Endogenous Recovery via Within-modality Prototypes for Incomplete Multimodal Hashing

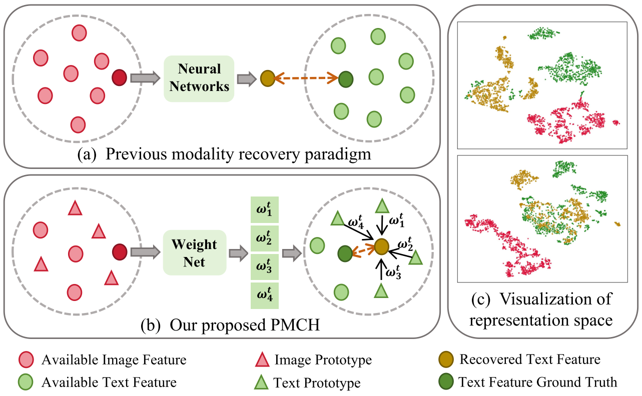
Sa Zhu[1,2,3] , **Dayan Wu**[1] * , **Chenming Wu**[4] , **Pengwen Dai**[5] and **Bo Li**[1,3]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]State Key Laboratory of Cyberspace Security Defense
[4]Baidu Research
[5]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
{zhusa, wudayan, libo}@iie.ac.cn, wuchenming@baidu.com, daipw@mail.sysu.edu.cn

Abstract

Multimodal hashing projects multimodal data into compact binary codes, enabling rapid and storage-efficient retrieval of large-scale multimedia content. In practical scenarios, the issue of missing modality frequently arises when dealing with multimodal data. Existing incomplete multimodal hashing techniques directly recover missing modalities by neural networks, resulting in a disjointed representation space between the recovered and true data. In this paper, we present a novel recovery paradigm, namely Prototype-based Modality Completion Hashing (PMCH). Instead of directly synthesizing it from available modalities, PMCH adaptively aggregates associated within-modality prototypes to recover missing modality data. Specifically, PMCH introduces an within-modality prototype learning module to optimize representative prototypes for each modality. These prototypes act as recovery anchors and reside within the same representation space as their corresponding modality data. Subsequently, PMCH adaptively aggregates the associated within-modality prototypes with coefficients derived from the modality-specific Weight-Net. By utilizing prototypes from the same modality, the semantic disparity between the reconstructed and authentic data can be substantially diminished. Extensive experiments on three widely used benchmark datasets demonstrate that PMCH can effectively recover the missing modality, and attain state-of-the-art performance in both complete and incomplete multimodal retrieval scenarios. Code is available at https://github.com/Sasa77777779/PMCH.git.

1 Introduction

Hashing [Wu *et al.*, 2019; Zhang12 *et al.*, 2019; Zhang *et al.*, 2020; Zhang *et al.*, 2021; Zhang *et al.*, 2022; Wu *et al.*, 2022a; Wu *et al.*, 2023; Wu *et al.*, 2024] aims to transform high-dimensional data into compact low-dimensional hash codes, thereby reducing retrieval complexity. As the scale of multimodal data rapidly expands, multimodal hashing [Shen *et al.*, 2018; Lu *et al.*, 2019b; Zheng *et al.*, 2020; An *et al.*, 2022] has garnered significant attention for its remarkable capability to enable large-scale multimodal retrieval with low storage costs and high retrieval efficiency. Unlike cross-modal hashing [Shen *et al.*, 2020; Kang *et al.*, 2023; Chen *et al.*, 2024; Liu *et al.*, 2024; Sun *et al.*, 2024] which focuses on learning a shared space for different modalities to enable cross-modal search, multimodal hashing aims to create a complementary space by fusing different modalities for concise retrieval. However, multimodal hashing often assumes complete modalities for each sample, which is often unavailable in real-world scenarios. For example, in social networks, the privacy or security constraints on uploaded images and texts make it challenging to access multimodal data.



(a) Previous modality recovery paradigm

(b) Our proposed PMCH

(c) Visualization of representation space

Figure 1: Two paradigms for missing modality recovery. (a) Typical paradigm exploits neural networks to generate missing data from the other (available) modality. (b) Our method aggregates within-modality prototypes for recovery. (c) T-SNE visualization of recovered and ground-truth modality representations between NCH [Tan *et al.*, 2023] (upper) and our PMCH (down). Clearly, ours exhibits closer representation spaces (overlapping of browns and greens).

The primary challenge of this incomplete multimodal hashing is *how to recover the miss modality data*. As shown in

[*] Corresponding author

Fig. 1(a), previous method like NCH [Tan *et al.*, 2023] directly recovers the missing modality features from available ones of another modality, which usually introduces a optimized neural network in a generative manner. However, this inevitably leads to the distribution gap between the generated data and the original data of miss modality, which is also verified by the empirical results of T-SNE visualizations in Fig. 1(c). Clearly, NCH shows distinct distribution gap between the representations of the recovered text features (depicted as brown points) and the ground-truth text features (depicted as green points).

To address the aforementioned challenges, we design a novel modality recovery paradigm for multimodal hashing, named Prototype-based Modality Completion Hashing (PMCH). PMCH performs data recovery by adaptively aggregating learned within-modality prototypes. Specifically, PMCH consists of two key components: Within-modality Prototype Learning Module (PLM), and Weight-Net based Prototype Aggregating Module (PAM). The PLM first designs a set of learnable prototypes for each modality, then utilizes a relation-driven prototype loss to transfer the modality- and category-specific information from available features. The PAM utilizes a modality-specific Weight-Net to dynamically learn the coefficients from the available modality data, enabling adaptive aggregation of missing modality prototypes. Since the within-modality prototypes reside in the same representation space as their corresponding features, the domain gap issue empirically observed in previous methods can be alleviated. Employing prototypes within the same modality for recovering missing data significantly minimizes the semantic discrepancy between the reconstructed and original data. The main contributions of this paper are summarized as follows:

- We propose a novel incomplete multimodal hashing framework called PMCH, that completes missing modalities by adaptively aggregating learned within-modality prototypes. PMCH can effectively mitigate the domain discrepancy between the recovered data and the authentic one.
- We meticulously devise a PLM module and a PAM module for modality and category-specific prototype learning and aggregation, respectively. The two modules can work seamlessly to directly reconstruct missing modality data with within-modality prototypes, thereby enhancing the authenticity of the recovered data.
- Extensive experiments conducted on three widely used benchmark datasets demonstrate that our PMCH achieves state-of-the-art performance in both complete and incomplete multimodal retrieval scenarios.

