Ambiguity-Tolerant Cross-Modal Hashing with Partial Labels

Chao Su[1,2], **Yanan Li**[3], **Xu Wang**[1,4], **Yingke Chen**[5], **Huiming Zheng**[6], **Dezhong Peng**[1,7*], **Yuan Sun**[1,2,8*]
[1]College of Computer Science, Sichuan University, Chengdu, China
[2]State Key Laboratory of AI Safety, Beijing, 100086
[3]Southwest Automation Research Institute, Mianyang, China
[4]Centre for Frontier AI Research (CFAR), A*STAR, Singapore
[5]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, UK, NE1 8ST
[6]Sichuan Newstrong UHD Video Technology Co., Ltd., Chengdu, China
[7]Tianfu Jincheng Laboratory, Chengdu, 610093, China
[8]National Key Laboratory of Fundamental Algorithms and Models for Engineering Numerical Simulation, Sichuan University, Chengdu, China
suchao.ml@gmail.com, liyanan@58suo.com, wangxu.scu@gmail.com, yke.chen@gmail.com, michaelzheng@uptcsc.com, pengdz@scu.edu.cn, sunyuan_work@163.com

Abstract

Cross-modal hashing (CMH) has achieved remarkable success in large-scale cross-modal retrieval due to its low storage cost and high computational efficiency. However, most existing CMH methods rely on accurately annotated training data, which is often impractical in real-world applications due to the high cost and limited scalability of data annotation. In practice, annotators typically assign a candidate label set rather than a single precise label to each sample pair, resulting in partial labels with inherent ambiguity. Such ambiguous supervision poses significant challenges to conventional CMH methods that assume reliable and unambiguous labels. In this paper, we investigate a less-touched yet meaningful problem, i.e., cross-modal hashing with partial labels (PLCMH). PLCMH faces two major challenges: label ambiguity and modality-alignment barriers induced by misleading supervision. To address these issues, we propose a new approach named Ambiguity-Tolerant Cross-Modal Hashing (ATCH). Specifically, ATCH presents a Local Consensus Disambiguation (LCD) mechanism that resolves label ambiguity by effectively inferring stable and accurate label confidence based on local consensus within the Hamming space. Moreover, ATCH proposes a Confidence-Aware Contrastive Hashing (CACH) mechanism that derives both pseudo labels and trustworthiness scores from the label confidence vectors to learn discriminative hash codes, leading to effective modality alignment. Extensive experiments on three multimodal datasets demonstrate the superiority of ATCH.

Code — https://github.com/Rose-bud/ATCH

Introduction

With the rapid growth of multimedia data on the Internet (Su et al. 2023; Chen et al. 2023; Lu et al. 2024; Liu et al. 2024; Wen et al. 2025; Luo et al. 2025; Lan et al. 2025; Yin et al. 2025), efficient retrieval across large-scale datasets has become increasingly important (Li et al. 2024; Feng et al. 2025; Su et al. 2025b). CMH has emerged as a key technique in this field by mapping data from different modalities into compact hash codes in Hamming space, enabling low storage overhead and highly efficient retrieval.

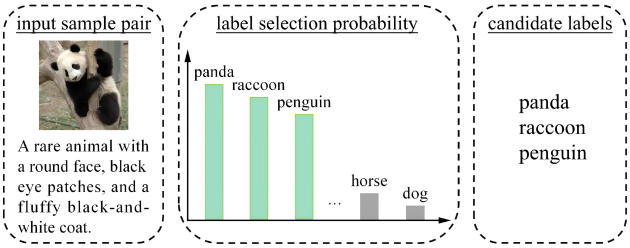


Figure 1: An input sample pair annotated with three candidate labels {panda, raccoon, penguin}. Among these, **panda** is the ground-truth, while the others are incorrect.

While unsupervised CMH approaches alleviate annotation cost by ignoring labels, they often suffer from limited semantic discriminability due to the lack of supervisory guidance. Consequently, most existing supervised CMH methods have become the dominant paradigm. However, they typically rely on an idealized assumption that the training data are annotated with precise labels. In real-world applications, obtaining such deterministic supervision is often costly and impractical. Instead, due to the complexity of the data or limited domain knowledge, annotators may struggle to assign a single definitive label. Consequently, they often provide a set of candidate labels that contains the correct label along with incorrect alternatives.

As illustrated in Fig. 1, the input sample pair can be assigned three uncertain candidate labels (panda, raccoon, penguin) if the annotators believe all of them are related to the input but are unsure which one is the true label. To address this issue, this paper studies a less-touched yet meaningful problem: cross-modal hashing with partial la-

*Co-corresponding authors.

bels (PLCMH). The main challenges of PLCMH arise from label ambiguity and modality-alignment barriers caused by wrong labels within the candidate label set.

To address these challenges, we propose a novel Ambiguity-Tolerant Cross-modal Hashing (ATCH) method for PLCMH in this paper. To be specific, ATCH includes two mechanisms, i.e., the local consensus disambiguation (LCD) mechanism and the confidence-aware contrastive hashing (CACH) mechanism. The LCD mechanism resolves label ambiguity by inferring a stable and accurate label confidence distribution based on local consensus in Hamming space. Moreover, CACH explicitly models the predictive reliability of sample pairs via dynamic trustworthiness scores, ensuring that the model prioritizes high-reliability pairs for precise modality alignment while mitigating the impact of misleading supervision from uncertain pairs. With the support of LCD, the proposed CACH reduces the modality gap and enhances the model's capability to learn a more precise label distribution. Meanwhile, the representations learned by CACH further facilitate more stable confidence estimation in LCD. This mutual reinforcement between LCD and CACH jointly boosts the overall performance of the ATCH method. The main contributions of our ATCH are summarized as follows:

- To study a less-touched yet meaningful problem: cross-modal hashing with partial labels (PLCMH), we propose a novel method named ATCH that addresses the inherent ambiguity in real-world data annotation.
- We present a local consensus disambiguation (LCD) mechanism that resolves label ambiguity by effectively inferring stable and accurate label confidence based on local consensus within the Hamming space.
- The proposed confidence-aware contrastive hashing loss explicitly models the prediction reliability of sample pairs by introducing trustworthiness scores, enabling precise modality alignment under ambiguous supervision.
- Extensive experiments on three multimodal datasets demonstrate the superiority of our proposed ATCH over other state-of-the-art baselines.

