neighbor, $\mathbf{y}_k$ represents the label vector of the k-th neighbor of sample i, and $\mathcal{N}_i$ denotes the set of neighbors of sample i.

We then compute a sample-specific confidence weight w_i that reflects how well the current sample's predictions agree with its soft label:

$$w_i = \gamma + (1 - \gamma) \cdot \frac{\mathbf{y}_i \cdot \mathbf{p}_i}{\|\mathbf{y}_i\| \cdot \|\mathbf{p}_i\|} \quad , \qquad (2)$$

where γ controls the base weight. This formulation measures the semantic consistency between the anchor and its soft label aggregated from cross-modal neighbors. A higher agreement yields a larger weight w_i, reflecting stronger supervision confidence and loss contribution.

Finally, we compute the weighted CSCC loss as:

$$L_{cscc} = -\frac{1}{2NC} \sum_{i=1}^{N} w_i \sum_{c=1}^{C} \sum_{m=1}^{2} [z_{ic}^m \log(y_{ic}) \\ + (1 - z_{ic}^m) \log(1 - y_{ic})] \quad , \qquad (3)$$

where $\mathbf{z}_i^m$ denotes the predicted probability vector over C classes for the i-th sample in modality m.

This CSCC module enables the model to retain all training samples while attenuating the impact of noisy or misleading labels, thereby providing robust and adaptive supervision in multi-label cross-modal settings.

Bidirectional Soft Contrastive Hashing

Although CSCC leverages clean samples and partially utilizes noisy ones, it overlooks the modality gap and the partial semantic overlaps in multi-label contrastive learning, leading to the loss of critical semantic information. To address these issues, we propose a novel bidirectional soft contrastive hashing (BSCH) module. which comprises two complementary components: an **Attraction** Component (L_{att}) that aligns semantically similar pairs, and a **Repulsion** Component (L_{rep}) that discourages alignment between mismatched samples.
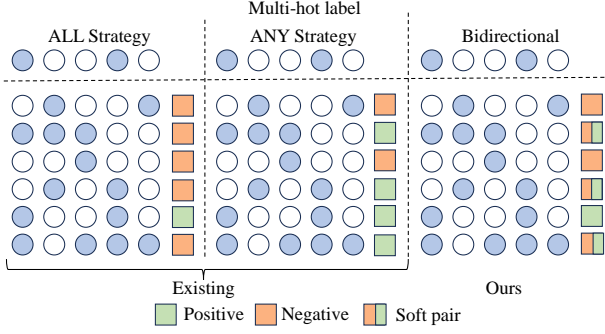
Figure 3: Illustration of three label-based contrastive pairing strategies in multi-label learning. Each row represents the multi-hot label vector of a sample with the first row corresponding to the anchor sample. Blue circle indicates a label is present with a value of one. Green boxes denote positive pairs with the anchor, orange boxes denote negative pairs, and half-green half-orange boxes represent soft pairs. Taking the last sample as an example, the ALL strategy constructs it as a negative pair, the ANY strategy regards it as a positive pair, while the Bidirectional strategy treats it as our soft pair due to partial label overlap.

Construct Contrastive Soft Pairs Existing multi-label contrastive methods (Audibert, Gauffre, and Amini 2024; Zhang and Wu 2024) typically adopt either the ALL or ANY strategy (as shown in Figure 3) to construct positive and negative pairs: **ALL** requires identical labels for positives, while **ANY** considers any shared label as positive. All other pairs are treated as negatives. However, these binary schemes overlook the semantic nuances of partially overlapping labels in multi-label data.

To better reflect these nuanced relationships, we introduce a **Bidirectional** strategy (Figure 3) that categorizes sample pairs as positive (identical labels), negative (no shared labels), or soft (partial overlap), providing a finer-grained view of semantic similarity. To quantify label similarity using the Jaccard Index. Given two multi-hot label vectors $\mathbf{y}_i, \mathbf{y}_j \in \{0,1\}^C$, the similarity score R_{ij} is defined as:

$$R_{ij} = \frac{\mathbf{y}_i^\top \mathbf{y}_j}{|\mathbf{y}_i|_1 + |\mathbf{y}_j|_1 - \mathbf{y}_i^\top \mathbf{y}_j} \quad , \qquad (4)$$

which measures the ratio of shared labels to the total activated labels across both samples.