2 Related Work

The multimodal hashing methods aim to learn unified and semantic-rich binary representations by fusing different modalities [Liu *et al.*, 2012; Song *et al.*, 2013; Shen *et al.*, 2015; Liu *et al.*, 2015; Wang *et al.*, 2015; Yang *et al.*, 2017; Liu *et al.*, 2018; Lu *et al.*, 2019b; Lu *et al.*, 2020; Tan *et al.*, 2022; Zheng *et al.*, 2022; Zhu *et al.*, 2023; Zhu *et al.*, 2024]. Until now, numerous attempts have been made to handle incomplete multimodal hashing retrieval scenarios. Depending on whether to restore missing modalities, existing solutions can be broadly categorized into non-recovery and recovery methods [Wang *et al.*, 2023]. Non-recovery methods primarily aim to enhance the fusion model's robustness against missing modalities. For instance, Flexible Online Multimodal Hashing (FOMH) [Lu *et al.*, 2019a] and Flexible Graph Convolutional Multimodal Hashing (FGCMH) [Lu *et al.*, 2021] employ a self-weighted strategy to seamlessly fuse heterogeneous multimodal data. Supervised Adaptive Partial Multi-View Hashing (SAPMH) [Zheng *et al.*, 2020] employs multimodal matrix factorization to learn shared and view-specific hash codes for complete and incomplete modalities respectively. While these non-recovery methods exhibit less sensitivity to modality missing compared to traditional complete multimodal hashing methods, there still exists a persistent fusion gap between complete-modality and partial-modality data. Recovery methods focus on explicitly estimating and reconstructing the missing modality data from the available modalities. Graph Convolutional Incomplete Multi-modal Hashing (GCIMH) [Shen *et al.*, 2023] develops Graph Convolutional Autoencoder to complete partial-modality data with effective exploitation of its semantic structure. Neighbor-aware Completion Hashing (NCH) [Tan *et al.*, 2023] designs a cross-modal generator implemented by Multi-layer perceptron (MLP) to directly generate missing modalities from available ones, and incorporates a neighbor-aware completion learning module to guide the learning of the recovery procedure.

Current recovery methods effectively bridge the semantic gap between complete and partial modality data, but overlook the domain gap between modalities. This results in recovered features occupying a distinct representation space from ground truth. NCH's neighbor-aware completion learning module generates domain-consistent representations for incomplete training data by randomly selecting complete training samples as anchors and aggregating them based on the similarity between the available modalities of the samples and anchors. However, it faces two issues: 1) Random anchors may miss categories, leading to recovered features deviating from ground truth [Tan *et al.*, 2023]. 2) It overlooks semantic inconsistencies across modalities [Li *et al.*, 2024] by directly transferring similarity relationships from the available modality to the missing modality. In this paper, we propose to aggregate within-modality prototypes for modality recovery. On the one hand, our learned within-modality prototypes ensure the recovered features align with authentic ones in the same space and integrates valuable information across all categories. On the other hand, our proposed modality specific Weight-Net mitigates the semantic inconsistency problem by adaptively learning the coefficients for aggregation from the available modality of data.

3 Methodology

3.1 Problem Formulation

Similar to previous multimodal hashing methods [Tan *et al.*, 2023; Shen *et al.*, 2023], we primarily focus on two modal-

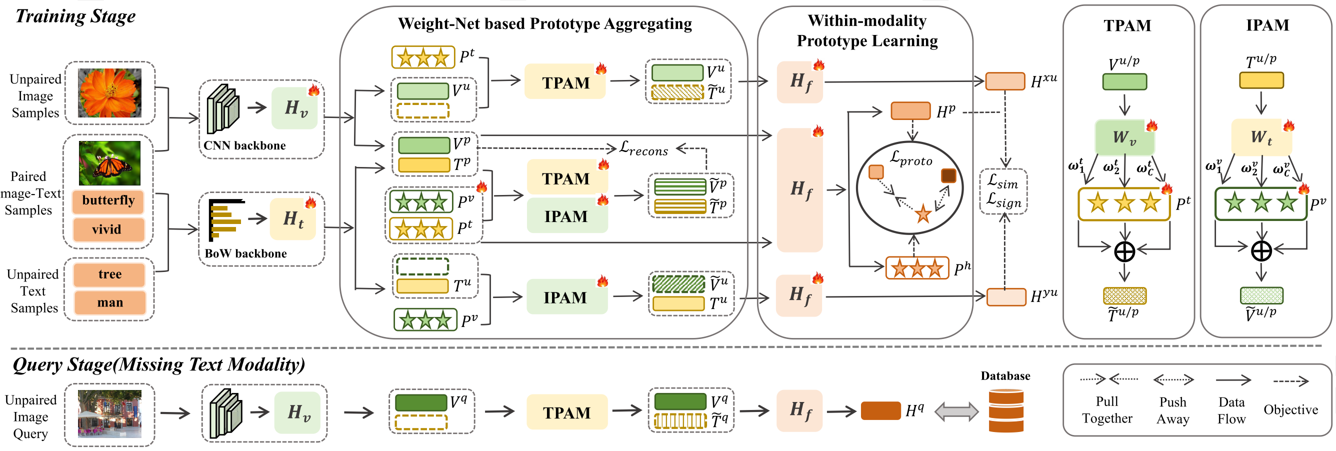


Figure 2: Illustration of our proposed Prototype-based Modality Completion Hashing (PMCH) framework, which primarily consists of two components: the Within-modality Prototype Learning Module (PLM) and the Weight-Net based Prototype Aggregating Module (PAM). PLM learns category-specific prototypes for each modality. PAM performs modality recovery by adaptively aggregating missing modality prototypes. Ultimately, the available and recovered features are fused for multimodal hashing retrieval.