Related Work

Partial Label Learning

Partial Label Learning (PLL) (Yan and Guo 2023; He et al. 2023; Si et al. 2024; Bao, Rui, and Zhang 2024; Gong, Bisht, and Xu 2024) is a classic weakly supervised problem where each training sample is associated with multiple candidate labels, only one of which is correct, reflecting the ambiguity commonly encountered in real-world annotations.

Conventional PLL methods are generally categorized into averaging-based and identification-based streams. The averaging-based methods (Cour, Sapp, and Taskar 2011; Zhang and Yu 2015) are computationally efficient since they treat all candidate labels indiscriminately, and thus are easily overwhelmed by false-positive labels. In contrast, the identification-based methods (Liu and Dietterich 2012; Wang et al. 2020) regard the true label as a latent variable and recover it iteratively.

Benefiting from deep representation learning, recent PLL research has evolved into diverse deep disambiguation paradigms. Early works (Feng et al. 2020; Lv et al. 2020) primarily focused on progressive truth identification by iteratively updating label confidence. Subsequent studies (Wang et al. 2022; Xia et al. 2022, 2023) introduced contrastive learning and prototype alignment, leveraging the discriminative structure of the feature space to assist label disambiguation. More recently, frontier research (Tian et al. 2024) has further explored consistency regularization and multi-network collaborative selection mechanisms, enhancing robustness against label ambiguity through mutual correction.

However, most existing methods overlook the predictive reliability of the disambiguation process itself. In the context of cross-modal hashing, which is intrinsically sensitive to misleading supervision, enforcing alignment based on unreliable predictions can easily mislead the model optimization, thereby severely compromising the retrieval performance.

Cross-Modal Hashing

Cross-modal hashing (CMH) (Sun et al. 2023; Pu et al. 2025a; Su et al. 2025a; Peng et al. 2025) aims to reduce the modality discrepancy by mapping heterogeneous data from different modalities into a shared Hamming space for efficient retrieval. Existing CMH methods are commonly classified into three categories according to the type of supervision they employ: unsupervised, strongly supervised, and robust supervised CMH methods.

Unsupervised approaches (Yang et al. 2020; Yu et al. 2021; Zhu et al. 2022; Hu et al. 2022) avoid the cost of manual annotation by exploiting intrinsic data structures, but their retrieval performance is often limited due to the lack of explicit semantic supervision. Strongly supervised methods (Yang et al. 2022a; Zhang et al. 2022; Gao et al. 2023) leverage label information to effectively align different modalities, but collecting large-scale accurate annotations is time-consuming and labor-intensive in real-world scenarios. More recently, robust hashing methods (Shu et al. 2024; Wang et al. 2024; Pu et al. 2025b) have been proposed to address noisy supervision by using different noisy label learning strategies. However, most of them are built on simplified noise assumptions and do not reflect realistic annotation scenarios, where the ground-truth label is hidden among the candidate labels.

In this paper, we focus on the challenging problem of cross-modal hashing with partial labels (PLCMH), striving to construct a robust hashing framework that effectively tolerates ambiguous supervision.

Methodology

Problem Formulation

Notations. Given an image–text multimodal dataset with K categories, the training set formally consists of N paired samples indexed by j, where each pair $\{x_j^1, x_j^2\}$ corresponds to the image and text modalities, respectively. Here, n denotes the mini-batch size. In PLCMH, each sample pair is weakly annotated with a candidate label set $y_j \in \{0,1\}^K$,

