Contrastive Incomplete Cross-Modal Hashing

Haoyang Luo, Zheng Zhang, and Liqiang Nie, *Senior Member, IEEE*

***Abstract*—The success of current deep cross-modal hashing admits a default assumption of the *fully-observed* cross-modal data. However, such a rigorous common policy is hardly guaranteed for practical large-scale cases, which directly disable the training of prevalent cross-modal retrieval methods with incomplete cross-modal instances and unpaired relations. The main challenges come from the collapsed semantic- and modality-level similarity learning as well as uncertain cross-modal correspondence. In this paper, we propose a Contrastive Incomplete Cross-modal Hashing (CICH) network, which simultaneously determines the cross-modal semantic coordination, unbalanced similarity calibration, and contextual correspondence alignment. Specifically, we design a prototypical semantic similarity coordination module to globally rebuild partially-observed cross-modal similarities under an asymmetric learning scheme. Meanwhile, a semantic-aware contrastive hashing module is established to adaptively perceive and remedy the unbalanced similarities across different modalities with the semantic transition for generating discriminative hash codes. Additionally, a contextual correspondence alignment module is conceived to maximally capture shared knowledge across modalities and eliminate the correspondence uncertainty via a dual contextual information bottleneck formula. To the best of our knowledge, this is *the first successful attempt* of enabling contrastive learning to incomplete deep cross-modal hashing. Extensive experiments validate the superiority of our CICH against state-of-the-art methods.**

***Index Terms*—Correspondence calibration, cross-modal hashing, incomplete contrastive learning, similarity reconstruction.**

I. Introduction

RECENTLY, multimodal data processing capabilities have been emerging as one of the most important potentials of data-driven artificial intelligence with its powerful fusion of real-world visual perception and conceptual language ability. Cross-modal hashing (CMH) has been recognized as an efficient technique in retrieving semantically relevant targets when responding to a query sample across different data modalities among massive data sources. Technically, CMH achieves this by constructing unified binary codes across modalities based on similarity measurements. To facilitate an accurate cross-modal similarity search, a challenging impediment involves measuring the similarity between intrinsic yet nontransferable data forms, commonly referred to as the heterogeneous gap.

A substantial number of CMH methods [1], [2] have been proposed to enable fast and accurate nearest-neighbor searches from various perspectives. To effectively bridge the semantic and heterogeneity gaps, most supervised CMH methods aim to learn a common space [3], [4], [5] to preserve desirable similarities. One paradigm is dedicated to directly merging modality feature spaces into a unified one, employing a pairwise similarity-preserving loss [4]. This learning scheme optimizes paired data modalities to maximize the compactness of similar samples and the separability of dissimilar ones. Another common strategy to address the semantic gap is to use semantic categories as hash proxies [5], bridging individual modality spaces and acquiring general class-level knowledge. Recent works [6] employ label-sample contrastive learning to align semantics, resulting in promising cross-modal retrieval performance.

Nevertheless, most CMH methods are built on the ideal yet default assumption that all available multi-modal training data are *religiously complete* without any missing points. Moreover, they emphasize fully paired semantic relationships between modalities, i.e., *fully paired instances*. In practice, meeting such a strict precondition is rarely feasible due to limited resources for labor-intensive tasks such as data annotation, pairing, and filtering. Therefore, a more realistic assumption is *incomplete multi-modal* datasets. For example, website videos may lack captions, and pictures are also uncommon to have paired audio samples. Real-world multi-modal data often violate the assumption of full pairing, exhibiting unaligned relationships, which enlarge the modal heterogeneity gaps and impede the learning of discriminative hash codes. Therefore, addressing the complexities of incompleteness becomes essential for practical and robust solutions in cross-modal and multi-modal learning.

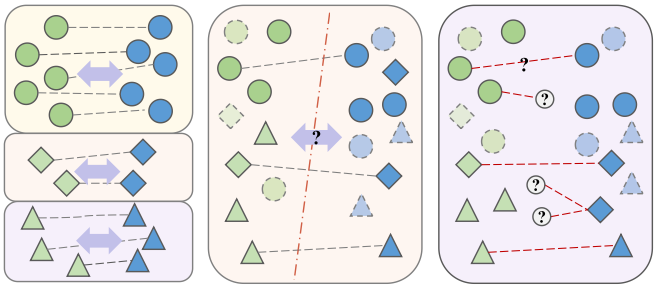
When addressing incompleteness, existing methods face several significant deficiencies. On the one hand, label distributions in incomplete modality sets are nonequivalent and unaligned due to random missing in different modalities. This results in partial and local relationships between samples, causing semantic shifts across modalities. Consequently, inaccurate semantic and modality measurements can severely degrade cross-modal similarity learning. On the other hand, they must contend with the challenge of vanished cross-modal correspondence. In cross-modal learning, the default assumption of one-to-one representation consistency between natural pairs can be challenged on both the instance and modal levels when dealing with limited

Manuscript received 28 August 2023; revised 21 March 2024; accepted 27 May 2024. Date of publication 14 June 2024; date of current version 27 September 2024. This work was supported in part by Shenzhen Science and Technology Program under Grant RCYX20221008092852077, in part by the National Natural Science Foundation of China under Grant 62372132, and in part by the Guangdong Natural Science Foundation under Grant 2023A1515010057. Recommended for acceptance by S. Salihoglu. *(Corresponding author: Zheng Zhang.)*

The authors are with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China, and also with Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: luohaoyang.lalutte@gmail.com; darrenzz219@gmail.com; nieliqiang@gmail.com).

Our code is available at https://github.com/DarrenZZhang/CICH.

Digital Object Identifier 10.1109/TKDE.2024.3410388



(a) Ideal cross-modal data. (b) Partially cross-modal data. (c) Correspondence uncertain data.

Fig. 1. An illustration of impediments in incomplete cross-modal hashing. Colors and shapes represent different modalities and classes, respectively, and horizontal lines indicate pairs. In an ideal situation (a), cross-modal data are perfectly paired with counterparts, forming clusters with other samples of similar semantics. However, when dealing with partially observed samples denoted by dashed borders in (b), the semantic and modal relationships become ambiguous, impacting the learning of cross-modal discriminative representations. Furthermore, the presence of partial samples illustrated in (c) introduces uncertain instance correspondence, leading to a depletion of the precise one-to-one feature alignment and hindering cross-modal knowledge interaction.

observations of natural pairs, leading to uncertain cross-modal correspondence. This also introduces uncertainty in the joint information distribution of modalities, as each modality observes a different knowledge space. Consequently, current CMH methods can only perform correspondence-uncertain similarity learning, lacking an emphasis on cross-modal transferability. The deficiencies highlighted in Fig. 1 allow us to propose a novel task in supervised CMH, termed Incomplete Cross-Modal Hashing (ICMH), for effective similarity learning on incomplete data.