ities: image and text. Suppose that our training dataset contains incomplete image and text multimodal data, their embeddings extracted by VGGNet [Simonyan and Zisserman, 2014] and Bag-of-Words (BoW) are denoted respectively as $X = [X^p, X^u] \subseteq \mathbb{R}^{n_x \times d_x}$ and $Y = [Y^p, Y^u] \subseteq \mathbb{R}^{n_y \times d_y}$, where n_p samples have complete modalities, i.e., $X^p \subseteq \mathbb{R}^{n_p \times d_x}$ and $Y^p \subseteq \mathbb{R}^{n_p \times d_y}$, n_{xu} samples have only image modality, i.e., $X^u \subseteq \mathbb{R}^{n_{xu} \times d_x}$ and n_{yu} samples have only text modality, i.e., $Y^u \subseteq \mathbb{R}^{n_{yu} \times d_y}$. d_x and d_y are dimensions of image and text feature, and we have $n_x = n_p + n_{xu}$, $n_y = n_p + n_{yu}$. Additionally, the multi-label annotation is represented as $L = [L^p, L^{xu}, L^{yu}] \subseteq \{0,1\}^{N \times C}$, where $N = n_p + n_{xu} + n_{yu}$, C denotes the total number of categories. We utilize the Frobenius norm denoted as $|| \cdot ||_F$ for various computations. The objective of our proposed PMCH framework is to learn compact and fused H-bit hash codes $b_i \in \{-1,1\}^{1 \times H}$ for both complete-modality and incomplete-modality data, as shown in Fig. 2.

3.2 Within-modality Prototype Learning

This module learns within-modality prototypes by correlating multimodal data, involving feature projection, fusion, and relation-driven prototype learning.

Multimodal feature projection and fusion. We first map the extracted visual embedding X and textual embedding Y to the feature representation that shares the same dimension through visual encoders H_v: $V = \mathcal{H}_v(X; \theta_{\mathcal{H}_v})$ and textual encoders H_t: $T = \mathcal{H}_t(Y; \theta_{\mathcal{H}_t})$, where $V = [V^p, V^u] \subseteq \mathbb{R}^{n_x \times K}$ denotes the visual feature, $T = [T^p, T^u] \subseteq \mathbb{R}^{n_y \times K}$ represents the textual feature. K is the feature dimension. Then we fuse the visual and textual feature of complete training data to generate the complete fusion hash code as follows:

$$H^p = \mathcal{H}_f(V^p + T^p; \theta_{\mathcal{H}_f}), \tag{1}$$

where $\mathcal{H}_f$ is the common hash function for hash projection, $H^p \subseteq \mathbb{R}^{n_p \times H}$ represents the fusion hash code of complete training data, H is the length of hash code.

Relation-driven prototype learning. Firstly, we devise a set of learnable prototypes tailored for each modality, denoted as $P^v = \{p_c^v\}_{c=1}^C \subseteq \mathbb{R}^{C \times K}$ for the visual modality and $P^t = \{p_c^t\}_{c=1}^C \subseteq \mathbb{R}^{C \times K}$ for the textual modality. Here, $p_c^v \in \mathbb{R}^{1 \times K}$ and $p_c^t \in \mathbb{R}^{1 \times K}$ represent learnable vectors that correspond to each category for their respective modalities, C denotes the class number. The dimension of the prototypes matches the dimension of their corresponding features, ensuring compatibility and efficient interactions during the learning process. Subsequently, we integrate prototypes from different modalities but belonging to the same category to create fusion prototypes. These fusion prototypes are then projected from their distinct representation space into a common hash space in a manner analogous to the projection of fusion features. Given learnable within-modality prototypes P^v and P^t, the fusion hash prototypes are obtained as follows:

$$P^h = \mathcal{H}_f(P^v + P^t; \theta_{\mathcal{H}_f}), \tag{2}$$

where $P^h \subseteq \mathbb{R}^{n_p \times H}$ represents the fusion hash prototype.

We further propose a relation-driven prototype loss to refine the hash prototypes. This loss function serves to pull relevant hash codes toward their corresponding prototypes while pushing away irrelevant hash codes, which is calculated in the following manner:

$$\begin{aligned} \mathcal{L}_{proto} = &- \frac{\sum_{i=1}^{n_p} \sum_{c=1}^{C} I(L_i^p = 1) cos(H_i^p, P_c^h)}{\sum_{i=1}^{n_p} I(L_i^p = 1)} \\ &+ \frac{\sum_{i=1}^{n_p} \sum_{c=1}^{C} I(L_i^p = 0) cos(H_i^p, P_c^h)}{\sum_{i=1}^{n_p} I(L_i^p = 0)}, \end{aligned} \tag{3}$$

where $I(\cdot)$ is an indicator function. The relation-driven prototype loss facilitates the effective transfer of semantic information from fusion hash features to fusion hash prototypes. As these fusion prototypes are constructed through the integration of within-modality prototypes, category information is subsequently propagated to the within-modality prototypes during the backpropagation process.

3.3 Weight-Net based Prototype Aggregating

This module focuses on aggregating within-modality prototypes for modality recovery, encompassing TPAM for text and IPAM for image prototype aggregation.

Text Prototype Aggregating Module (TPAM). This module completes patrial samples containing only image modality by aggregating text prototypes. The coefficients required for this aggregation process are dynamically generated by the image Weight-Net, leveraging the accessible image features. During the training stage, this process can be expressed as follows:

$$\{\omega_c^t\}_{c=1}^{C} = W_v(V_i^u; \theta_{w_v}), \tag{4}$$

where V_i^u denotes the incomplete training data with only image modality, $\omega_c^t \in \mathbb{R}^{1\times C}$ represents the generated coefficient corresponding to each text prototype. W_v refers to the image Weight-Net parameterized by θ_{w_v}. In our implementation, we adopt a MLP model with two hidden layers as W_v.

Once we have obtained the coefficients, we then utilize them to combine the text prototypes for feature recovery. The feature of the text modality can be reconstructed as follows:

$$\tilde{T}_i^u = \sum_{c=1}^{C} \omega_c^t \cdot p_c^t, \tag{5}$$

where $p_c^t \in \mathbb{R}^{1\times K}$ represents the learnable text prototypes, $\tilde{T}_i^u$ denotes the recovered text feature.

Image Prototype Aggregating Module (IPAM). This module is designed to handle the missing of image modality. Similarly, the text Weight-Net W_t is initially employed to generate the coefficients ω_c^v. These coefficients are then utilized to aggregate the image prototypes, thereby reconstructing the image feature, denoted as $\tilde{V}_i^u$

Modality reconstruction loss. We further propose a modality reconstruction loss to facilitate the learning of the modality specific Weight-Net and within-modality prototypes using the complete training data $[V^p, T^p]$, formulated as follows:

$$\mathcal{L}_{recons} = \|\tilde{V}^p - V^p\|_F^2 + \|\tilde{T}^p - T^p\|_F^2, \tag{6}$$

where V^p and T^p are the authentic visual and textual feature, $\tilde{V}^p$ and $\tilde{T}^p$ are the recovered visual and textual feature.