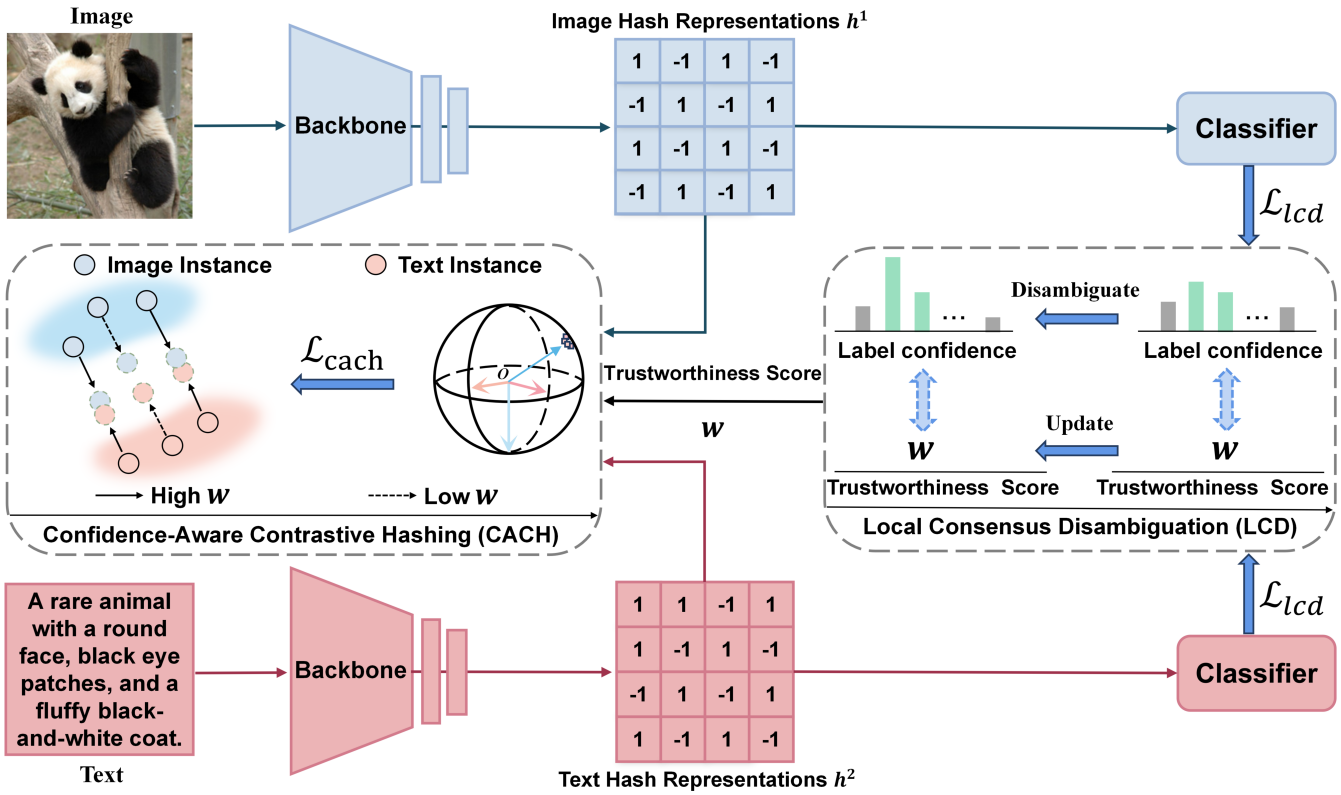


Figure 2: The framework of our proposed ATCH. LCD ($\mathcal{L}_{lcd}$) resolves label ambiguity based on the local consensus within the Hamming space. Meanwhile, CACH ($\mathcal{L}_{cach}$) enhances the model by using pseudo labels and trustworthiness scores derived from label confidence to learn discriminative hash codes, bringing positive sample pairs closer in the Hamming space and achieving precise modality alignment.

in which just one label represents the correct class. The objective of PLCMH is to map multimodal data into a unified Hamming space while preserving cross-modal semantic consistency. To this end, modality-specific hash functions $f^i(\cdot)$ are used to produce continuous embeddings $h_j^i = \tanh(f^i(x_j^i))$, where $i = 1$ and $i = 2$ represent the image and text modality, respectively. In the retrieval process, these continuous representations are discretized into binary hash codes $\boldsymbol{b}_j^i \in \{-1, 1\}^L, i \in \{1, 2\}$ according to $\boldsymbol{b}_j^i = \text{sign}(\boldsymbol{h}_j^i)$, where L denotes the length of hash codes.

Overview. In PLCMH, ambiguous annotations severely impair semantic supervision and cross-modal alignment. To address this, ATCH jointly incorporates label disambiguation and accurate modality alignment in a unified framework, with the overall objective defined as:

$$\mathcal{L} = \mathcal{L}_{lcd} + \alpha \mathcal{L}_{cach}, \tag{1}$$

where $\mathcal{L}_{lcd}$ encourages consistent label confidence estimation by exploiting local semantic consensus in the Hamming space, while $\mathcal{L}_{cach}$ guides reliable cross-modal alignment by weighting sample pairs according to their prediction trustworthiness. The balancing parameter α controls the relative influence of the two objectives. Details of these components are introduced in the following subsections.

Local Consensus Disambiguation

Recent deep partial label learning methods aim to address ambiguous supervision by estimating soft label distributions over the candidate label space. Such distributions reflect the uncertainty associated with each category and provide weak supervisory signals for model training. While this strategy has shown promising results in unimodal settings, label disambiguation becomes more challenging in cross-modal scenarios due to modality heterogeneity. In particular, relying solely on individual sample representations may result in unreliable confidence estimation, thereby hindering effective cross-modal alignment.

To alleviate this issue, we introduce a local consensus disambiguation mechanism. The underlying intuition is that samples exhibiting consistent semantic responses under ambiguous supervision are more likely to agree on their underlying labels. Leveraging such consistency allows the model to obtain more accurate confidence estimates under ambiguous supervision.

Given the hash representation h_j^i of the j-th sample from modality i, we first project it into the label space through a linear transformation:

$$v_j^i = \text{linear}(h_j^i). \tag{2}$$

For each sample pair x_j, we further incorporate locally

consistent semantic information to estimate its label confidence. Specifically, we aggregate label evidence from samples that exhibit similar semantic responses in the joint embedding space. The resulting local consensus confidence for the j-th sample pair is computed as:

$$r_j = \frac{1}{2}\,\mathrm{Normalize}\left(\sum_{c\in N_c}\sum_{i=1}^{2}(s_{jc}^{i}\cdot y_c)\cdot y_j\right), \quad (3)$$

where N_c denotes the set of neighboring pairs indexed by c that form a local consensus with the j-th pair in the Hamming space. s_{jc}^{i} measures the semantic similarity between sample j and sample c from modality i. The normalization operation $\mathrm{Normalize}$ ensures comparability across samples. The element-wise product with y_j restricts the confidence estimation to the candidate label set associated with x_j.

The obtained local consensus confidence serves as ambiguity-aware soft supervision, encouraging ATCH to learn stable and robust hash representations. Based on the estimated local consensus confidence, the disambiguation loss is defined as:

$$\mathcal{L}_{lcd} = -\sum_{j=1}^{n}\sum_{i=1}^{2}(r_j\cdot\log \mathrm{sm}(v_j^{i})\cdot y_j), \quad (4)$$

where $\mathrm{sm}(\cdot)$ denotes the softmax function, $i=1$ and $i=2$ denote the image and text modalities, respectively.