Attract The attraction loss L_{att} encourages semantic alignment by attracting the similarity of cross-modal positive and soft pairs, and promoting consistency between modalities of the same instance.

$$L_{att} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left[\exp(\xi - S_{ij}) M_{ij} \right] - \frac{1}{n} \sum_{i=1}^{n} S_{ii} \quad , \quad (5)$$

where $S_{ij} = \mathbf{h}_i^1 \cdot \mathbf{h}_j^2$ measures cross-modal feature similarity between samples i and j, and M_{ij} is a binary mask indicating positive pairs with $R_{ij} > 0$, while S_{ii} enforces intra-instance modality alignment as a regularizer. The hyperparameter ξ controls the sharpness of positive similarity,

guiding the model to perform fine-grained alignment based on label semantics for more coherent cross-modal fusion.

Repel Although soft and negative pairs generally have low label similarity, a subset of them presents high feature similarity, making them difficult to distinguish. To handle these cases, we adopt a margin-based strategy to repel dissimilar pairs with deceptively high similarity:

$$N_{ij}^d = S_{ij}^d - \xi \cdot max(0, (S_{ii} - m) - S_{ij}^d), \qquad (6)$$

where S_{ij}^d denotes the raw cross-modal similarity in direction d, with $S_{ij}^{12} = S_{ij}$ for image-to-text and $S_{ij}^{21} = S_{ji}$ for text-to-image retrieval. The hyperparameter m serves as a margin threshold to distinguish hard negatives, while ξ is a decay term used to down-weight weakly aligned pairs.

Then the repulsion loss L_{rep} is defined as:

$$L_{rep} = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left[\sum_{d=1}^{2} \exp\left(N_{ij}^d (1 - R_{ij}) \right) \right] \quad . \quad (7)$$

This component repels high feature similarity for semantically dissimilar pairs, with adaptive weighting: high label similarity R_{ij} reduces the penalty, while high feature similarity N_{ij}^d under low label agreement increases it.

The final bidirectional soft contrastive hashing loss integrates both attraction and repulsion components, along with a regularization term:

$$L_{bsch} = L_{att} + L_{rep} + \frac{\beta}{n \cdot L} \sum_{m=1}^{2} \sum_{i=1}^{n} \sum_{l=1}^{L} (|h_{il}^m| - 1), \quad (8)$$

the last term acts as a regularizer, encouraging each element of the hash codes h_{il}^m to approach binary values, with β as a hyperparameter governing the regularization strength.

This bidirectional attraction-repulsion design allows a sample pair to be treated as soft pairs depending on label overlap. It pulls semantically aligned instances closer, while pushing mismatched ones apart. This mechanism enhances the robustness and discriminability of hash codes in noisy multi-label cross-modal scenarios.

Optimization

By combining the above loss terms, the final objective function of our model is defined as:

$$L = \begin{cases} L_c + \alpha L_{bsch} & \text{if } t \leq E_w \\ L_{cscc} + \alpha L_{bsch}, & \text{if } E_w < t < E_m \end{cases} , \quad (9)$$

where α is a hyperparameter, t is the current epoch, E_w is the warm-up epoch, and E_m is the total training epochs. For $t \leq E_w$, we disable the semantic-consistency weighting strategy and use unweighted classification loss L_c by setting $w_i = 1$ in L_{cscc}. Once $t > E_w$, we switch to L_{cscc} with confidence-based weights for adaptive supervision.

Experiments

Experimental Setup

To evaluate the performance of our method, we conducted extensive experiments across four widely-used benchmark