As the modality reconstruction loss aims to minimize the distance between the true features and those recovered through an adaptive aggregation of their respective prototypes, it can also implicitly ensure that the within-modality prototypes and their corresponding features reside in the same representation space.

3.4 Multimodal Hashing Learning

After restoring the missing modality features, we then integrate them with their associated available modality features to generate the hash code as follows: $H^{yu} = \mathcal{H}_m(\tilde{V}^u + T^u; \theta_{\mathcal{H}_m})$ and $H^{xu} = \mathcal{H}_m(V^u + \tilde{T}^u; \theta_{\mathcal{H}_m})$. Then the hash codes for the entire training set can be denoted as $H = [H^p, H^{xu}, H^{yu}] \subseteq \mathbb{R}^{N\times H}$, where H^p represents the hash codes of paired training data computed using Eq 1. Subsequently, we employ the pairwise similarity loss to learn discriminate hash codes as follows:

$$\mathcal{L}_{sim} = \sum_{i=1}^{N}\sum_{j=1}^{N} \|cos(H_i, H_j) - S_{ij}\|_F^2, \tag{7}$$

where S_{ij} denotes the semantic similarity between the i-th sample and j-th sample. S_{ij} is constructed as follows:

$$S_{ij} = \frac{2}{1 + e^{-L_i L_j^T}} - 1. \tag{8}$$

To minimize the quantization errors caused by $sign(\cdot)$ operator, the discrete hash code $sign(H_i)$ is used to guide the learning process of the continuous hash code H_i as follows:

$$\mathcal{L}_{sign} = \sum_{i=1}^{N} \|H_i - sign(H_i)\|_F^2. \tag{9}$$

Finally, the overall objective function of our proposed PMCH architecture can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_{sim} + \alpha\mathcal{L}_{recons} + \beta\mathcal{L}_{proto} + \gamma\mathcal{L}_{sign}, \tag{10}$$

where α, β, and γ serve as hyper-parameters.

3.5 Out-of-Sample Extension

Our proposed PMCH architecture shown in Fig 2 can effectively handle both queries with complete modalities and those with missing modalities. Specifically, given a query containing both visual modality v^q and textual modality t^q, its corresponding hash code can be generated by the trained multimodal hashing network $\mathcal{H}_f$ as follows:

$$b^q = sign(\mathcal{H}_f(v^q + t^q; \theta_{\mathcal{H}_f})). \tag{11}$$

When confronted with a query containing only the visual modality, i.e., $(v^q, *)$ or solely the textual modality, i.e., $(*, t^q)$, PMCH employs modality specific Weight-Net to generate the coefficients for prototype aggregation from the available modality :

$$\{\omega_c^m\}_{c=1}^{C} = W_a(a^q; \theta_{w_a}), \tag{12}$$

where m represents the missing modality, a represents the available modality.

Subsequently, the missing modality is reconstructed by aggregating its corresponding prototypes with the generated coefficients, as outlined below:

$$\tilde{m}^q = \sum_{c=1}^{C} \omega_c^m \cdot p_c^m. \tag{13}$$

Finally, we integrate the recovered missing modality feature with the available modality feature to produce the final hash code, as detailed below:

$$b^q = sign(\mathcal{H}_f(a^q + \tilde{m}^q; \theta_{\mathcal{H}_f})). \tag{14}$$

Task	Method	MIR Flickr				NUS-WIDE				MS COCO			
		16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
T_{only}	DCMVH(TIP'20)	0.6516	0.6702	0.6893	0.7001	0.5309	0.5586	0.5770	0.5923	0.5038	0.5223	0.5279	0.5341
	OASIS(AAAI'22)	0.7070	0.7316	0.7434	0.7522	0.5681	0.5852	0.6097	0.6253	0.5299	0.5457	0.5501	0.5527
	BSTH(SIGIR'22)	0.7125	0.7477	0.7516	0.7578	0.5798	0.5935	0.6241	0.6372	0.5343	0.5472	0.5544	0.5595
	SFISMH(TMM'24)	0.7290	0.7337	0.7401	0.7519	0.6159	0.6406	0.6486	0.6558	0.5326	0.5552	0.5590	0.5668
	STBMH(TKDE'24)	0.7334	0.7443	0.7490	0.7537	0.5948	0.6285	0.6419	0.6511	0.5474	0.5705	0.5878	0.6035
	FOMH(MM'19)	0.6209	0.6343	0.6432	0.6692	0.5652	0.5952	0.6101	0.6186	0.5111	0.5247	0.5261	0.5335
	FGCMH(MM'21)	0.6715	0.6921	0.7084	0.7189	0.5865	0.6152	0.6451	0.6497	0.5674	0.5776	0.6096	0.6329
	SAPMH(TMM'21)	0.7249	0.7353	0.7462	0.7554	0.6251	0.6524	0.6777	0.6907	0.5692	0.5862	0.6167	0.6419
	GCIMH(MM'23)	0.7553	0.7725	0.7781	0.7828	0.6535	0.6721	0.6940	0.7056	0.5782	0.5989	0.6266	0.6516
	NCH(TMM'23)	0.7590	0.7743	0.7800	0.7847	0.6819	0.6995	0.7176	0.7310	0.5907	0.6116	0.6370	0.6639
	PMCH(ours)	**0.7854**	**0.7985**	**0.8040**	**0.8061**	**0.7019**	**0.7181**	**0.7303**	**0.7378**	**0.6054**	**0.6454**	**0.6700**	**0.6844**
I_{only}	DCMVH(TIP'20)	0.7671	0.7744	0.7902	0.8010	0.6647	0.6683	0.6751	0.6889	0.4556	0.4611	0.4652	0.4715
	OASIS(AAAI'22)	0.8017	0.8093	0.8136	0.8272	0.6817	0.6886	0.6935	0.7019	0.4873	0.5032	0.5094	0.5130
	BSTH(SIGIR'22)	0.8055	0.8116	0.8269	0.8313	0.6985	0.7096	0.7179	0.7215	0.4993	0.5107	0.5115	0.5180
	SFISMH(TMM'24)	0.8056	0.8128	0.8255	0.8327	0.7007	0.7205	0.7381	0.7447	0.5112	0.5118	0.5145	0.5149
	STBMH(TKDE'24)	0.8027	0.8125	0.8293	0.8317	0.6852	0.7168	0.7310	0.7417	0.4970	0.5081	0.5168	0.5292
	FOMH(MM'19)	0.7502	0.7600	0.7870	0.7930	0.6348	0.6649	0.6869	0.6893	0.4448	0.4494	0.4533	0.4536
	FGCMH(MM'21)	0.7568	0.7724	0.7956	0.8101	0.6387	0.6708	0.6774	0.6875	0.4977	0.5075	0.5214	0.5372
	SAPMH(TMM'21)	0.7796	0.7931	0.8049	0.8127	0.6727	0.6944	0.7056	0.7269	0.5076	0.5175	0.5281	0.5361
	GCIMH(MM'23)	0.7920	0.8105	0.8285	0.8315	0.6993	0.7202	0.7368	0.7430	0.5213	0.5408	0.5427	0.5439
	NCH(TMM'23)	0.8072	0.8231	0.8316	0.8353	0.7149	0.7307	0.7470	0.7601	0.5210	0.5354	0.5433	0.5510
	PMCH(ours)	**0.8390**	**0.8516**	**0.8545**	**0.8582**	**0.7398**	**0.7584**	**0.7695**	**0.7707**	**0.5302**	**0.5419**	**0.5490**	**0.5554**