Confidence-Aware Contrastive Hashing

Although partial labels provide weak supervision, effectively aligning cross-modal representations under ambiguous annotations remains challenging in practice. Such ambiguity in supervision can introduce unreliable predictions, which may negatively affect contrastive learning and degrade representation quality.

To address this challenge, we propose a confidence-aware contrastive hashing mechanism that explicitly accounts for the reliability of the inferred supervisory signals. By incorporating instance-level confidence information into contrastive optimization, the model can better mitigate the adverse impact of ambiguous supervision. Based on the local consensus confidence, we assign each sample pair a pseudo label corresponding to the most probable category:

$$\hat{k}_j = \arg\max(r_j), \quad (5)$$

where the pseudo label $\hat{k}_j$ is used to provide supervision for contrastive learning. During training, different samples exhibit varying levels of reliability in their confidence estimation. Accurately assessing this reliability is essential for adaptive contrastive learning. While previous work (He, Yang, and Feng 2023) demonstrates that information entropy can serve as a proxy for evaluating the reliability of a sample's confidence vector, it is often sensitive to noise across all candidate labels and may fail to capture the model's relative confidence in distinguishing among the most probable classes.

We therefore introduce a trustworthiness score that evaluates instance-level reliability by measuring the separability between the most confident label prediction and its closest competitor. Specifically, we define:

$$\delta_j = r_j^{max} - r_j^{sec}, \quad (6)$$

where r_j^{max} and r_j^{sec} represent the highest and second-highest confidence values in the local consensus confidence of x_j, respectively. Thus, the trustworthiness score for the j-th sample pair is calculated as follows:

$$w_j = \epsilon + (1-\epsilon)\cdot \mathrm{Normalize}\left(\delta_j\right), \quad (7)$$

where the hyperparameter ϵ sets the lower bound of the trustworthiness score. In our experiments, ϵ is set to 0.9. The normalization operation $\mathrm{Normalize}(\cdot)$ ensures comparability across samples within each mini-batch. A larger trustworthiness score w_j indicates a more decisive confidence distribution and thus a higher degree of reliability.

We compute the similarity matrix for the image-to-text and text-to-image modalities as follows:

$$\mathcal{R}_{pq}^{12} = \left\langle \boldsymbol{h}_p^1, \boldsymbol{h}_q^2\right\rangle, \mathcal{R}_{pq}^{21} = \left\langle \boldsymbol{h}_p^2, \boldsymbol{h}_q^1\right\rangle, \quad (8)$$

where $\boldsymbol{h}^1$ and $\boldsymbol{h}^2$ represent the representations of image and text modalities, respectively. The notation $\langle\cdot,\cdot\rangle$ denotes the inner product operation, which is used to calculate the similarity between two hash representations. The superscript 12 indicates the direction from image to text, while 21 represents the direction from text to image.

Following previous work (Hu et al. 2022), we refine the similarity matrix to $\mathcal{D}_{pq}^{*}$:

$$D_{pq}^{*} = \begin{cases} \mathcal{R}_{pq}^{*}, & \mathcal{R}_{pp}^{*} - \mathcal{R}_{pq}^{*} \le \eta, \\ \mathcal{R}_{pq}^{*} - \mu, & \text{otherwise}, \end{cases} \quad (9)$$

where $*\in\{12,21\}$, η is the positive margin value and μ is used to modify the value of D^{*}. By refining the similarity matrix, Eq. (9) leverages the consistent similarity of identical samples across different modalities (as shown on the diagonal) to mitigate the impact of unreliable pairs in other parts of the similarity matrix. By considering all negative pairs within a soft margin, this strategy effectively reduces the negative influence of erroneous pairs.

To eliminate the modality gap and learn discriminative hash codes in a Hamming space, we adopt a triplet loss following (Lai et al. 2015; Deng et al. 2018) which encourages similar samples to be closer together and dissimilar samples from different modalities to be further apart. Moreover, considering the differences in the reliability of disambiguation results for different pairs, we use the trustworthiness score from Eq. (7) as the instance-level weight. We formulate the CACH loss as follows:

$$\mathcal{L}_{cach}^{*} = \frac{1}{n^3}\sum_{p=1}^{n}\sum_{j=1}^{n}\sum_{q=1}^{n} w_j \Phi_{pj}\left(1-\Phi_{pq}\right)\cdot \max\left(0, \eta + D_{pq}^{*} - D_{pj}^{*}\right), \quad (10)$$

where $*\in\{12,21\}$, $\Phi_{pq}\in\{0,1\}$ is an indicator showing whether the p-th image and the q-th text have the same pseudo label $\hat{k}$.

The trustworthiness score is used to modulate the contribution of each instance in contrastive optimization. This design enables the model to focus on reliable supervisory signals while reducing the influence of low-trustworthiness pairs, leading to more stable cross-modal alignment.