Dataset	Method	Year	20%				50%				80%			
			16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
WIDE	DJSRH	2019	42.7	46.5	49.7	52.4	42.7	46.5	49.7	52.4	42.7	46.5	49.7	52.4
	DGCPN	2021	59.2	61.5	63.3	64.4	59.2	61.5	63.3	64.4	59.2	61.5	63.3	64.4
	PIP	2021	56.7	58.1	59.7	60.2	56.7	58.1	59.7	60.2	56.7	58.1	59.7	60.2
	CIRH	2022	56.3	58.6	59.3	60.4	56.3	58.6	59.3	60.4	56.3	58.6	59.3	60.4
	UCCH	2023	60.3	62.7	63.7	64.4	60.3	62.7	63.7	64.4	60.3	<u>62.7</u>	63.7	64.4
	CPAH	2020	63.3	66.5	67.7	67.9	57.3	60.5	61.3	60.5	49.2	52.9	54.7	53.8
	CMMQ	2022	62.2	63.4	64.2	64.7	59.7	60.7	61.5	61.9	50.9	50.6	51.7	52.4
	MIAN	2023	60.3	61.9	63.2	63.5	54.2	56.5	57.8	58.0	46.8	48.2	49.6	48.6
	LtCMH	2023	59.4	60.5	62.7	63.5	49.3	49.8	50.7	51.7	44.6	43.8	45.6	45.8
	NRCH	2024	<u>65.3</u>	<u>67.6</u>	<u>69.1</u>	<u>68.8</u>	63.1	65.1	67.0	68.1	<u>61.0</u>	62.3	<u>63.9</u>	<u>64.7</u>
	DHRL	2024	57.6	61.6	61.6	64.2	52.8	54.3	56.7	61.6	32.0	40.4	50.8	56.0
	RSHNL	2025	63.3	65.3	66.5	67.5	<u>63.3</u>	<u>65.3</u>	66.0	67.2	57.8	59.9	62.6	64.1
	Our SCBCH		**66.7**	**68.2**	**69.6**	**69.7**	**63.9**	**66.4**	**67.5**	**68.1**	**62.0**	**63.9**	**65.0**	**66.0**
COCO	DJSRH	2019	47.8	50.6	54.4	57.1	47.8	50.6	54.4	57.1	47.8	50.6	54.4	57.1
	DGCPN	2021	60.5	63.2	63.1	64.4	60.5	63.2	63.1	64.4	60.5	63.2	63.1	64.4
	PIP	2021	54.2	56.2	57.7	59.4	54.2	56.2	57.7	59.4	54.2	56.2	57.7	59.4
	CIRH	2022	61.1	63.1	64.0	64.1	61.1	63.1	64.0	64.1	61.1	63.1	64.0	64.1
	UCCH	2023	55.3	61.4	61.2	62.5	55.3	61.4	61.2	62.5	55.3	61.4	61.2	62.5
	CPAH	2020	<u>64.7</u>	<u>68.2</u>	<u>68.8</u>	<u>69.2</u>	61.3	66.2	<u>68.5</u>	69.0	57.4	64.7	66.4	64.8
	CMMQ	2022	56.8	58.6	60.3	61.0	54.3	56.2	58.0	57.7	49.7	50.4	52.8	53.1
	MIAN	2023	58.7	61.4	62.3	63.1	56.9	57.6	59.0	59.8	52.0	56.0	57.0	57.5
	LtCMH	2023	55.8	59.7	62.6	65.3	53.1	56.6	57.9	56.3	51.7	52.9	55.2	58.0
	NRCH	2024	63.5	68.0	67.5	67.8	<u>65.4</u>	<u>68.2</u>	68.4	68.6	<u>63.9</u>	<u>67.1</u>	<u>68.3</u>	<u>69.0</u>
	DHRL	2024	33.8	33.4	45.4	61.4	33.8	33.4	57.3	60.7	33.8	33.4	49.2	58.1
	RSHNL	2025	58.2	61.1	63.4	64.0	59.4	61.3	64.4	64.4	59.3	61.9	63.9	64.8
	Our SCBCH		**67.2**	**68.8**	**69.7**	**70.6**	**66.8**	**68.2**	**70.0**	**70.4**	**65.2**	**67.7**	**69.7**	**69.5**

Table 1: The performance comparison in terms of average MAP scores of I2T and T2I tasks on the NUS-WIDE (NUS) and MS-COCO (COCO) datasets is presented, with highest and the second-highest scores are in **bold** and <u>underlined</u>, respectively.
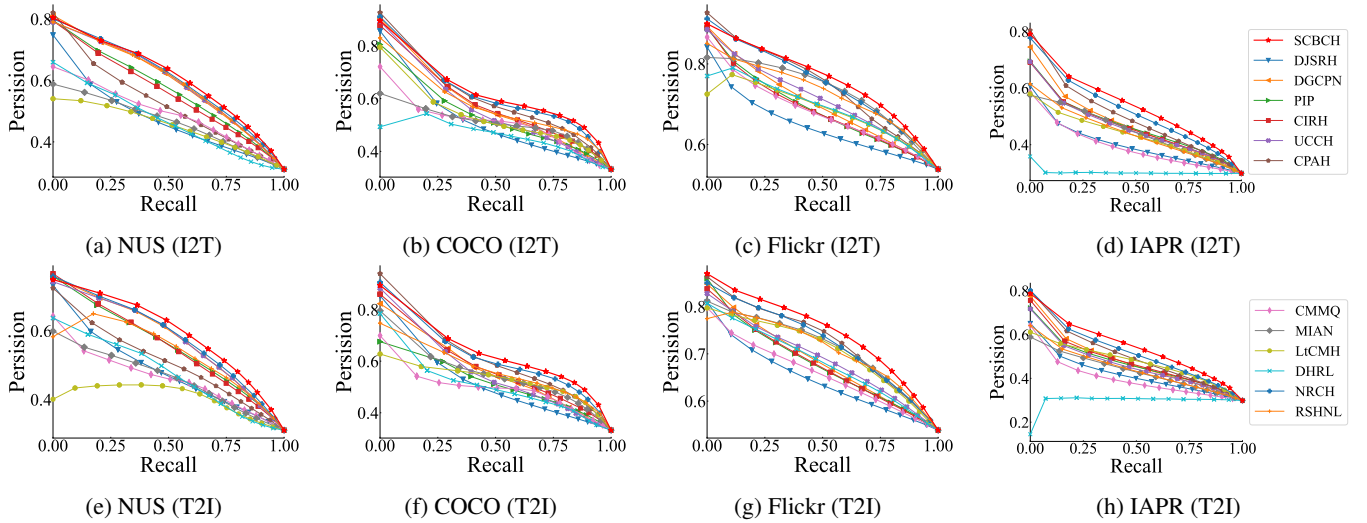