To the best of our knowledge, only a few works have investigated the ICMH problem. For example, the dual-aligned variational autoencoders (DAVAE) [7] incorporates variational autoencoders (VAEs) into cross-modal retrieval for embedding learning and missing feature generation. However, it is challenging to recover precise features based only on the limited knowledge of distribution metrics, which may compromise modal discriminability and transferability, especially when dealing with multiple labels. Apart from generative methods, the multimodal coordinated clustering network (MCCN) [8] introduces unsupervised prototypes to substitute similarity learning and recover trainable samples. However, such a learning scheme neglects the imbalanced nature of incomplete sample classes and is prone to multi-class complexity, resulting in inferior semantic discriminability. Consequently, current ICMH approaches are significantly limited in their ability to learn sufficient modal and semantic similarities.

In this work, we propose a novel Contrastive Incomplete Cross-modal Hashing (CICH), which jointly performs balanced similarity learning and recovers informative and precise cross-modal features. Our CICH regulates three cooperative losses, including Prototypical Semantic Similarity Coordination (PSSC), Semantic-aware Contrastive Hashing (SaCH), and Contextual Correspondence Alignment (CCA). Specifically, PSSC learns semantic similarity asymmetrically by coordinating modal binary codes and learnable intermediate prototypes. Inspired from contrastive learning, SaCH further achieves balanced cross-modal similarity learning through a novel contrastive hashing scheme to improve the modal structure and semantic relevance. Unlike current contrastive hashing [6], [9], [10], [11] that makes strong assumptions on data characteristics or inadequately construct similarity relationships, our SaCH calibrates the destructed cross-modal similarity distribution through a relation propagation of positive pairs to construct stable similarity learning for recovered samples. Furthermore, our CCA adopts a dual learning strategy to align and recover modal representation with a novel contextual variational information bottleneck, transferring maximal modality-common information while absorbing cross-modal contextual knowledge for informative representation transformation and recovery. The main contributions of this work are as follows:

1) We propose a principled Contrastive Incomplete Cross-modal Hashing (CICH) framework, which jointly performs semantic and similarity calibration as well as missing feature restoration, for efficient incomplete cross-modal retrieval. To the best of our knowledge, there is no prior study on supervised CMH with incomplete instances, not to mention the contrastive learning-based ICMH.
2) To learn discriminative hash codes on incomplete cross-modal data, the prototypical similarity coordination loss and semantic-aware contrastive loss are designed to recover global cross-modal similarity with asymmetric transition and stabilize the imbalanced contrastive sample pairs with semantic inference.
3) To eliminate the correspondence uncertainty, a contextual correspondence alignment loss is constructed to ensure valid correspondence and maximal knowledge sharing between modalities via a contextual information bottleneck.
4) Extensive experiments verify that our model can outperform competitive CMH methods in various incomplete settings. It confirms the superiority of our method against increasing levels of incompleteness.

The following parts of this paper are arranged in the following manner. In Section II, we introduce related works to our proposed method. In Section III, we present our proposed CICH in detail and briefly develop our learning algorithm. In Section IV, we give extensive experimental results and corresponding analysis comprehensively. In Section V, we draw a conclusive statement of our proposed method.

II. Related Work

A. Cross-Modal Hashing

Cross-modal hashing (CMH) efficiently facilitates cross-modal retrieval by preserving similarity in binary codes. CMH methods [3], [4], [5], [12], [13], [14], [15] are broadly classified as shallow CMH and deep CMH.

As a representative shallow CMH method, discrete cross-modal hashing (DCH) [3] attempts to learn modality-private feature transformations for classification while producing common codes with the learned features. Semantics-Preserving hashing

(SePH) [16] performs CMH on a label-constructed affinity matrix. SCRATCH [12] utilizes collective matrix factorization and label semantic embedding to learn latent representations and codes. These shallow CMH methods typically employ a two-step training paradigm, i.e., modal feature extraction, and cross-modal hashing.

On the other hand, deep CMH methods have enabled end-to-end training that incorporates feature compression into hashing to leverage the powerful representation ability of deep neural networks. DCMH [4] is the first work to incorporate a semantic similarity matrix into CMH learning. SSAH [5] employs a self-learned label encoder to learn similarity with adversarial training for consistent codes. Deep adversarial discrete hashing (DADH) [13] proposes to learn about weighted cosine triplets with adversarial training to rank similarities. Recent work [14] identifies multiple ways for semantic supervision based on affinity matrix, label encoding, and class prototypes. However, these methods are designed under the fully observed assumption, limiting their ability to learn semantic- and modality-similarity with clear correspondence with incomplete samples.

In a nutshell, existing CMH methods are vulnerable in the more modal-imbalanced situations with incomplete samples. Our work aims to recover a balanced pair distribution by learning a globally balanced semantic similarity. Through complementarily designed contrastive hashing, our approach adjusts the biased modal relationships while preserving cross-modal transferability. Simultaneously, our model determines clear cross-modal correspondence and optimally recovers informative modal features.

B. Contrastive Learning

Contrastive learning is a prominent paradigm in self-supervised representation learning [17]. Numerous studies have explored diverse contrastive objectives and methods of leveraging data pairs. For instance, deep InfoMax (DIM) [18] introduces the infoNCE loss for contrastive local and global image representations. MoCo [19] employs momentum updates for contrastive learning on large datasets. BYOL [20] introduces cross-view prediction for positive-only contrastive learning with momentum updates. Contrastive learning has also been integrated into pre-training large-scale vision-language models [21], [22] based on various training schemes.

In retrieval, several unsupervised methods utilize contrastive learning for hash code generation. CIBHash [9] utilizes the NT-Xent contrastive loss to optimize stochastic hash codes and reconsiders the learning process into an information bottleneck scheme. CIMON [10] considers unsupervised consistency learning with semantic and contrastive objectives. However, their assumptions about one-to-one correspondence and data distribution availability cannot be guaranteed in the incomplete scenarios. UCMFH [11] constructs a multi-modal fusion transformer and applies contrastive learning in both feature and hash spaces. Nonetheless, the proposed multi-modal fusion is impractical with incomplete modalities. In a supervised approach [6], contrastive learning is introduced between modal binary codes and their corresponding labels. However, the utilization of semantic information in pointwise correspondence with modalities can be insufficient for modal alignment.