Dataset	Method	Rate	0.1				0.3				0.5			
		Ref.	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
Wiki	DJSRH	ICCV'19	15.0	17.8	19.2	20.2	15.0	17.8	19.2	20.2	15.0	17.8	19.2	20.2
	DGCPN	AAAI'21	25.6	29.0	30.7	30.4	25.6	29.0	30.7	30.4	25.6	29.0	30.7	30.4
	PIP	SIGIR'21	19.8	23.1	22.4	22.5	19.8	23.1	22.4	22.5	19.8	23.1	22.4	22.5
	CIRH	TKDE'22	28.1	28.8	28.9	26.5	28.1	28.8	28.9	26.5	28.1	28.8	28.9	26.5
	UCCH	TPAMI'23	26.2	32.2	33.1	36.7	26.2	32.2	33.1	36.7	26.2	32.2	33.1	36.7
	CMMQ	CVPR'22	34.2	36.5	37.0	36.5	16.4	17.6	18.3	19.6	11.7	11.6	11.6	11.7
	MIAN	TKDE'23	33.3	34.7	38.8	37.4	16.4	19.0	15.5	17.3	11.3	11.3	11.5	11.5
	LtCMH	AAAI'23	39.7	38.0	38.0	37.8	24.6	24.0	23.7	21.4	13.6	13.4	14.2	13.2
	DHRL	TBD'24	33.2	45.2	47.2	41.1	24.9	32.2	33.8	34.0	20.5	25.1	24.9	24.5
	NRCH	ACMMM'24	39.5	44.1	43.8	46.1	27.2	29.9	29.3	33.3	11.4	11.7	12.1	11.9
	RSHNL	AAAI'25	53.0	53.7	57.9	60.0	31.2	36.8	43.3	46.4	20.0	20.8	36.6	39.4
	ATCH	Ours	**61.0**	**64.1**	**64.5**	**65.3**	**53.0**	**56.1**	**58.6**	**58.4**	**44.2**	**50.3**	**54.5**	**51.9**
XMedia	DJSRH	ICCV'19	5.7	7.6	7.6	6.5	5.7	7.6	7.6	6.5	5.7	7.6	7.6	6.5
	DGCPN	AAAI'21	22.3	64.3	41.2	39.3	22.3	64.3	41.2	39.3	22.3	64.3	41.2	39.3
	PIP	SIGIR'21	17.8	20.6	23.8	12.3	17.8	20.6	23.8	12.3	17.8	20.6	23.8	12.3
	CIRH	TKDE'22	60.1	76.9	81.0	82.8	60.1	76.9	81.0	82.8	60.1	76.9	81.0	82.8
	UCCH	TPAMI'23	40.5	61.7	76.9	88.6	40.5	61.7	76.9	88.6	40.5	61.7	76.9	88.6
	CMMQ	CVPR'22	39.5	46.1	49.7	51.5	8.7	9.6	10.0	10.4	5.8	6.1	7.0	7.0
	MIAN	TKDE'23	24.3	28.3	50.1	53.7	5.3	5.3	5.2	5.2	5.2	5.2	5.2	5.2
	LtCMH	AAAI'23	7.4	9.6	16.2	21.8	6.2	6.0	7.1	8.3	6.4	5.6	6.0	6.7
	DHRL	TBD'24	5.2	49.8	69.8	72.7	16.1	29.9	42.3	58.3	19.6	23.0	30.1	36.9
	NRCH	ACMMM'24	89.0	89.8	92.4	92.4	34.9	61.9	62.4	66.0	5.3	5.4	5.4	5.7
	RSHNL	AAAI'25	81.8	88.8	88.5	89.0	49.2	73.0	72.2	77.3	24.4	51.6	51.2	62.6
	ATCH	Ours	**91.3**	**92.5**	**92.9**	**94.0**	**90.4**	**91.3**	**92.3**	**92.4**	**91.0**	**91.4**	**91.8**	**92.1**

Table 1: Performance comparison of average MAP scores (%) for I2T and T2I tasks under various partial rates and bit lengths on the Wikipedia (Wiki) and XMedia datasets. The highest and second highest results are shown in **bold** and in underline, respectively.

Dataset	Method	Rate	0.01				0.03				0.05			
		Ref.	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit	16bit	32bit	64bit	128bit
INRIA	DJSRH	ICCV'19	2.7	3.8	5.1	6.5	2.7	3.8	5.1	6.5	2.7	3.8	5.1	6.5
	DGCPN	AAAI'21	3.1	17.3	30.9	34.1	3.1	17.3	30.9	34.1	3.1	17.3	30.9	34.1
	PIP	SIGIR'21	2.2	5.3	13.8	1.9	2.2	5.3	13.8	1.9	2.2	5.3	13.8	1.9
	CIRH	TKDE'22	18.8	27.1	32.9	35.8	18.8	27.1	32.9	35.8	18.8	27.1	32.9	35.8
	UCCH	TPAMI'23	15.1	22.9	28.3	32.7	15.1	22.9	28.3	32.7	15.1	22.9	28.3	32.7
	CMMQ	CVPR'22	4.4	13.3	17.8	21.2	3.7	8.6	11.2	11.7	3.5	5.9	7.7	8.4
	MIAN	TKDE'23	1.1	1.1	1.3	1.3	1.1	1.3	1.3	1.3	1.2	1.2	1.2	1.2
	LtCMH	AAAI'23	1.2	1.2	1.2	1.3	1.2	1.2	1.2	1.3	1.2	1.2	1.2	1.3
	DHRL	TBD'24	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1	1.1
	NRCH	ACMMM'24	24.9	30.1	35.1	37.4	25.4	30.3	34.3	37.0	25.6	30.0	33.6	36.0
	RSHNL	AAAI'25	28.2	38.2	44.4	48.2	22.6	31.6	39.0	43.4	15.1	23.8	29.6	35.2
	ATCH	Ours	**39.3**	**48.4**	**50.9**	**53.3**	**41.5**	**48.6**	**51.5**	**53.1**	**40.8**	**47.8**	**49.8**	**50.9**