Figure 4: Precision-recall curves on four datasets under 64-bit hash codes and a 0.8 noise rate.

datasets, **NUS-WIDE** (Chua et al. 2009), **MS-COCO** (Lin et al. 2014), **MIRFlickr-25K** (Huiskes and Lew 2008), and **IAPR TC-12** (Escalante et al. 2010). And we conduct two cross-modal retrieval tasks: image-to-text (I2T) and text-to-image (T2I), where retrieval is based on Hamming distance. Similar to (Hu et al. 2023), we adopt the widely used Mean

Average Precision (MAP) as the evaluation metric. MAP is computed as the mean of Average Precision (AP) scores over all queries, effectively reflecting both the precision and ranking quality of the retrieved results. MAP scores are reported under different hash code lengths (16, 32, 64, and 128 bits). To assess robustness, we inject mixed symmetric label noise

Dataset	Method	Year	20%				50%				80%			
			16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
Flickr	DJSRH	2019	62.5	63.6	65.2	66.7	62.5	63.6	65.2	66.7	62.5	63.6	65.2	66.7
	DGCPN	2021	68.9	69.5	70.7	71.3	68.9	69.5	70.7	71.3	68.9	69.5	70.7	71.3
	PIP	2021	68.0	68.0	69.3	70.0	68.0	68.0	69.3	70.0	68.0	68.0	69.3	70.0
	CIRH	2022	68.0	68.4	69.3	69.9	68.0	68.4	69.3	69.9	68.0	68.4	69.3	69.9
	UCCH	2023	70.1	71.4	72.0	72.2	70.1	71.4	72.0	72.2	70.1	71.4	72.0	72.2
	CPAH	2020	<u>76.9</u>	<u>78.6</u>	<u>79.2</u>	<u>79.2</u>	<u>75.0</u>	<u>76.9</u>	<u>77.4</u>	76.5	<u>74.1</u>	<u>75.0</u>	75.7	74.6
	CMMQ	2022	68.0	71.2	73.1	73.8	66.1	70.0	71.4	72.1	63.3	67.5	68.2	69.2
	MIAN	2023	76.1	77.3	78.1	78.3	74.4	75.2	76.2	<u>76.6</u>	71.5	73.4	74.4	74.9
	LtCMH	2023	61.5	75.0	75.7	76.3	68.9	72.4	73.9	74.2	68.4	71.0	71.6	72.0
	NRCH	2024	75.8	76.0	76.5	76.4	74.7	75.3	76.2	76.0	73.7	74.3	<u>75.8</u>	<u>76.0</u>
	DHRL	2024	69.0	71.2	73.2	71.5	67.3	70.7	71.3	70.4	66.1	67.7	70.5	69.0
	RSHNL	2025	73.0	74.2	75.5	75.9	71.9	73.1	74.2	74.8	71.2	72.4	74.0	73.9
	Our SCBCH		**78.0**	**79.2**	**79.7**	**79.8**	**76.2**	**77.9**	**78.6**	**78.8**	**74.7**	**76.1**	**77.4**	**77.9**
IAPR	DJSRH	2019	36.7	38.8	41.4	43.3	36.7	38.8	41.4	43.3	36.7	38.8	41.4	43.3
	DGCPN	2021	46.1	47.6	48.8	49.6	46.1	47.6	48.8	49.6	46.1	47.6	48.8	49.6
	PIP	2021	45.1	46.6	47.7	48.4	45.1	46.6	47.7	48.4	45.1	46.6	47.7	48.4
	CIRH	2022	44.4	46.3	47.8	49.0	44.4	46.3	47.8	49.0	44.4	46.3	47.8	49.0
	UCCH	2023	44.7	46.1	47.1	47.6	44.7	46.1	47.1	47.6	44.7	46.1	47.1	47.6
	CPAH	2020	<u>51.0</u>	<u>53.6</u>	54.9	55.6	<u>49.9</u>	52.2	53.0	54.0	48.5	50.7	51.4	52.2
	CMMQ	2022	38.3	39.0	41.7	43.3	38.1	38.6	41.3	42.9	37.5	37.6	40.3	41.7
	MIAN	2023	46.2	48.7	50.2	51.5	44.8	46.8	48.3	49.5	43.0	45.4	46.1	47.4
	LtCMH	2023	44.4	45.8	47.6	48.6	44.2	45.6	47.2	47.8	44.0	45.1	46.6	47.3
	NRCH	2024	50.7	53.1	<u>55.0</u>	<u>55.7</u>	<u>49.9</u>	<u>52.8</u>	<u>54.4</u>	<u>55.5</u>	<u>49.2</u>	<u>52.0</u>	<u>53.8</u>	<u>55.0</u>
	DHRL	2024	30.0	30.1	30.5	43.8	30.0	30.1	30.5	43.1	30.0	30.1	30.5	42.3
	RSHNL	2025	44.4	45.5	45.7	46.0	43.8	45.8	45.9	45.8	44.2	44.5	45.3	45.5
	Our SCBCH		**53.0**	**55.6**	**58.1**	**59.2**	**52.0**	**54.5**	**57.3**	**58.3**	**50.7**	**53.8**	**56.1**	**57.2**