Notably, none of these methods explore the similarity reconstruction potential of contrastive hashing in ICMH. To learning cross-modal balanced binary representations with incomplete instances, we distinguish our method from current contrastive hashing schemes by explicitly performing contrastive learning between modalities.

C. Incomplete Cross-Modal Retrieval

Limited works specifically target the challenge of incomplete cross-modal retrieval. For instance, an unsupervised method [23] constructs modality-specific and unified global anchor graphs to recover affinity relationships for learning binary codes with incomplete data. DAVAE [7] replaces pairwise semantic learning with class recognition and utilizes variational autoencoders for feature learning across modalities. Wu et al. [24] propose modality-cyclic generative models to compensate for missing features directly. Additionally, retrieval methods like MCCN [8] and PAN [25] focus on constructing class-specific prototypes to aid similarity learning while regenerating modal features based on similarity metrics. MCCN [8] recovers unobserved features by coordinating available modal features whereas PAN [25] leverages class prototypes with similar samples to restore incomplete modalities in a feature propagation manner.

Although the aforementioned methods make efforts to tackle the incomplete issue, their limitations are evident. These methods struggle to establish balanced and stable similarity learning between modalities and overlook the cross-modal transferability introduced by modal correspondence knowledge. Addressing these deficiencies should be prioritized when dealing with incomplete CMH.

III. The Proposed Method

In this section, we first briefly introduce the notations and problem formulation for the incomplete cross-modal hashing problem. Then, we introduce our proposed CICH in detail, including calibrated contrastive network and contextual knowledge alignment, followed by optimization. The overall framework of CICH is illustrated in Fig. 2.

A. Problem Definition

Without loss of generality, we consider the two-modality (i.e., image and text) situation, which can be easily adapted to multiple modalities. The conventional CMH problem involves a set of paired modalities with one-to-one mappings, along with the corresponding label set for image-text retrieval, denoted as $\mathcal{O} = \{\boldsymbol{o}_i\}_{i=1}^{n} = \{(\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{l}_i)\}_{i=1}^{n}$. Here, $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ represent the textual annotation and their corresponding visual image of the i-th instance while $\boldsymbol{l}_i$ is its semantic label. However, in real scenarios, cross-modal data may only partially exist, represented as $\mathcal{O} = \mathcal{F} \cup \mathcal{U} \cup \mathcal{V}$, where $\mathcal{F} = \{(\boldsymbol{u}_i, \boldsymbol{v}_i, \boldsymbol{l}_i)\}_{i=1}^{n_f}$ is the complete part, whilst $\mathcal{U} = \{(\boldsymbol{u}_i, \boldsymbol{l}_i)\}_{i=n_f+1}^{n_f+n_u}$ and $\mathcal{V} = \{(\boldsymbol{v}_i, \boldsymbol{l}_i)\}_{i=n_f+n_u+1}^{n}$ are the incomplete single-modal data for text and image, respectively, where $n = n_f + n_u + n_v$ is the

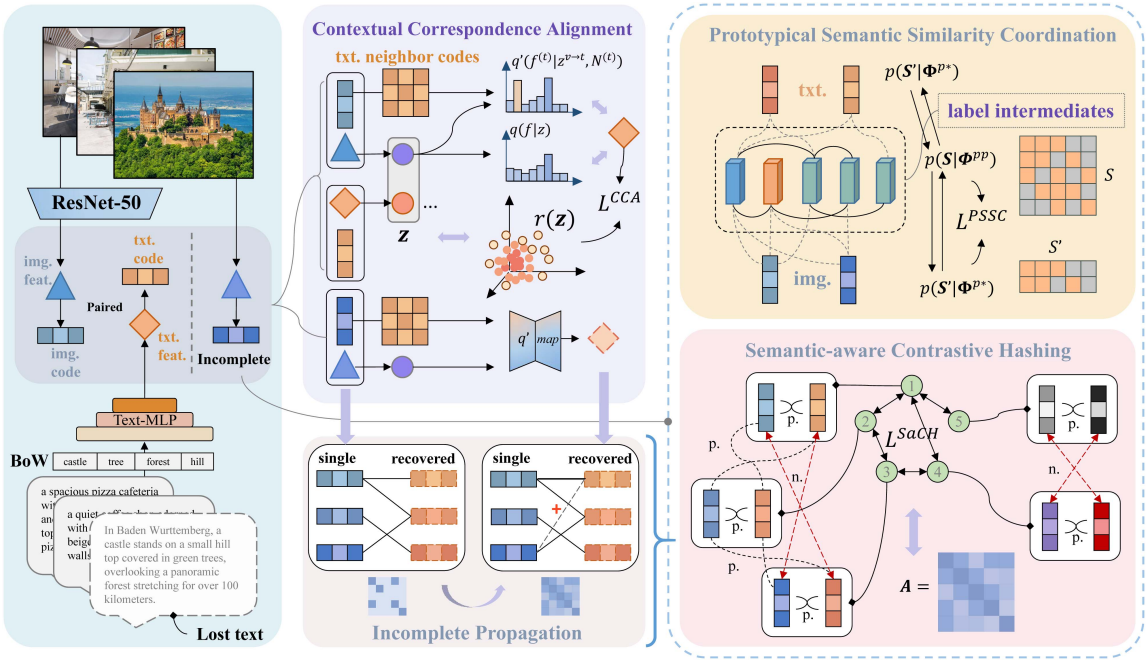


Fig. 2. An outline of our proposed Contrastive Incomplete Cross-modal Hashing (CICH). The distinctive shapes represent real-valued features of different modalities, while the grids represent vectors or matrices. CICH features three coordinated modules, i.e., prototypical semantic similarity coordination, semantic-aware contrastive hashing, and contextual correspondence alignment. The prototypical semantic similarity coordination can reconstruct semantic similarity by asymmetrically learning between global intermediates and modal codes, while semantic-aware contrastive hashing enables modal similarity learning through semantic transition and propagation of positive pairing. Furthermore, contextual correspondence alignment eliminates uncertainty in cross-modal correspondence.

total number of training instances. The goal of supervised incomplete CMH is to learn hash functions $H^{u,v}$ to encode text and image samples into binary codes $\boldsymbol{b}_i^u = sign(\boldsymbol{h}_i^u) = sign(H^u(\boldsymbol{u}_i))$ and $\boldsymbol{b}_i^v = sign(\boldsymbol{h}_i^v) = sign(H^v(\boldsymbol{v}_i))$, respectively. Here, $H^{u,v}(\cdot) = G^{u,v}(F^{u,v}(\cdot))$ uses modal feature extractors $F^{u,v}$ and hash layers $G^{u,v}$ to encode each sample and preserve the Hamming ranks given the similarity matrix S where $S_{ij} = 1$ if $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$ have a common label, otherwise $S_{ij} = 0$. The incomplete CMH is defined as follows:

Definition 1: Incomplete Cross-modal Hashing: Formally, a complete cross-modal instance comprises M modalities denoted as $\mathbb{O} = \{o^{(m)}\}_{m=1}^M$ along with its associated semantic label l. A partially-observed cross-modal instance is a non-empty subset $\overline{\mathbb{O}}$ of its fully-observed complete cross-modal instance (i.e., $\overline{\mathbb{O}} \subseteq \mathbb{O}$ and $\overline{\mathbb{O}} \neq \varnothing$) with $\overline{M}$ modalities, where $1 \leq \overline{M} \leq M$. Provided with an incomplete cross-modal dataset $\{(o_i^{(m)}, l_i)\}_{i=1}^N$, incomplete cross-modal hashing seeks to learn binary codes $\{\boldsymbol{B}^{(m)}\}_{m=1}^M$ for these instances and hash functions $\{H^{(m)}(\cdot)\}_{m=1}^M$ for each modality.

This insufficiency in samples significantly limits cross-modal relationships. In another type of problem, unpaired cross-modal hashing, only the sample relationship is partially known, whereas incomplete cross-modal hashing even sacrifices samples, resulting in a more challenging problem.

B. Prototypical Semantic Similarity Coordination

Without sufficient pairs, the semantic similarity cannot be directly learned as the semantic distributions become imbalanced and shifted due to the unequal presence of samples. Therefore, we seek to leverage the globally observed label semantics as a surrogate for the cross-modal counterpart. Toward this goal, we train an intermediary code network $\boldsymbol{\psi}_i = g(\boldsymbol{l}_i) = \boldsymbol{W}_p^\top \boldsymbol{l}_i$ as prototypical codes of labels. We introduce a prototypical semantic similarity coordination function to optimize the similarity between prototypical codes and modalities, i.e.,

$$\min_{\boldsymbol{B}^{p,u}} \sum_{i=1}^{n} \sum_{j=1}^{N_u} -(\boldsymbol{\psi}_i)^\top \boldsymbol{\psi}_j S_{ij} - (\boldsymbol{\psi}_i)^\top \boldsymbol{h}_j^u S_{ij}$$
$$s.t.\ \boldsymbol{b}_i^p = sgn(\boldsymbol{\psi}_i), \boldsymbol{b}_i^u = sgn(\boldsymbol{h}_i^u), \boldsymbol{b}_i^{p,u} \in \{-1, 1\}^r, \quad (1)$$

where $N_u = n_f + n_u$ is the number of available text samples, $\boldsymbol{\psi}_i$ is the prototypical code alternative produced by $g(\cdot)$, and $\boldsymbol{h}^u$ is the text binary code. $\boldsymbol{b}_i^{p,u}$ is the i-th column of $\boldsymbol{B}^{p,u}$, respectively. The above objective is similar for image modality and can be transferred into a negative log-likelihood objective. We first formulate the probability of the predicted similarity matrix S under the condition of inner product similarity as

$$p(\boldsymbol{S}|\boldsymbol{\Phi}) = \begin{cases} \delta(\Phi_{ij}), & S_{ij} = 1 \\ 1 - \delta(\Phi_{ij}), & S_{ij} = 0 \end{cases}, \quad (2)$$

where $\Phi_{ij} = \frac{1}{2}\boldsymbol{b}_i^\top \boldsymbol{b}_j$ and $\delta(\Phi_{ij}) = \frac{1}{1+e^{-\Phi_{ij}}}$. To facilitate the reconstruction of prototypical semantic similarity learning, the negative log-likelihood objective is subsequently formulated in an asymmetric manner as

$$\min_{\boldsymbol{W}_p, \theta^{u,v}} \mathcal{L}^{PSSC} = \sum_{k \in M} (\mathcal{L}_P^k + \mathcal{L}_Q^k)$$

$$= \sum_{k\in M}\left(-\sum_{i\in I_k}\sum_{j=1}^{n}\left(S_{ij}\Lambda_{ij}^{k} - \log\left(1+e^{\Lambda_{ij}^{k}}\right)\right) + \sum_{i\in I_k}\|\boldsymbol{h}_i^k - \boldsymbol{b}_i\|_F^2\right) s.t.\ \boldsymbol{B}\in\{-1,1\}^{r\times n}, \quad (3)$$

where $M=\{p,u,v\}$, $\boldsymbol{H}^p=\boldsymbol{\Psi}$, $\Lambda_{ij}^k=\frac{1}{2}(\boldsymbol{h}_i^k)^\top\boldsymbol{\psi}_j$, and I_k is the index set for a mini-batch of training samples of modality k. The global visibility of prototypical label codes is utilized in teaching a mini-batch of m image and text hash codes using all n available codes. The learned similarity knowledge in $\boldsymbol{\psi}$ is distilled through Λ_{ij}^k to asymmetrically transfer comprehensive semantic knowledge from implicit label intermediates to image and text codes, so that the disrupted semantic similarity learning can be efficiently rebuilt with global relation preservation.

C. Semantic-Aware Contrastive Hashing

Although similarity learning between prototypical intermediates and modal codes successfully recovers semantic similarity, precise modal similarity learning remains challenging. To this end, we introduce a contrastive hashing framework to tackle the imbalanced comparison between modal samples. Different from label contrastive hashing method [6], which is susceptible to class imbalance and semantic shifts in the presence of missing samples, our proposed semantic-aware contrastive hashing explicitly operates between modalities. It transforms the one-to-one positive pairing strategy into a calibrated contrastive relationship, thereby improving the distribution of positive pairs.