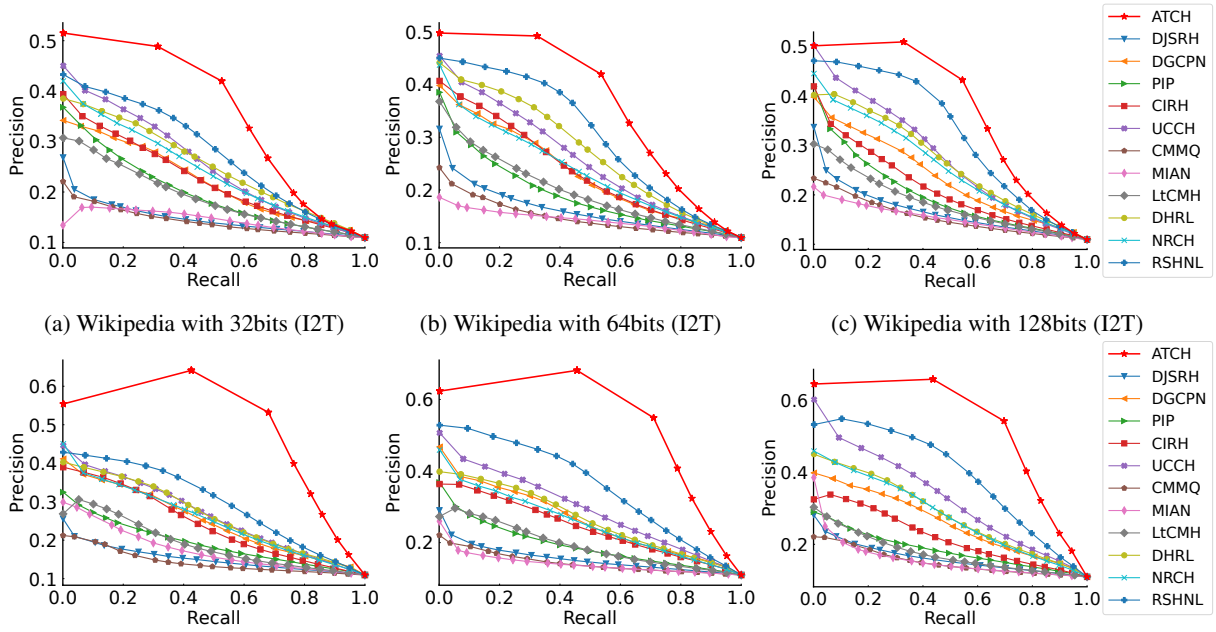
Table 2: Performance comparison of average MAP scores (%) for I2T and T2I tasks under various partial rates and bit lengths on the INRIA-Websearch (INRIA) dataset. The highest and second highest results are shown in **bold** and underline respectively.

Experiments

Dataset

To demonstrate ATCH's superiority, we carry out comprehensive evaluations across three multimodal datasets. The details of all datasets used are as follows:

- **Wikipedia** contains 2,866 image-text pairs from 10 different categories. We select 231 sample pairs as the query (test) dataset, and the remaining data serve as the retrieval



(a) Wikipedia with 32bits (I2T) (b) Wikipedia with 64bits (I2T) (c) Wikipedia with 128bits (I2T)

(d) Wikipedia with 32bits (T2I) (e) Wikipedia with 64bits (T2I) (f) Wikipedia with 128bits (T2I)

Figure 3: Precision-recall curves for different bit lengths on the Wikipedia dataset under a partial rate of 0.3.

(database) dataset. From this retrieval (database) dataset, we further extract a training subset of 2173 sample pairs.

- **XMedia** is a multimodal dataset comprising data collected from the Web across five different modalities. In this paper, we only select image and text data for CMH, including 5,000 sample pairs. From these data, we choose 500 pairs as the query (test) dataset, and the remaining pairs serve as the retrieval (database) dataset. Furthermore, 4,000 pairs from the retrieval (database) dataset are selected as the training dataset.
- **INRIA-Websearch** is a large-scale multimodal dataset consisting of 71,478 images and 71,478 text descriptions. In this work, we use a subset (Wei et al. 2017) comprising 14,698 image-text pairs from 100 different classes. From this subset, we further select 1332 image-text pairs to serve as the query (test) dataset, while the remaining data are used as the retrieval (database) dataset. Additionally, we extracted a training dataset of 4,366 image-text pairs from the retrieval dataset.

Implementation Detail

In this paper, we use the RMSprop (Tieleman 2012) optimizer with a weight decay of 1e-6. For all experiments, we set the learning rates to 1e-4, 1e-5, and 5e-4 and the maximum training epochs to 100, 50, and 50 for the Wikipedia, XMedia, and INRIA-Websearch datasets, respectively. The mini-batch size is set to 128 to ensure consistency. The values of η and μ in Eq. (9) are set to 0.2 and 1. As in prior work (Hu et al. 2021), we use the pre-trained VGG-19 (Simonyan and Zisserman 2014) model as the backbone to extract image features and the pre-trained Doc2Vec (Lau and Baldwin 2016) model for text representation on the Wikipedia dataset. For the XMedia dataset, we directly adopt the official feature representations provided with the dataset. For INRIA-Websearch, AlexNet (Krizhevsky, Sutskever, and Hinton 2012) extracts image features and LDA embeds the text descriptions. To learn shared representations across modalities, we employ three hidden layers stacked on the image backbone and two hidden layers stacked on the text backbone. Each fully connected (FC) layer, except for the final one, is followed by a Rectified Linear Unit (ReLU). The hidden layers in these FC structures consistently contain 8,192 units. The output layer maps the features into a shared space with a dimensionality of L, where L corresponds to the bit length. Additionally, the input to the disambiguation loss is projected into a K-dimensional label space through an additional fully connected layer, where K represents the number of categories. Our ATCH is implemented on the PyTorch (Paszke et al. 2019) framework and all experiments are performed on a Nvidia V100 GPU.

Experimental Setup

In this work, we perform two retrieval tasks: the image-to-text retrieval (I2T) task and the text-to-image retrieval (T2I) task. These tasks aim to retrieve relevant samples from a different modality using an image or a text as the query. We report performance using the mean average precision (MAP) computed across all retrieval results. The bit lengths are configured to 16, 32, 64, and 128. Following the previous works (Wang et al. 2022; Su et al. 2025b), we preprocess each dataset by setting distinct partial label rates: $\{0.1, 0.3, 0.5\}$ for Wikipedia and XMedia, while $\{0.01, 0.03, 0.05\}$ for INRIA-Websearch.