Table 2: The performance comparison in terms of average MAP scores of I2T and T2I tasks on the MIRFlickr-25K (Flickr) and IAPR TC-12 (IAPR) datasets is presented, with highest and the second-highest scores are in **bold** and <u>underlined</u>, respectively.

(Burgert, Ravanbakhsh, and Demir 2022) into the training data at noise rates of 20%, 50%, and 80%.

Implementation Details

In SCBCH, we use the pre-trained VGG19 (Jiang et al. 2017) on ImageNet for image features and Doc2Vec (Lau and Baldwin 2016) for text features. Three FC layers are added to the image branch and two to the text branch, each with ReLU activations except the last. Hidden layers have 8,192 units, and outputs are projected to dimension L for hash codes. Training employs Adam (Kingma and Ba 2014) with a learning rate of 1×10^{-4}, batch size 128, and 50 epochs. Additional specific parameters are set to $\gamma = 0.5$, $\alpha = 0.7$, and $\beta = 0.3$. To ensure a fair comparison, all backbone networks are kept frozen during training. Our SCBCH is implemented using the PyTorch framework and trained on a single GeForce RTX3090 24GB GPU.

Comparison with State-of-the-Arts

To validate the effectiveness of our proposed method, we conduct comprehensive comparisons against 12 state-of-the-art methods on four widely adopted cross-modal benchmark datasets. These baselines include five unsupervised approaches: DJSRH (Su, Zhong, and Zhang 2019), DGCPN (Yu et al. 2022), PIP (Zhang et al. 2021), CIRH (Zhu et al. 2022), and UCCH (Hu et al. 2023), as well as seven supervised methods: CPAH (Xie et al. 2020), CMMQ (Yang et al.

2022), MIAN (Zhang et al. 2023), LtCMH (Gao et al. 2023), NRCH (Wang et al. 2024), DHRL (Shu et al. 2024), and RSHNL (Pu et al. 2025). For each method, we track performance across training epochs and report the best MAP scores. Tables 1 and 2 report the average MAP scores for both I2T and T2I retrieval tasks. Additionally, Figure 4 plots the precision-recall curves under a 64-bit code length and 80% noise setting. Based on the results in Tables 1 and 2, we draw the following key observations:

- Longer hash codes generally boost performance via enhanced semantic encoding, but supervised methods suffer under noise, whereas unsupervised ones resist noise yet show limited gains without labels.

- As shown in Tables 1 and 2, the proposed SCBCH consistently outperforms all competing methods. For example, with a hash code length of 64 and a noise rate of 0.8, SCBCH achieves performance gains of 1.1%, 1.4%, 1.6%, and 2.3% over the second-best method on the NUS-WIDE, MS-COCO, MIRFlickr-25K, and IAPR TC-12 datasets, respectively.

- As shown in Figure 4, SCBCH consistently achieves the highest area under the precision-recall curves for both I2T and T2I tasks, demonstrating its stable and superior performance compared to state-of-the-art methods.