In conventional supervised CMH, the contrastive objective can be formulated based on samples and their corresponding labels as

$$\mathcal{L}_i^k = -log\frac{e^{\delta(\frac{1}{2}(\boldsymbol{h}_i^k)^\top\boldsymbol{\psi}_i)/\tau}}{\sum_{j=1}^{n}e^{\delta(\frac{1}{2}(\boldsymbol{h}_i^k)^\top\boldsymbol{\psi}_j)/\tau}} \quad (4)$$

for the i-th example in modal k, where τ is a temperature parameter. While effective for ordinary CMH, this approach is not helpful for semantic similarity learning between incomplete modalities due to the unequal loss of class samples. This imbalance leads to a biased contrastive power for each modality. Additionally, the implicit sample-label relationship is suboptimal for CMH since the retrieval is performed between modalities. To address the issue of insufficient positive pairs, both in terms of the sample-label relationship and natural cross-modal pairs, explicit preservation of modal similarity is necessary. Thus, we propose a novel semantic-aware contrastive hashing that infers and transfers the contrastive relationship through a semantic adjacency matrix between asymmetrically sampled modal sets $\{(\boldsymbol{u}_i,\boldsymbol{l}_i)\}_{i\in I_u}$ and $\{(\boldsymbol{v}_j,\boldsymbol{l}_j)\}_{j\in I_v}$. The semantic-aware contrastive hashing objective is defined as

$$\mathcal{L}_i^{u\to v} = \sum_{r\in I_v} -\boldsymbol{A}_{ir}^{uv} log\frac{e^{\delta(\frac{1}{2}(\boldsymbol{h}_i^u)^\top\boldsymbol{h}_r^v)/\tau}}{\sum_{j\in I_v}e^{\delta(\frac{1}{2}(\boldsymbol{h}_i^u)^\top\boldsymbol{h}_j^v)/\tau}}, \quad (5)$$

where u and v represent different modalities, $\delta(\Phi)=\frac{1}{1+e^{-\Phi}}$ is the sigmoid function, and $\boldsymbol{A}^{uv}\in\{0,1\}^{|I_u|\times|I_v|}$ is a cross-modal adjacency matrix defined differently for complete points and recovered ones. By considering recovered samples in Section III-D, the I_u and I_v here are sampled from the entire training set (i.e., $I_{u,v}\subset[n]$). This objective function transitions and expands the positive pair relationship from a single imbalanced positive pair to a range of adjacent pairs for each anchor sample, reconstructing a balanced contrastive relationship and establishing explicit cross-modal hashing. Here, we define $\boldsymbol{A}^{uv}$ between complete pairs as the corresponding segment of $\boldsymbol{S}$ for simplicity.

Concerning the recovered incomplete samples acquired in Section III-D, their semantic relationship may be ambiguous, particularly when the level of incompleteness is significant. To address this, we employ a propagating strategy to further expand the scope of positive pairs. The segments associated with recovered samples in $\boldsymbol{A}^{uv}$ are redefined as

$$\boldsymbol{A}^{uv} = \mathbb{I}\left((\boldsymbol{S}^{uv}\boldsymbol{S}^{uv\top})\boldsymbol{S}^{uv}\right), \quad (6)$$

where v represents the incomplete modality and $\mathbb{I}(\cdot)$ is an element-wise operator that assigns a value of 1 to positive elements and 0 to non-positive elements. This propagation strategy encourages both pairwise similarity learning and multi-hop similarity learning between samples. The propagated adjacency matrix facilitates the transfer of neighboring conditions, thereby stabilizing the learning process for recovered samples with ambiguous semantics. The objective function for both modalities is formulated as

$$\mathcal{L}^{SaCH} = \sum_{i\in I_u}\mathcal{L}_i^{u\to v} + \sum_{j\in I_v}\mathcal{L}_j^{v\to u}. \quad (7)$$

D. Contextual Correspondence Alignment

In a semi-observed data environment, another factor that collapses the training process is the vague and uncertain correspondence relationship learned by the model due to the insufficient instance-level aligned pairs. To address this, it is necessary to align heterogeneous instances and modalities in an information-efficient way that alleviates uncertainty and simultaneously recovers representations. Thus, we conceive to conduct modal alignment at the instance level rather than adversarial learning at the general distribution level. Inspired by the critical information bottleneck (IB) principle [26], we introduce a modal transfer variable $\boldsymbol{z}_i^{v\to u}$ as a compressed intermediate between visual feature $\boldsymbol{f}_i^v$ and textual feature $\boldsymbol{f}_i^u$. This variable can act as an information selector to purify the source given the target and retrieve cross-modal common knowledge between the two. The correspondence alignment objective is formulated as

$$\max\mathcal{L}_{CA} = I\left(\boldsymbol{z}^{v\to u},\boldsymbol{f}^u\right) - \beta I\left(\boldsymbol{z}^{v\to u},\boldsymbol{f}^v\right), \quad (8)$$

where $I(\cdot,\cdot)$ indicates the mutual information between two variables, and β is a hyperparameter quantifying the source forgetting. By introducing a variational strategy [27], [28], we optimize the variational lower bound of (8), where the lower bound of the first term and a variational upper bound of the second term are optimized. The variational objective is formulated

as

$$\mathcal{L}^{CA} = -\frac{1}{|I_v|}\sum_{i\in I_v}\mathbb{E}_{\varepsilon\sim p(\varepsilon)} log\left(q\left(\boldsymbol{f}_i^u|\boldsymbol{z}_i^{v\to u}\right)\right) + \beta\mathbb{E}_{p(\boldsymbol{f}_i^v)}\left[KL\left(p(\boldsymbol{z}_i^{v\to u}|\boldsymbol{f}_i^v), r\left(\boldsymbol{z}_i^{v\to u}\right)\right)\right], \quad (9)$$

where $KL(\cdot,\cdot)$ is the Kullback-Leibler divergence, $\boldsymbol{z}_i^{v\to u} = \xi(\boldsymbol{f}_i^v, \varepsilon)$ is the reparameterized variable of $\boldsymbol{f}_i^v$ given Gaussian variable ε, and $q(\boldsymbol{f}_i^u|\boldsymbol{z}_i^{v\to u})$ approximates variationally to $p(\boldsymbol{f}_i^u|\boldsymbol{z}_i^{v\to u})$. The marginal probability $p(\boldsymbol{z})$ is replaced by its Gaussian approximation $r(\boldsymbol{z})$.