Comparison with State-of-the-Art Methods

To demonstrate the superiority of our ATCH, we compare the proposed ATCH with 11 state-of-the-art baselines on three multimodal datasets. These methods include unsupervised CMH methods: DJSRH (Su, Zhong, and Zhang 2019), DGCPN (Yu et al. 2021), PIP (Zhang et al. 2021), CIRH (Zhu et al. 2022), and UCCH (Hu et al. 2022) ; and supervised CMH methods: CMMQ (Yang et al. 2022b), MIAN (Zhang et al. 2022), LtCMH (Gao et al. 2023), DHRL (Shu et al. 2024), NRCH (Wang et al. 2024), and RSHNL (Pu et al. 2025b).

We present the average MAP scores for the I2T and T2I tasks with different settings in Table 1 and Table 2. In addition, we provide precision recall curves on the Wikipedia dataset under the partial rate of 0.3 in Fig. 3. From the experimental results, we can draw the following observations:

- The proposed ATCH outperforms all the compared CMH methods across all experimental settings. Thanks to both the LCD and CACH mechanisms, the proposed ATCH effectively reduces label ambiguity, endowing the model with the ability to learn discriminative hash codes from different modalities within a shared Hamming space.
- As the partial label rates increase, the performance of supervised methods declines rapidly because they cannot handle label ambiguity as effectively as ATCH. In contrast, unsupervised methods are not affected by this issue, as they operate without utilizing any label information.
- The bit length remarkably influences the performance of all methods. In most cases, as the length increases, the retrieval accuracy improves significantly. This is because longer hash codes contain richer information, enabling the model to learn more discriminative features.
- Most methods perform poorly on datasets with a large number of categories, especially on the INRIA-Websearch dataset, which contains 100 classes. In contrast, thanks to the CACH module, ATCH achieves precise cross-modal alignment, effectively reducing the modality gap and improving retrieval performance.

Ablation Study

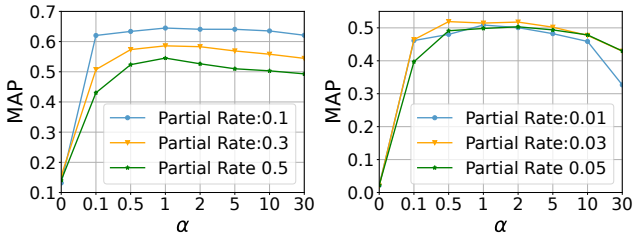
To explore the contributions of each component, we conduct ablation studies with different bit lengths on INRIA-Websearch. Specifically, we compare the full ATCH with its three variants: (1) ATCH-1 represents removing the loss $\mathcal{L}_{cach}$; (2) ATCH-2 represents removing the loss $\mathcal{L}_{lcd}$; (3) ATCH-3 represents removing the trustworthiness scores in ATCH. As shown in Table 3, one can observe that the full version of ATCH achieves the best performance while other variants show suboptimal results. This shows that all components of ATCH are essential for its effectiveness.

Parameter Analysis

To investigate the influence of the coefficient α in Eq. (1), we conduct fair experiments using 64-bit hash codes under various partial rates on the Wikipedia and INRIA-Websearch datasets, respectively. The performance of ATCH under different settings is shown in Fig. 4, where the model shows stable performance within the range of $\alpha \in [0.5, 2]$. Based on this observation, we also investigate the impact of the proposed CACH module, which is shown to further enhance the retrieval performance of ATCH.

Method	I2T			
	16bit	32bit	64bit	128bit
ATCH-1	2.4	2.4	2.4	1.5
ATCH-2	30.8	39.3	35.9	35.0
ATCH-3	39.5	46.3	49	50.7
Full ATCH	**40.2**	**47.3**	**49.3**	**50.9**
	T2I			
ATCH-1	2.3	2.2	1.5	1.9
ATCH-2	32.1	40.2	39.2	38
ATCH-3	40.6	49.1	52.8	55.2
Full ATCH	**42.7**	**49.9**	**53.6**	**55.4**

Table 3: Ablation study with a partial rate of 0.03 under different bit lengths on INRIA-Websearch. The best results are highlighted in bold.



(a) Wikipedia (b) INRIA-Websearch

Figure 4: The performance of our proposed ATCH under varying α values.

Conclusion

In this paper, we study a less-touched yet meaningful problem, i.e., cross-modal hashing with partial labels (PLCMH). To overcome this problem, we propose a novel cross-modal hashing paradigm, termed Ambiguity-Tolerant Cross-Modal Hashing (ATCH), which can learn discriminative hash codes from partial labels with ambiguity. Specifically, our ATCH presents local consensus disambiguation (LCD) to mitigate label ambiguity by providing more stable and reliable confidence estimates derived from local consensus. To overcome modality-alignment barriers caused by ambiguous supervision, we propose a confidence-aware contrastive hashing (CACH) mechanism to learn discriminative hash codes, which can bring positive sample pairs closer in the Hamming space under the guidance of both pseudo labels and trustworthiness scores. Extensive experiments on three multimodal benchmarks clearly demonstrate the effectiveness of our ATCH for cross-modal hashing with partial labels.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62372315, 62306197), Sichuan Science and Technology Planning Project (2025ZNSFSC1507, 2024ZHCG0005, 2024YFG0007, 2024ZDZX0004, 2024NSFTD0049), Central Government's Guide to Local Science and Technology Development Fund (2025ZYDF101), China Postdoctoral Science Foundation (2021TQ0223, 2022M712236), Chengdu Science and Technology Project (2023-XT00-00004-GX), Postdoctoral Joint Training Program of Sichuan University (SCDXLHPY2307), Open Funding Programs of State Key Laboratory of AI Safety.

References

Bao, W.-X.; Rui, Y.; and Zhang, M.-L. 2024. Disentangled partial label learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11007–11015.

Chen, Z.; Fu, L.; Yao, J.; Guo, W.; Plant, C.; and Wang, S. 2023. Learnable graph convolutional network and feature fusion for multi-view learning. *Information Fusion*, 95: 109–119.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from partial labels. *The Journal of Machine Learning Research*, 12: 1501–1536.