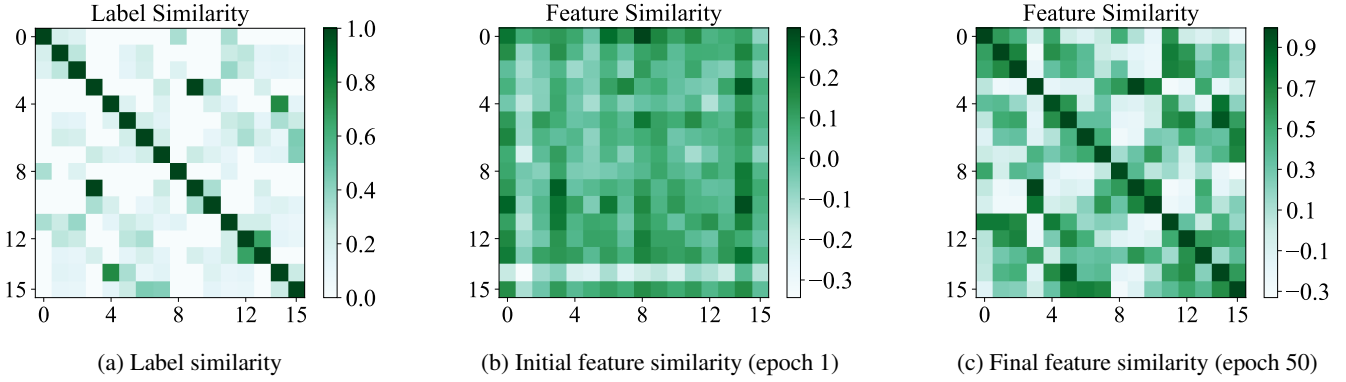|	Label Similarity	Feature Similarity	Feature Similarity
(a) Label similarity	(b) Initial feature similarity (epoch 1)	(c) Final feature similarity (epoch 50)	

Figure 5: Heatmaps of label similarities and cross-modal feature similarities among 16 randomly selected samples on the MIRFlickr-25K dataset under a 50% noise rate. Subfigure (a) shows the label similarity matrix. Subfigures (b) and (c) illustrate the evolution of feature similarities from epoch 1 to epoch 50.

Dataset	COCO			IAPR		
Noise	0.2	0.5	0.8	0.2	0.5	0.8
SCBCH-1	68.0	68.2	68.6	57.3	56.1	55.3
SCBCH-2	36.8	37.0	36.0	30.4	30.4	30.4
SCBCH-3	69.4	69.5	69.3	57.9	56.9	55.8
SCBCH-4	65.9	64.6	64.2	47.2	47.1	46.9
SCBCH	**69.7**	**70.0**	**69.7**	**58.1**	**57.3**	**56.1**

Table 3: Ablation results on the MS-COCO and IAPR TC-12 dataset with 64-bit codes under all noise rates are presented.

Ablation Study

To thoroughly investigate the contribution of each component, we compare SCBCH with its four variants: (1) removing the CSCC loss L_{cscc}, (2) removing the BSCH loss L_{bsch}, (3) removing the semantic-consistency weighting strategy, and (4) removing the attraction component loss L_{att}.

To ensure fair comparison, all variants are trained under the same configuration as SCBCH and evaluated on the test datasets. As shown in Table 3, the complete SCBCH achieves the best performance, while the variants yield inferior results. This confirms that each component plays a critical role in the overall effectiveness of SCBCH.

Model Analysis

Contrastive Similarity Analysis Figure 5 shows label and cross-modal feature similarity matrices for 16 MIRFlickr-25K samples under 50% noise with 64-bit codes. Subfigure (a) displays the label similarity heatmap, capturing partial semantic overlaps among samples. Subfigures (b) and (c) show the feature similarity matrices at epoch 1 and epoch 50, respectively. Initially disordered, they become well aligned with labels after training, demonstrating the effectiveness of our bidirectional contrastive strategy in enhancing noise-robust semantic consistency.

Weight Strategy Analysis As illustrated in Figure 6, under 50% noise rate on the MS-COCO dataset with 64 bits, the proposed weighting strategy effectively distinguishes

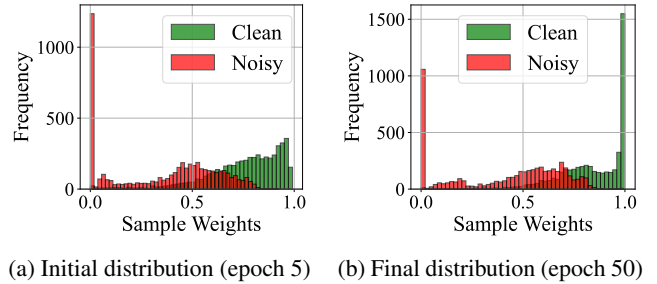| (a) Initial distribution (epoch 5)	(b) Final distribution (epoch 50)

Figure 6: Sample weight distributions for clean and noisy samples on the MS-COCO dataset with 64-bit hash codes under a 50% noise rate.

clean from noisy samples. The transition from an initially uniform distribution to a bimodal pattern concentrated at both extremes confirms the precision of confidence assignment: reliable samples are assigned high weights, while potentially mislabeled instances are suppressed, thereby enhancing robustness against label noise.