To leverage the reconstructed knowledge correspondence, one can directly recover a text feature by resorting to the variational approach q. However, recovery based solely on shared modal knowledge may encounter an unknown contextual distribution in the target modality since the intermediate variable is optimized to be target-agnostic as a knowledge subset between modalities. Therefore, we seek to utilize cross-modal neighbor context as a complementary source to fill out the knowledge emptiness in the intermediate variables. The final objective for contextual correspondence alignment is expressed as

$$\mathcal{L}^{CCA} = -\frac{1}{|I_v|}\sum_{i\in I_v}\mathbb{E}_{\varepsilon\sim p(\varepsilon)}\left(log\left(q\left(\boldsymbol{f}_i^u|\boldsymbol{z}_i^{v\to u}\right)\right) + log\left(q'\left(\boldsymbol{f}_i^u|\boldsymbol{z}_i^{v\to u}, \mathcal{N}^t\left(\boldsymbol{f}_i^v\right)\right)\right)\right) + \beta\mathbb{E}_{p(\boldsymbol{f}_i^v)}\left[KL\left(p(\boldsymbol{z}_i^{v\to u}|\boldsymbol{f}_i^v), r\left(\boldsymbol{z}_i^{v\to u}\right)\right)\right], \quad (10)$$

where $\mathcal{N}^t(\boldsymbol{f}_i^v)$ is the features of the top$K$ similar text neighbors ranked in hash space *w.r.t.* the i-th visual sample, and q' is a context-conditioned approximator designed to fully recover the cross-modal correspondence. It is noteworthy that the q and q' play distinct roles. q extracts maximal overlapping knowledge between modalities, while q' tries to fuse the modal-agnostic features with modal private knowledge to recover sufficiently informative samples. Thus, the uncertain correspondence is hopefully solidified with such a dual approximation strategy to recover instance-level transferability. The objective for $u \to v$ is similar to (10) with alternated modal superscripts.

E. Optimization

Considering the (3), (7), and (10) together, we formulate the overall loss function for our Contrastive Incomplete Cross-modal Hashing (CICH) as

$$\mathcal{L}_{CICH} = \mathcal{L}^{PSSC} + \alpha\mathcal{L}^{SaCH} + \delta\mathcal{L}^{CCA}. \quad (11)$$

Herein, $\mathcal{L}^{PSSC}$ and $\mathcal{L}^{SaCH}$ represent the prototypical semantic similarity coordination loss and semantic-aware contrastive hashing loss, respectively. They work in a complementary manner to fully capture both semantic and modal similarity. $\mathcal{L}^{CCA}$ stands for the contextual correspondence alignment loss. α and δ are hyperparameters for module balance. For incomplete samples, we employ the learned mapper q' to recover the missing feature, completing the learning process for the semantic-aware contrastive hashing loss. In such a coordinated manner, modal features are jointly aligned and recovered to rescue the collapsed similarity learning and reconstruct the instance correspondence.

Algorithm 1: Optimization Algorithm for CICH.

Require: Training dataset $\boldsymbol{F}, \boldsymbol{U}, \boldsymbol{V}$; hash code length r; pairwise similarity matrix $\boldsymbol{S}$.
Ensure: Optimal discrete hashing codes $\boldsymbol{B}$, intermediate projection parameter $\boldsymbol{W}_p$ and modal network parameters $\boldsymbol{\theta}^{u,v}$, and parameters in q and q'.
1: **Initialization**
Randomly initialize $\boldsymbol{W}_p$, $\boldsymbol{\theta}^{u,v}$, q, q', δ, β, τ, and K.
Configure the learning-rate μ; the mini-batch size $m = 64$; maximum iterations $Epoch_n$.
2: **while** $iter < Epoch_n$ & not converged **do**
3: Compute loss $\mathcal{L}^{PSSC}$ and update $\boldsymbol{W}_p$ by optimizing problem (3);
4: Compute loss $\mathcal{L}_{CICH}$ and update $\theta^{u,v}$, q, and q' by optimizing problem (11);
5: Generate recovered sample features via q';
6: Update $\boldsymbol{B}$ by $\boldsymbol{B} = sign(\boldsymbol{H}^u + \boldsymbol{H}^v + \boldsymbol{\Psi})$;
7: **end while**

TABLE I
THE STATISTICS FOR DATASETS IN OUR EXPERIMENTS

Dataset	Train	Test	Database	Total	Text dim.	Classes
MIRFLICKR-25K	18,015	2,000	20,015	18,015	1,386	24
MS COCO	82,081	5,000	87,081	82,081	2,000	80
NUS-WIDE-10K	8,000	2,000	10,000	8,000	1,000	10
IAPR TC12	18,000	2,000	20,000	18,000	2,885	275
NUS-WIDE	50,000	2,085	52,085	50,000	1,000	21

We select larger training sets and adopt them as retrieval databases in incomplete CMH.

We optimize our overall objective function in an alternative manner, with detailed procedures outlined in Algorithm 1.

IV. EXPERIMENT

A. Experiment Settings

1) Evaluation Settings: We evaluate our method through extensive experiments on 5 widely-used datasets: i.e., MIRFLICKR-25K [29], MS COCO [30], NUS-WIDE-10K [31], IAPR TC-12 [32], and NUS-WIDE [31]. The statistics are outlined in Table I.

Following the configurations in [7], we employ three levels of incompleteness for the evaluation of incomplete CMH. In the easy setting, 50% of instances are unpaired, with at least one modality missing, while the remaining 50% are paired samples. Specifically, the training set is subdivided into 50% image-text pairs, 25% single images, and 25% single text (denoted as $(50\%, 25\%, 25\%)$). The unpaired missing modality is excluded from both training and evaluation. Similarly, the medium setting divides the training set into $(30\%, 35\%, 35\%)$, and the hard setting involves nearly unpaired instances $(10\%, 45\%, 45\%)$.

While the number of available training samples remains constant in each setting, we extend our evaluation by considering three detailed partition settings for both the training and testing phases: (1) *Discard* requires only paired samples for both the training and retrieval database and is designed for comparison with the later two settings. (2) *Enhance* utilizes incomplete samples in training as enhancement with only samples of full modalities for the retrieval test. (3) *Extend* performs training