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*, 27(8): 3893–3903.

Feng, L.; Lv, J.; Han, B.; Xu, M.; Niu, G.; Geng, X.; An, B.; and Sugiyama, M. 2020. Provably consistent partial-label learning. *Advances in neural information processing systems*, 33: 10948–10960.

Feng, Y.; Qin, Y.; Peng, D.; Zhu, H.; Peng, X.; and Hu, P. 2025. Pointcloud-text matching: Benchmark dataset and baseline. *IEEE Transactions on Multimedia*.

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.

Gong, X.; Bisht, N.; and Xu, G. 2024. Does label smoothing help deep partial label learning? In *Forty-first International Conference on Machine Learning*.

He, S.; Wang, C.; Yang, G.; and Feng, L. 2023. Candidate label set pruning: A data-centric perspective for deep partial-label learning. In *The Twelfth International Conference on Learning Representations*.

He, S.; Yang, G.; and Feng, L. 2023. Candidate-aware selective disambiguation based on normalized entropy for instance-dependent partial-label learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1792–1801.

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.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lai, H.; Pan, Y.; Liu, Y.; and Yan, S. 2015. Simultaneous feature learning and hash coding with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3270–3278.

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*, 2122–2130.

Lau, J. H.; and Baldwin, T. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, 78–86.

Li, Y.; Qin, Y.; Sun, Y.; Peng, D.; Peng, X.; and Hu, P. 2024. RoMo: Robust unsupervised multimodal learning with noisy pseudo labels. *IEEE Transactions on Image Processing*.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3612–3620.

Liu, L.; and Dietterich, T. 2012. A conditional multinomial mixture model for superset label learning. *Advances in neural information processing systems*, 25.

Lu, J.; Wu, Z.; Chen, Z.; Cai, Z.; and Wang, S. 2024. Towards multi-view consistent graph diffusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 186–195.

Luo, X.; Chen, P.; Liu, C.; Jin, X.; Wen, J.; Liu, Y.; and Wang, J. 2025. Enhancing Multimodal Protein Function Prediction Through Dual-Branch Dynamic Selection with Reconstructive Pre-Training. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 7598–7606.

Lv, J.; Xu, M.; Feng, L.; Niu, G.; Geng, X.; and Sugiyama, M. 2020. Progressive identification of true labels for partial-label learning. In *International Conference on Machine Learning*, 6500–6510. PMLR.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Peng, L.; Su, C.; Wu, W.; Sun, Y.; Peng, D.; Peng, X.; and Wang, X. 2025. Semantic-Consistent Bidirectional Contrastive Hashing for Noisy Multi-Label Cross-Modal Retrieval. *arXiv preprint arXiv:2511.07780*.

Pu, R.; Qin, Y.; Song, X.; Peng, D.; Ren, Z.; and Sun, Y. 2025a. SHE: Streaming-media Hashing Retrieval. In *Forty-second International Conference on Machine Learning*.

Pu, R.; Sun, Y.; Qin, Y.; Ren, Z.; Song, X.; Zheng, H.; and Peng, D. 2025b. 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*.

Si, C.; Jiang, Z.; Wang, X.; Wang, Y.; Yang, X.; and Shen, W. 2024. Partial Label Learning with a Partner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15029–15037.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Su, C.; Li, Z.; Lei, T.; Peng, D.; and Wang, X. 2023. 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 20610–20618.

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, Y.; Ren, Z.; Hu, P.; Peng, D.; and Wang, X. 2023. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26: 824–836.

Tian, S.; Wei, H.; Wang, Y.; and Feng, L. 2024. CroSel: Cross Selection of Confident Pseudo Labels for Partial-Label Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19479–19488.

Tieleman, T. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2): 26.

Wang, H.; Qiang, Y.; Chen, C.; Liu, W.; Hu, T.; Li, Z.; and Chen, G. 2020. Online partial label learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 455–470. Springer.

Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; and Zhao, J. 2022. Pico: Contrastive label disambiguation for partial label learning. In *International Conference on Learning Representations*.

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.

Wei, Y.; Zhao, Y.; Lu, C.; Wei, S.; Liu, L.; Zhu, Z.; and Yan, S. 2017. Cross-Modal Retrieval With CNN Visual Features: A New Baseline. *IEEE Transactions on Cybernetics*, 47(2): 449–460.

Wen, J.; Long, J.; Lu, X.; Liu, C.; Fang, X.; and Xu, Y. 2025. Partial Multiview Incomplete Multilabel Learning Via Uncertainty-Driven Reliable Dynamic Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xia, S.; Lv, J.; Xu, N.; and Geng, X. 2022. Ambiguity-Induced Contrastive Learning for Instance-Dependent Partial Label Learning. In *Proceedings of International Joint Conference on Artificial Intelligence*, 3615–3621.

Xia, S.; Lv, J.; Xu, N.; Niu, G.; and Geng, X. 2023. Towards effective visual representations for partial-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15589–15598.

Yan, Y.; and Guo, Y. 2023. Mutual partial label learning with competitive label noise. In *The Eleventh International Conference on Learning Representations*.

Yang, D.; Wu, D.; Zhang, W.; Zhang, H.; Li, B.; and Wang, W. 2020. Deep semantic-alignment hashing for unsupervised cross-modal retrieval. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, 44–52.

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

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9535–9543.

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*, volume 35, 4626–4634.

Zhang, M.-L.; and Yu, F. 2015. Solving the partial label learning problem: An instance-based approach. In *Proceedings of International Joint Conference on Artificial Intelligence*, 4048–4054.

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*, 634–643.

Zhang, Z.; Luo, H.; Zhu, L.; Lu, G.; and Shen, H. T. 2022. Modality-invariant asymmetric networks for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 35(5): 5091–5104.

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.