Conclusion

In this paper, we propose a novel Semantic-Consistent Bidirectional Contrastive Hashing (SCBCH) framework to address the challenge of noisy supervision in multi-label cross-modal hashing. SCBCH consists of two key modules: Cross-modal Semantic-Consistent Classification (CSCC) and Bidirectional Soft Contrastive Hashing (BSCH). Specifically, CSCC leverages cross-modal semantic consistency to generate adaptive soft supervision, retaining both clean and noisy samples for discriminative learning while preserving potentially informative data. BSCH improves robustness by dynamically constructing balanced soft pairs based on label overlap, thereby making full use of fine-grained semantic correlations and reducing the adverse impact of noisy annotations. Experiments on four benchmarks show SCBCH outperforms 12 state-of-the-art methods under noise, setting a new robust cross-modal hashing baseline.

Acknowledgements

References

Audibert, A.; Gauffre, A.; and Amini, M.-R. 2024. Multi-Label Contrastive Learning: A Comprehensive Study. *TBD*.

Burgert, T.; Ravanbakhsh, M.; and Demir, B. 2022. On the Effects of Different Types of Label Noise in Multi-Label Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–13.

Cao, Y.; Wang, X.; Wang, Q.; Yuan, Z.; Shi, Y.; and Peng, D. 2024. Graph and text multi-modal representation learning with momentum distillation on Electronic Health Records. *Knowledge-Based Systems*, 302: 112373.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. NUS-WIDE. In *Proceedings of the ACM International Conference on Image and Video Retrieval*.

Deng, C.; Chen, Z.; Liu, X.; Gao, X.; and Tao, D. 2018. Triplet-Based Deep Hashing Network for Cross-Modal Retrieval. *IEEE Transactions on Image Processing*, 3893–3903.

Escalante, H. J.; Hernández, C. A.; Gonzalez, J. A.; López-López, A.; Montes, M.; Morales, E. F.; Enrique Sucar, L.; Villaseñor, L.; and Grubinger, M. 2010. The segmented and annotated IAPR TC-12 benchmark. *Computer Vision and Image Understanding*, 419–428.

Frenay, B.; and Verleysen, M. 2014. Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 845–869.

Gao, Z.; Wang, J.; Yu, G.; Yan, Z.; Domeniconi, C.; and Zhang, J. 2023. Long-tail cross modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7642–7650.

Hu, P.; Zhu, H.; Lin, J.; Peng, D.; Zhao, Y.-P.; and Peng, X. 2023. 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*.

Jiang, L.; Zhou, Z.; Leung, T.; Li, L.; and Li, F. 2017. MentorNet: Learning Data-Driven Curriculum for Very Deep Neural Networks on Corrupted Labels. *Cornell University - arXiv,Cornell University - arXiv*.

Jiang, Q.-Y.; and Li, W.-J. 2017. Deep Cross-Modal Hashing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *arXiv: Learning,arXiv: Learning*.

Lan, Y.; Xu, S.; Su, C.; Ye, R.; Peng, D.; and Sun, Y. 2025. Multi-view Hashing Classification. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, 2122–2130. New York, NY, USA: Association for Computing Machinery. ISBN 9798400720352.

Lau, J.; and Baldwin, T. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. *Cornell University - arXiv,Cornell University - arXiv*.

Li, J.; Socher, R.; and Hoi, S. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. *International Conference on Learning Representations*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. *Microsoft COCO: Common Objects in Context*, 740–755. Springer.

Liu, H.; Ma, Y.; Yan, M.; Chen, Y.; Peng, D.; and Wang, X. 2024. DiDA: disambiguated domain alignment for cross-domain retrieval with partial labels. In *AAAI*, AAAI'24/IAAI'24/EAAI'24. AAAI Press. ISBN 978-1-57735-887-9.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-Learning Regularization Prevents Memorization of Noisy Labels. *Cornell University - arXiv,Cornell University - arXiv*.

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*.