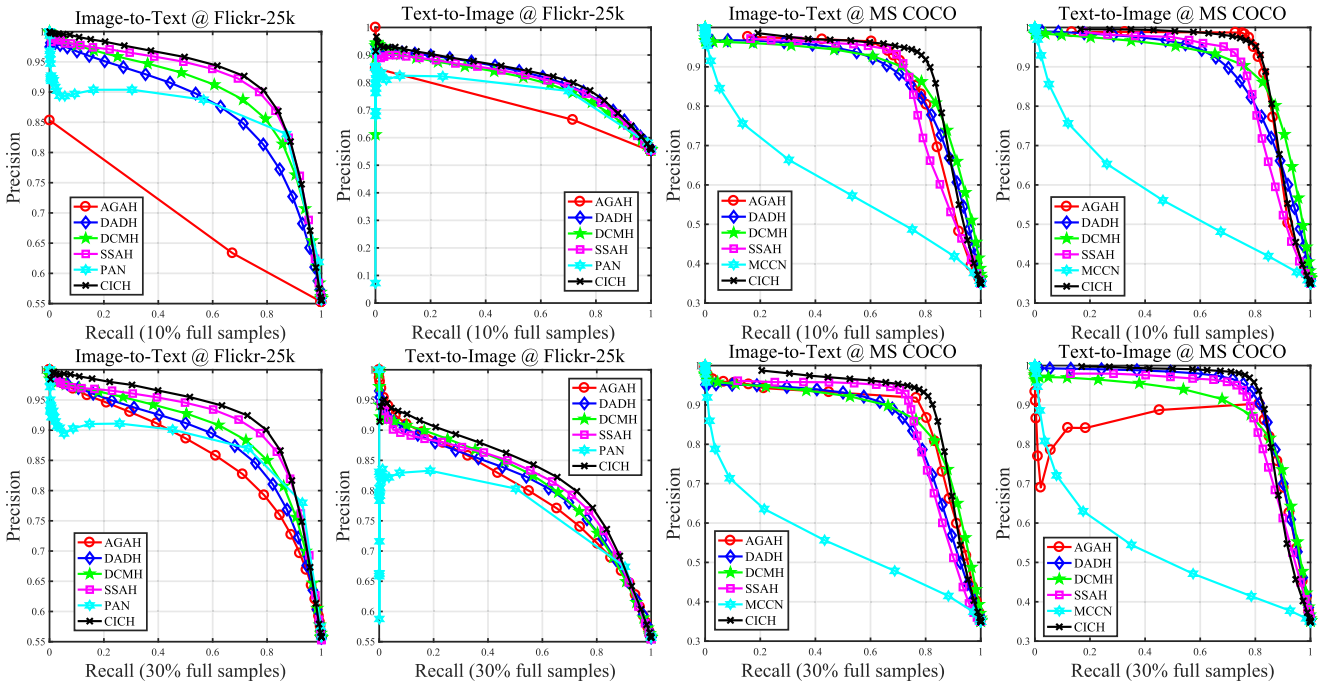


Fig. 3. The precision-recall curves of deep cross-modal retrieval methods w.r.t. different incomplete levels on the MIRFLICKR-25K and MS COCO datasets.

and retrieval validation on both the complete and incomplete instances. By default, *Extend* is used throughout the following experiments for a more comprehensive evaluation.

2) Evaluation Metrics: We adopt the mean average precision (mAP) as a comparison metric for all experiments in the following sections. The mAP is a common evaluation metric in the research of CMH, computed by averaging the AP of all query results. Specifically, given an image query $\boldsymbol{q}_i^v$, the text database is ranked as a sorted list $L_i^t = \{\boldsymbol{d}_j^t\}_{j=1}^{n_d}$. The AP of this query is computed as

$$AP = \frac{1}{R_{n_d}} \sum_{j=1}^{n_d} \frac{R_j}{j} \odot \mathrm{rel}\left(\boldsymbol{d}_j^t\right), \tag{12}$$

where R_j is the number of relevant instances in the first j retrieved results, and $\mathrm{rel}(\cdot)$ is an instance relevance indicator which is valued as 1 if $\boldsymbol{d}_j^t$ is relevant to $\boldsymbol{q}_i^v$ and otherwise 0. The text-to-image AP is calculated similarly. We evaluate image-to-text and text-to-image mAPs for all experimented methods. In addition, we also utilize the widely-used precision-recall (PR) curves and the NDCG@topK metric to demonstrate the model efficiency in the hash lookup task. The NDCG@topK metric is an effective metric for evaluating multi-label tasks.

3) Implementation Details: For feature extraction, we adopt ResNet-50 as the image backbone network or feature extractor for all compared methods, while the raw text feature is further fed into a 2-layer multi-linear perceptron (MLP). The image features extracted by the pretrained ResNet-50 network are then projected into a 1,024-dimensional intermediate feature which can be divided into 512-dimensional $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ for reparameterization to acquire $\boldsymbol{z}^{v\to t}$. The text MLP features a 4,096-dimensional hidden layer with a 1,024-dimensional output, which is processed similarly to acquire $\boldsymbol{z}^{t\to v}$. The predicted label and hash code are both computed after a corresponding linear layer on $\boldsymbol{\mu}$, with no activation function and a tanh activation, respectively. The q and q' are implemented as two simple 1-layer MLPs where we concatenate $\boldsymbol{z}$ and all the top-k nearest neighbor features as the input for q'. The label's prototypical encodings are acquired by a single linear projection into a vector of 32 dimensions, which is the hash bit number we set throughout the paper. We search for α and β together and search for the rest of the hyperparameters δ, τ, and K independently with a cross-validation strategy to get the best performance. The optimal parameters found are also simply adopted for MS COCO and NUS-WIDE-10K evaluations.

B. Experiment Results

1) Incomplete Cross-Modal Hashing: To validate the hamming ranking and incompleteness resistance abilities of our method, we compare it with 10 representative baseline methods, including 3 state-of-the-art shallow hashing methods, namely DCH [3], JIMFH [33], and SCRATCH [12], and 7 deep cross-modal retrieval methods. We expect our method to perform consistently well and superior across all evaluated incompleteness levels. Among the deep methods, 5 of them are based on deep hashing (DCMH [4], SSAH [5], AGAH [34], DADH [13], DCHMT [35]), while 2 of them (MCCN [8], PAN [25]) are representative incomplete cross-modal retrieval methods based on real values. Since no code is available for MCCN and PAN, we do our best to implement them and use available implementations for other methods. The mAP evaluation of the compared methods and our method on MIRFLICKR-25K, MS COCO,