Table 1: mAP results under two missing conditions on query set.

4 Experiments

4.1 Experimental Settings

Datasets. In this study, we conduct experiments on three widely used benchmark datasets, i.e., MIR Flickr, NUS-WIDE, and MSCOCO. The multimodal data includes image and text modalities. The images are extracted using VGGNet, while the text modality is represented using bag-of-words vectors [Shen *et al.*, 2023; Tan *et al.*, 2023]. MIR Flickr [Huiskes *et al.*, 2010] comprises 20,015 image-text pairs annotated with one or more of 24 categories, which are crawled from the Flickr website. NUS-WIDE [Chua *et al.*, 2009] comprises 195,834 web images-text pairs associated with the 21 most prevalent concept labels. MS COCO [Lin *et al.*, 2014] comprises 82,783 training samples and 40,504 validation samples, each associated with at least one of the 80 categories. We follow Tan [Tan *et al.*, 2023] to split the three datasets into the training, validation, and test sets.

Implementation. In our proposed PMCH, the modality-specific Weight-Net comprises two fully-connected (FC) layers, with the latent feature dimension set to 2048. Additionally, the visual and textual encoders are implemented using FC layers, with latent dimensions of 2,048 and 1,024 respectively. The outputs of these encoders have a dimension of 512. Furthermore, the hash fusion model incorporates a linear projection layer that takes in an input dimension of 512 and outputs a hash code length as specified. During the training stage, we set the batch size to 512 and run the iterations for 50 epochs. For optimization, we employ the Adam optimizer [Kingma and Ba, 2014]. Empirically, we set the learning rate to 0.05 for prototype learning and 0.001 for the remaining components.

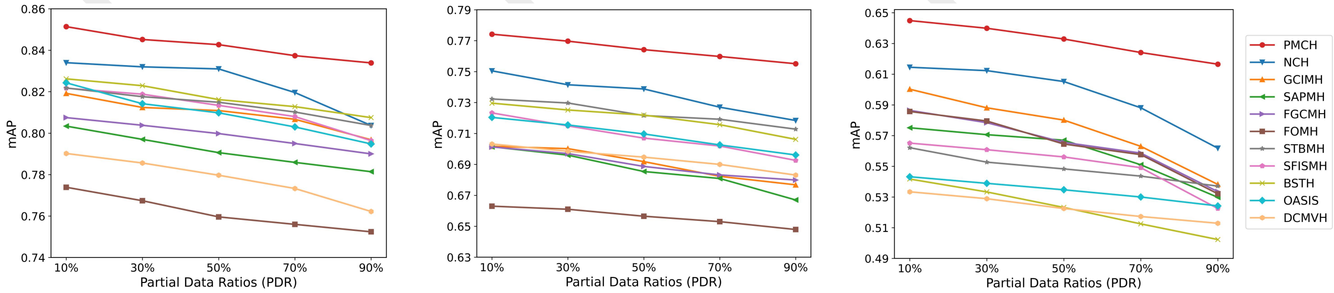
Baselines. We compare our proposed method against ten state-of-the-art (SOTA) multimodal hashing techniques, including five complete multimodal hashing methods which limited to handling complete multimodal data, i.e., DCMVH [Zhu *et al.*, 2020], OASIS [Wu *et al.*, 2022b], BSTH [Tan *et al.*, 2022], SFISMH [Zhu *et al.*, 2024], STBMH [Tu *et al.*, 2024] and five incomplete approaches which are capable of processing not only complete but also incomplete multimodal data, i.e., FOMH [Lu *et al.*, 2019a], FGCMH [Lu *et al.*, 2021], SAPMH [Zheng *et al.*, 2020], NCH [Tan *et al.*, 2023], and GCIMH [Shen *et al.*, 2023]. For the complete multimodal hashing methods, we directly model the incomplete multimodal data for training and query without recovery. We carefully reproduce these methods using their publicly available codes and adhere to the parameter settings specified in the original papers.

Evaluation Metric. Consistent with established multimodal hashing retrieval methods [Tan *et al.*, 2022; Shen *et al.*, 2023], we consider the widely-used metrics, i.e., mean Average Precision (mAP) to quantitatively assess the retrieval performance of our proposed approach. The mAP is calculated using all the samples in the database.

4.2 Comparison with State-of-the-Art Methods

We compare the proposed PMCH with state-of-the-arts on both incomplete and complete multimodal retrieval tasks.

Incomplete Query Set. This section evaluates the performance of PMCH when query data is incomplete. Following [Shen *et al.*, 2023], we consider two missing-modality conditions: missing visual modality and missing textual modality. According to Table 1, we can obtain the following observations: 1) Generally, recovery methods show better performance than non-recovery methods. This demonstrates that reconstructing the missing modality data could bridge the semantic gap with complete modality data, thereby obtaining higher performance. 2) Our PMCH outperforms all baselines



(a) MIR Flickr (b) NUS WIDE (c) MS COCO

Figure 3: mAP results with different PDR on training set.