Shu, Z.; Bai, Y.; Yong, K.; and Yu, Z. 2024. Deep cross-modal hashing with ranking learning for noisy labels. *IEEE Transactions on Big Data*.

Song, H.; Kim, M.; Park, D.; Shin, Y.; and Lee, J.-G. 2022. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 1–19.

Su, C.; Li, Z.; Lei, T.; Peng, D.; and Wang, X. 2024. MetaVG: A Meta-Learning Framework for Visual Grounding. *IEEE Signal Processing Letters*, 31: 236–240.

Su, C.; Peng, L.; Sun, Y.; Peng, D.; Peng, X.; and Wang, X. 2025a. Neighbor-aware Contrastive Disambiguation for Cross-Modal Hashing with Redundant Annotations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Su, C.; Zheng, H.; Peng, D.; and Wang, X. 2025b. DiCA: disambiguated contrastive alignment for cross-modal retrieval with partial labels. In *AAAI*, AAAI'25/IAAI'25/EAAI'25. AAAI Press. ISBN 978-1-57735-897-8.

Su, S.; Zhong, Z.; and Zhang, C. 2019. Deep Joint-Semantics Reconstructing Hashing for Large-Scale Unsupervised Cross-Modal Retrieval. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Wang, D.-B.; Wen, Y.; Pan, L.; and Zhang, M.-L. 2022. Learning from Noisy Labels with Complementary Loss

Functions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 10111–10119.

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.

Wang, Y.; Ma, X.; Chen, Z.; Luo, Y.; Yi, J.; and Bailey, J. 2019. Symmetric Cross Entropy for Robust Learning with Noisy Labels. *Cornell University - arXiv,Cornell University - arXiv*.

Wang, Y.; Ou, X.; Liang, J.; and Sun, Z. 2021. Deep semantic reconstruction hashing for similarity retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(1): 387–400.

Wang, Y.; and Peng, Y. 2022. MARS: Learning modality-agnostic representation for scalable cross-media retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4765–4777.

Wang, Y.; Wu, Y.; Dai, Z.; Tian, C.; Long, J.; and Chen, J. 2025. Noisy Correspondence Rectification via Asymmetric Similarity Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 21384–21392.

Wei, H.; Feng, L.; Chen, X.; and An, B. 2020. Combating noisy labels by agreement: A joint training method with co-regularization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, W.; Liu, Z.; Chen, Y.; Su, C.; Peng, D.; and Wang, X. 2025. Improving the Transferability of Adversarial Examples by Inverse Knowledge Distillation. *arXiv preprint arXiv:2502.17003*.

Xie, D.; Deng, C.; Li, C.; Liu, X.; and Tao, D. 2020. Multi-Task Consistency-Preserving Adversarial Hashing for Cross-Modal Retrieval. *IEEE Transactions on Image Processing*, 3626–3637.

Yang, E.; Yao, D.; Liu, T.; and Deng, C. 2022. Mutual Quantization for Cross-Modal Search with Noisy Labels. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7541–7550.

Yin, Z.; Feng, Y.; Yan, M.; Song, X.; Peng, D.; and Wang, X. 2025. RoDA: robust domain alignment for cross-domain retrieval against label noise. In *AAAI*, AAAI'25/IAAI'25/EAAI'25. AAAI Press. ISBN 978-1-57735-897-8.

Yu, J.; Zhou, H.; Zhan, Y.; and Tao, D. 2022. Deep Graph-neighbor Coherence Preserving Network for Unsupervised Cross-modal Hashing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 4626–4634.

Yu, X.; Han, B.; Yao, J.; Niu, G.; Tsang, I.; and Sugiyama, M. 2019. How does disagreement help generalization against label corruption. *International Conference on Machine Learning,International Conference on Machine Learning*.

Zhang, P.; and Wu, M. 2024. Multi-Label Supervised Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16786–16793.

Zhang, P.-F.; Li, Y.; Huang, Z.; and Yin, H. 2021. Privacy Protection in Deep Multi-modal Retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Zhang, Z.; Luo, H.; Zhu, L.; Lu, G.; and Shen, H. T. 2023. Modality-Invariant Asymmetric Networks for Cross-Modal Hashing. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 5091–5104.

Zhang, Z.; and Sabuncu, M. 2018. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *arXiv: Learning,arXiv: Learning*.

Zhu, L.; Wu, X.; Li, J.; Zhang, Z.; Guan, W.; and Shen, H. T. 2022. Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 8838–8851.

Zhu, L.; Zheng, C.; Guan, W.; Li, J.; Yang, Y.; and Shen, H. T. 2024. Multi-Modal Hashing for Efficient Multimedia Retrieval: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 239–260.