Method	16bits	32bits	64bits	128bits
BSTH(SIGIR'22)	0.8208	0.8334	0.8467	0.8526
GCIMH(MM'23)	0.8201	0.8426	0.8455	0.8517
SFISMH(TMM'24)	0.8228	0.8389	0.8488	0.8548
STBMH(TKDE'24)	0.8256	0.8401	0.8486	0.8580
Ours	**0.8404**	**0.8528**	**0.8552**	**0.8590**

Table 2: mAP results on complete multimodal retrieval.

Task	Method	16bits	32bits	64bits	128bits
T_{only}	w/o Completion	0.7180	0.7359	0.7404	0.7505
	+GCAE w/o PMCH	0.7443	0.7585	0.7640	0.7682
	+MLP w/o PMCH	0.7440	0.7598	0.7685	0.7718
	Ours	**0.7606**	**0.7725**	**0.7782**	**0.7815**
I_{only}	w/o Completion	0.7833	0.8032	0.8163	0.8239
	+GCAE w/o PMCH	0.8061	0.8159	0.8271	0.8294
	+MLP w/o PMCH	0.8017	0.8171	0.8278	0.8285
	Ours	**0.8264**	**0.8391**	**0.8452**	**0.8459**

Table 3: Ablation study on modality completion method.

Task	Method	16bits	32bits	64bits	128bits
T_{only}	+anchors w/o PLM	0.7173	0.7328	0.7450	0.7475
	Ours	**0.7606**	**0.7725**	**0.7782**	**0.7815**
I_{only}	+anchors w/o PLM	0.7776	0.7988	0.8131	0.8202
	Ours	**0.8264**	**0.8391**	**0.8452**	**0.8459**

Table 4: Ablation study on Prototype Learning Module.

Task	Method	16bits	32bits	64bits	128bits
T_{only}	+KNN w/o PAM	0.7290	0.7524	0.7605	0.7648
	+GAT w/o PAM	0.7156	0.7309	0.7406	0.7458
	+TEs w/o PAM	0.7449	0.7621	0.7678	0.7716
	Ours	**0.7606**	**0.7725**	**0.7782**	**0.7815**
I_{only}	+KNN w/o PAM	0.8035	0.8223	0.8289	0.8319
	+GAT w/o PAM	0.8006	0.8192	0.8215	0.8241
	+TEs w/o PAM	0.8113	0.8285	0.8337	0.8373
	Ours	**0.8264**	**0.8391**	**0.8452**	**0.8459**

Table 5: Ablation study on Prototype Aggregating Module.

in both missing cases. This proves that PMCH could recover more authentic features for incomplete query data, as it alleviates the domain gap issue by adaptively aggregating within-modality prototypes.

Incomplete Training Set. This section section evaluates the performance of PMCH when training data is incomplete. Following [Tan *et al.*, 2023], we construct incomplete training data using five Partial Data Ratios (PDR). For example, "90%" indicates 90% of the data misses a modality, with visual and textual modalities each having a PDR of 45%. The mAP results of all baselines with respect to 32 bits are shown in Fig. 3. From Fig. 3, we can observe that PMCH obviously outperforms the baselines across all PDRs. In addition, as PDR increases, PMCH achieves relatively stable performance. These empirical results imply that the proposed PMCH can effectively handle the training of incomplete multimodal hashing by minimizing the semantic discrepancy between the reconstructed and original data.

Complete Multimodal Retrieval. This section evaluates the performance of PMCH on complete multimodal retrieval. We present comparison results on MIR Flickr with the top four baselines in Table 2. According to Table 2, although PMCH is not specifically designed for complete multimodal retrieval, it still shows competitive performance in this setting. The reason lies in the joint learning of hash codes and their corresponding prototypes in PLM module, which not only learns within-modality prototypes, but also effectively draws relevant data closer, thereby improving complete multimodal retrieval performance.

4.3 Ablation Study

In this section, we perform ablation studies by substituting the modality completion strategies and key components with their respective variants. The experiment is conducted on Mir Flickr with 50% PDR applied to the training set, and either image or text modality is missing in the query set.

Variants on Modality Completion Method. In this section, we design three variants to prove the superiority of the proposed modality completion paradigm, including: 1) **w/o completion**. In this variant, we remove our completion learning and directly model incomplete multimodal data. 2) **+GCAE w/o PMCH** In this variant, following GCIMH [Shen *et al.*, 2023], we use Graph Convolutional Autoencoder (GCAE) to complete the partial modality data. 3) **+MLP w/o PMCH** In this variant, following NCH [Tan *et al.*, 2023], we introduce a Multi-layer perceptron (MLP) to

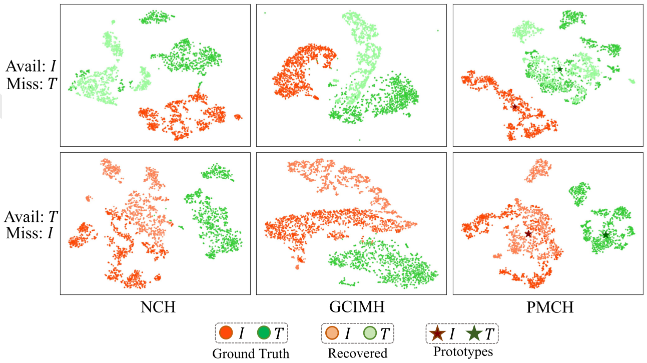


Figure 4: Visualization of the recovered feature and ground truth for NCH, GCIMH and our PMCH under missing patterns with only one modality available.

generate the missing modality from the available one. From the retrieval performance shown in Table 3, two observations can be obtained: 1) Removing total completion learning will result in incomplete multimodal semantics and thus reduce retrieval performance. 2) Compared with the generation-based modality recovery method, our proposed prototype-based completion method performs the best. The main reason is that compared with the cross-modal generation, PMCH can avoid the separated representation space with true features when recovering missing modalities.

Effectiveness of Prototype Learning Module (PLM). In our PMCH, we design PLM to learn category-specific prototypes for each modality. In order to evaluate its effectiveness, following the setting of [Tan *et al.*, 2023], we replace the learnable within-modality prototypes with complete anchors randomly selected from the training set. The number of anchors is set to 300. According to the results shown in Table 4, our PMCH performs better. The reason is that random anchors may miss categories. If all the anchors are dissimilar to the arrived incomplete samples, the recovered features will be distinct from the ground truth. While our within-modality prototypes learned by PLM integrate information across all categories, enabling the generation of appropriate features for partial samples belonging to any category.

Effectiveness of Prototype Aggregating Module (PAM). PAM employs the modality specific Weight-Net to dynamically learn the coefficients for aggregating missing modality prototypes. In order to evaluate its effectiveness, we replace it with three similarity-based aggregation methods mentioned in [Tan *et al.*, 2023], which respectively employ K-Nearest Neighbors (KNN), Graph Attention Network (GAT) and Transformer Encoders (TEs) to compute similarities between the available modality of samples and the available modality prototypes. These similarities are then utilized to aggregate missing modality prototypes. As shown in Table 5, we can find that our PAM shows better performance than the other similarity-based prototype aggregating variants, which implies that the modality specific Weight-Net could dynamically learn the coefficients from the available modality of samples, enabling adaptive aggregation of missing modality prototypes, thus recovering more authentic features.

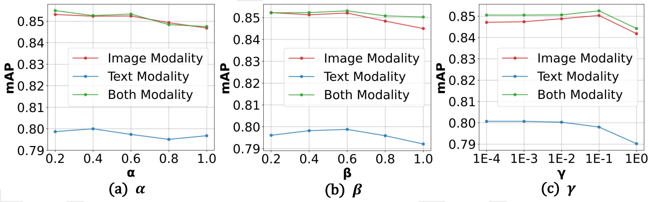


Figure 5: Parameter sensitivity analysis.

4.4 Further Analysis

Visualization. We visualize the recovered features and ground truth to qualitatively compare the generation-based recovery method NCH, GCIMH, and our proposed PMCH. To achieve this, we randomly select one label from each of the datasets, with a code length set to 16. We then project the features of the selected samples into a 2D space using t-SNE [Van der Maaten and Hinton, 2008]. Additionally, we visualize the learned within-modality prototypes for the chosen label. The results on MIR Flickr are presented in Fig. 4. By analyzing the t-SNE results, we make two key observations: 1) The prototypes learned by PMCH reside in the same representation space as their associated modality features. 2) The representation space of features recovered by PMCH is significantly closer to the ground truth compared to NCH and GCIMH. These results underscore the effectiveness of our proposed PMCH method in avoiding the domain gap during the feature recovery process.

Parameter Sensitivity Analysis. We investigate the impact of three crucial hyper-parameters in our proposed PMCH method: α, β, and γ, using a hash code length of 32-bits on the MIR Flickr dataset. Fig. 5 depicts the performance variation curves on image modality, text modality and both modalities, respectively. Overall, our method demonstrates robustness as the mAP performance remains relatively stable within a certain range of hyper-parameter variations.

5 Conclusion

In this paper, we introduce a prototype-based modality completion method called PMCH for incomplete multimodal hashing learning. Specifically, we devise an Within-modality Prototype Learning Module that learns prototypes for each modality. Subsequently, we delicately design a modality specific Weight-Net to dynamically generate coefficients from available modality, enabling adaptive aggregation of missing modality prototypes. PCMH ensures recovered features align with authentic ones in the same space and adaptively integrates valuable information across all categories for precise recovery. Extensive experiments on three widely used multimodal retrieval datasets underscore the superiority of our proposed method from various aspects in both complete and incomplete multimodal retrieval scenarios.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant 2022YFB3103500, the National Natural Science Foundation of China under Grant 62472420, the Natural Science Foundation of Guangdong, 2025A1515011224.

References

[An *et al.*, 2022] Junfeng An, Haoyang Luo, Zheng Zhang, Lei Zhu, and Guangming Lu. Cognitive multi-modal consistent hashing with flexible semantic transformation. *Information Processing & Management*, 59(1):102743, 2022.

[Chen *et al.*, 2024] Bingzhi Chen, Zhongqi Wu, Yishu Liu, Biqing Zeng, Guangming Lu, and Zheng Zhang. Enhancing cross-modal retrieval via visual-textual prompt hashing. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 623–631, 2024.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009.

[Huiskes *et al.*, 2010] Mark J Huiskes, Bart Thomee, and Michael S Lew. New trends and ideas in visual concept detection: The mir flickr retrieval evaluation initiative. In *Proceedings of the international conference on Multimedia information retrieval*, pages 527–536, 2010.

[Kang *et al.*, 2023] Xiao Kang, Xingbo Liu, Xuening Zhang, Xiushan Nie, and Yilong Yin. Online discriminative cross-modal hashing. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Li *et al.*, 2024] Tieying Li, Xiaochun Yang, Yiping Ke, Bin Wang, Yinan Liu, and Jiaxing Xu. Alleviating the inconsistency of multimodal data in cross-modal retrieval. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*, pages 4643–4656. IEEE, 2024.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2012] Xianglong Liu, Junfeng He, Di Liu, and Bo Lang. Compact kernel hashing with multiple features. In *Proceedings of the 20th ACM international conference on multimedia*, pages 881–884, 2012.

[Liu *et al.*, 2015] Li Liu, Mengyang Yu, and Ling Shao. Multiview alignment hashing for efficient image search. *IEEE Transactions on image processing*, 24(3):956–966, 2015.

[Liu *et al.*, 2018] Ruoyu Liu, Shikui Wei, Yao Zhao, Zhenfeng Zhu, and Jingdong Wang. Multiview cross-media hashing with semantic consistency. *IEEE MultiMedia*, 25(2):71–86, 2018.

[Liu *et al.*, 2024] Kaiming Liu, Yunhong Gong, Yu Cao, Zhenwen Ren, Dezhong Peng, and Yuan Sun. Dual semantic fusion hashing for multi-label cross-modal retrieval. In *International Joint Conferences on Artificial Intelligence Organization, IJCAI*, pages 4569–4577, 2024.

[Lu *et al.*, 2019a] Xu Lu, Lei Zhu, Zhiyong Cheng, Jingjing Li, Xiushan Nie, and Huaxiang Zhang. Flexible online multi-modal hashing for large-scale multimedia retrieval. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1129–1137, 2019.

[Lu *et al.*, 2019b] Xu Lu, Lei Zhu, Jingjing Li, Huaxiang Zhang, and Heng Tao Shen. Efficient supervised discrete multi-view hashing for large-scale multimedia search. *IEEE Transactions on Multimedia*, 22(8):2048–2060, 2019.

[Lu *et al.*, 2020] Xu Lu, Li Liu, Liqiang Nie, Xiaojun Chang, and Huaxiang Zhang. Semantic-driven interpretable deep multi-modal hashing for large-scale multimedia retrieval. *IEEE Transactions on Multimedia*, 23:4541–4554, 2020.

[Lu *et al.*, 2021] Xu Lu, Lei Zhu, Li Liu, Liqiang Nie, and Huaxiang Zhang. Graph convolutional multi-modal hashing for flexible multimedia retrieval. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1414–1422, 2021.

[Shen *et al.*, 2015] Xiaobo Shen, Fumin Shen, Quan-Sen Sun, and Yun-Hao Yuan. Multi-view latent hashing for efficient multimedia search. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 831–834, 2015.

[Shen *et al.*, 2018] Xiaobo Shen, Fumin Shen, Li Liu, Yun-Hao Yuan, Weiwei Liu, and Quan-Sen Sun. Multiview discrete hashing for scalable multimedia search. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 9(5):1–21, 2018.

[Shen *et al.*, 2020] Heng Tao Shen, Luchen Liu, Yang Yang, Xing Xu, Zi Huang, Fumin Shen, and Richang Hong. Exploiting subspace relation in semantic labels for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 33(10):3351–3365, 2020.

[Shen *et al.*, 2023] Xiaobo Shen, Yinfan Chen, Shirui Pan, Weiwei Liu, and Yuhui Zheng. Graph convolutional incomplete multi-modal hashing. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7029–7037, 2023.

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

[Song *et al.*, 2013] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Jiebo Luo. Effective multiple feature hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia*, 15(8):1997–2008, 2013.

[Sun *et al.*, 2024] Yuan Sun, Jian Dai, Zhenwen Ren, Yingke Chen, Dezhong Peng, and Peng Hu. Dual self-paced cross-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15184–15192, 2024.

[Tan *et al.*, 2022] Wentao Tan, Lei Zhu, Weili Guan, Jingjing Li, and Zhiyong Cheng. Bit-aware semantic transformer hashing for multi-modal retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 982–991, 2022.

[Tan *et al.*, 2023] Wentao Tan, Lei Zhu, Jingjing Li, Zheng Zhang, and Huaxiang Zhang. Partial multi-modal hashing via neighbor-aware completion learning. *IEEE Transactions on Multimedia*, 2023.

[Tu *et al.*, 2024] Rong-Cheng Tu, Xian-Ling Mao, Jinyu Liu, Wei Wei, Heyan Huang, et al. Similarity transitivity broken-aware multi-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Wang *et al.*, 2015] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. Deep multimodal hashing with orthogonal regularization. In *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[Wang *et al.*, 2023] Yuanzhi Wang, Zhen Cui, and Yong Li. Distribution-consistent modal recovering for incomplete multimodal learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22025–22034, 2023.

[Wu *et al.*, 2019] Dayan Wu, Qi Dai, Jing Liu, Bo Li, and Weiping Wang. Deep incremental hashing network for efficient image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9069–9077, 2019.

[Wu *et al.*, 2022a] Dayan Wu, Qinghang Su, Bo Li, and Weiping Wang. Efficient hash code expansion by recycling old bits. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 572–580, 2022.

[Wu *et al.*, 2022b] Xiao-Ming Wu, Xin Luo, Yu-Wei Zhan, Chen-Lu Ding, Zhen-Duo Chen, and Xin-Shun Xu. Online enhanced semantic hashing: Towards effective and efficient retrieval for streaming multi-modal data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 4263–4271, 2022.

[Wu *et al.*, 2023] Dayan Wu, Qi Dai, Bo Li, and Weiping Wang. Deep uncoupled discrete hashing via similarity matrix decomposition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19:1–22, 2023.

[Wu *et al.*, 2024] Dayan Wu, Qinghang Su, Bo Li, and Weiping Wang. Pairwise-label-based deep incremental hashing with simultaneous code expansion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9169–9177, 2024.

[Yang *et al.*, 2017] Rui Yang, Yuliang Shi, and Xin-Shun Xu. Discrete multi-view hashing for effective image retrieval. In *Proceedings of the 2017 ACM on international conference on multimedia retrieval*, pages 175–183, 2017.

[Zhang *et al.*, 2020] Wanqian Zhang, Dayan Wu, Yu Zhou, Bo Li, Weiping Wang, and Dan Meng. Deep unsupervised hybrid-similarity hadamard hashing. In *Proceedings of the 28th ACM international conference on multimedia*, pages 3274–3282, 2020.

[Zhang *et al.*, 2021] Wanqian Zhang, Dayan Wu, Yu Zhou, Bo Li, Weiping Wang, and Dan Meng. Binary neural network hashing for image retrieval. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 1318–1327, 2021.

[Zhang *et al.*, 2022] Wanqian Zhang, Dayan Wu, Chule Yang, Bo Li, and Weiping Wang. Clustering and separating similarities for deep unsupervised hashing. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1655–1659. IEEE, 2022.

[Zhang12 *et al.*, 2019] Wanqian Zhang12, Dayan Wu, Bo Li, Xiaoyan Gu, Weiping Wang, and Dan Meng. Fast and multilevel semantic-preserving discrete hashing. 2019.

[Zheng *et al.*, 2020] Chaoqun Zheng, Lei Zhu, Zhiyong Cheng, Jingjing Li, and An-An Liu. Adaptive partial multi-view hashing for efficient social image retrieval. *IEEE Transactions on Multimedia*, 23:4079–4092, 2020.

[Zheng *et al.*, 2022] Chaoqun Zheng, Lei Zhu, Zheng Zhang, Jingjing Li, and Xiaomei Yu. Efficient semi-supervised multimodal hashing with importance differentiation regression. *IEEE Transactions on Image Processing*, 31:5881–5892, 2022.

[Zhu *et al.*, 2020] Lei Zhu, Xu Lu, Zhiyong Cheng, Jingjing Li, and Huaxiang Zhang. Deep collaborative multi-view hashing for large-scale image search. *IEEE Transactions on Image Processing*, 29:4643–4655, 2020.

[Zhu *et al.*, 2023] Lei Zhu, Chaoqun Zheng, Weili Guan, Jingjing Li, Yang Yang, and Heng Tao Shen. Multi-modal hashing for efficient multimedia retrieval: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[Zhu *et al.*, 2024] Jian Zhu, Yu Cui, Zhangmin Huang, Xingyu Li, Lei Liu, Lingfang Zeng, and Li-Rong Dai. Adaptive confidence multi-view hashing for multimedia retrieval. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7900–7904. IEEE, 2024.