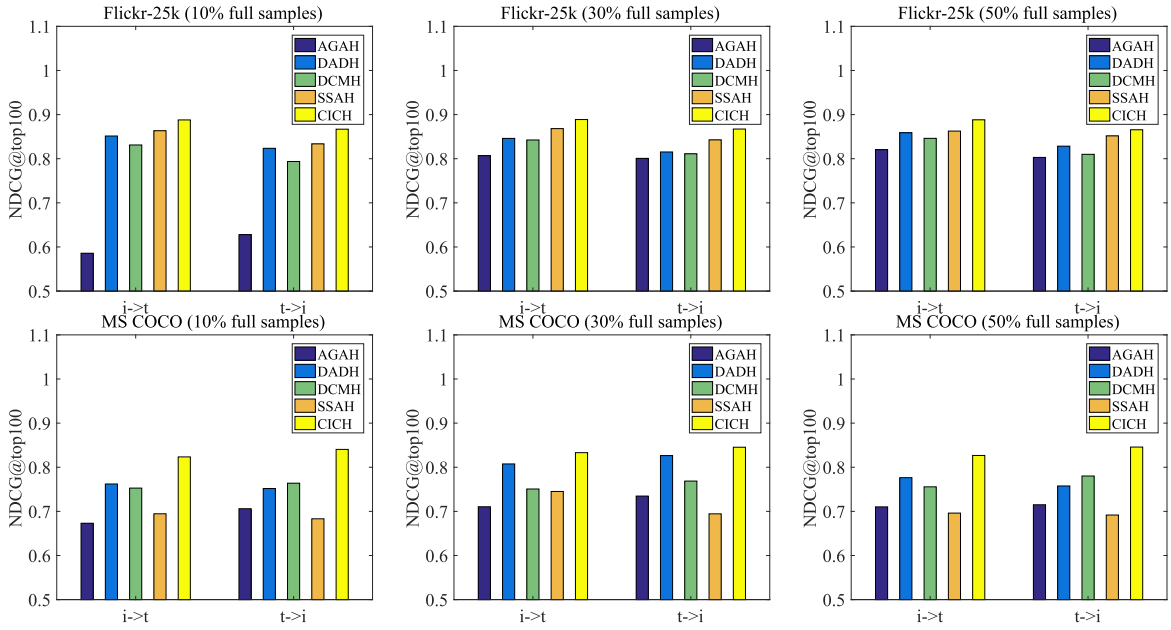


Fig. 4. The NDCG histogram w.r.t. different levels of incompleteness on the MIRFLICKR-25K and MS COCO datasets.

TABLE II
THE MAP ACCURACY COMPARISONS ON THE MIRFLICKR-25K, MS COCO, AND NUS-WIDE-10K DATASETS *W.R.T.* DIFFERENT LEVELS OF COMPLETENESS

Datasets	Methods	hard(10%,45%,45%)			medium(30%,35%,35%)			easy(50%,25%,25%)			Δ (easy - hard)(%)		
		i→ t	t→ i	mean	i→ t	t→ i	mean	i→ t	t→ i	mean	i→ t	t→ i	mean
MIRFLICKR-25K	DCH	0.597	0.622	0.610	0.713	0.683	0.698	0.733	0.692	0.713	13.60	7.00	10.30
	JIMFH	0.560	0.589	0.575	0.589	0.598	0.594	0.599	0.600	0.600	3.90	1.10	2.50
	SCRATCH	0.795	0.733	0.764	0.813	0.737	0.775	0.828	0.748	0.788	3.30	1.50	2.40
	DCMH	0.890	0.817	0.854	0.891	0.826	0.859	0.894	0.829	0.862	0.40	1.20	0.80
	SSAH	0.895	0.822	0.859	0.895	0.832	0.864	0.896	0.845	0.871	**0.10**	2.30	1.20
	AGAH	0.617	0.626	0.622	0.841	0.811	0.826	0.876	0.830	0.853	25.90	20.40	23.15
	DADH	0.876	0.814	0.856	0.883	0.828	0.859	0.884	0.835	0.864	0.80	2.10	1.45
	PAN	0.896	0.807	0.852	0.901	0.806	0.854	0.903	0.811	0.857	0.70	0.40	0.55
	DCHMT	0.860	0.802	0.832	0.862	0.805	0.835	0.864	0.809	0.835	0.40	0.70	0.55
	CICH(ours)	**0.917**	**0.845**	**0.881**	**0.918**	**0.845**	**0.882**	**0.924**	**0.847**	**0.886**	0.70	**0.20**	**0.45**
MS COCO	DCH	0.527	0.628	0.578	0.535	0.643	0.589	0.539	0.645	0.592	1.20	1.70	1.45
	JIMFH	0.414	0.461	0.438	0.462	0.485	0.474	0.479	0.492	0.486	6.50	3.10	4.8
	SCRATCH	0.539	0.600	0.570	0.603	0.668	0.636	0.609	0.671	0.640	7.00	7.10	7.05
	DCMH	0.681	0.710	0.696	0.689	0.718	0.704	0.695	0.724	0.710	1.40	**1.40**	1.4
	SSAH	0.613	0.623	0.618	0.615	0.638	0.627	0.625	0.645	0.635	1.20	2.20	1.7
	AGAH	0.576	0.600	0.588	0.595	0.618	0.607	0.621	0.634	0.628	4.51	3.40	3.955
	DADH	0.686	0.670	0.678	0.689	0.688	0.689	0.692	0.689	0.691	**0.60**	1.90	**1.25**
	MCCN	0.672	0.617	0.645	0.673	0.621	0.647	0.704	0.636	0.670	3.20	1.90	2.55
	DCHMT	0.656	0.631	0.644	0.662	0.636	0.649	0.668	0.649	0.659	1.20	1.80	1.5
	CICH(ours)	**0.708**	**0.739**	**0.724**	**0.723**	**0.744**	**0.734**	**0.739**	**0.764**	**0.752**	3.10	2.50	2.8
NUS-10K	DCH	0.507	0.585	0.546	0.531	0.599	0.565	0.553	0.598	0.576	4.60	1.30	2.95
	JIMFH	0.132	0.205	0.169	0.154	0.207	0.181	0.154	0.213	0.184	2.20	0.80	1.50
	SCRATCH	0.607	0.571	0.589	0.606	0.581	0.594	0.619	0.584	0.602	1.20	1.30	1.25
	DCMH	0.605	0.521	0.563	0.614	0.541	0.578	0.622	0.545	0.584	1.70	2.40	2.05
	SSAH	0.514	0.496	0.505	0.563	0.505	0.534	0.579	0.555	0.567	6.50	5.90	6.20
	AGAH	0.304	0.356	0.330	0.470	0.413	0.442	0.616	0.574	0.595	31.20	21.80	26.50
	DADH	0.569	0.588	0.579	0.598	0.594	0.596	0.610	**0.613**	0.612	4.10	2.50	3.30
	MCCN	0.563	0.537	0.550	0.588	0.550	0.569	0.587	0.555	0.571	2.40	1.80	2.10
	DCHMT	0.625	0.542	0.584	**0.649**	0.553	0.601	**0.671**	0.560	0.616	4.60	1.80	3.20
	CICH(ours)	**0.628**	**0.602**	**0.615**	0.644	**0.604**	**0.624**	0.639	0.607	**0.623**	**1.10**	**0.50**	**0.80**

The best performances are labeled in boldface. Δ measures the performance drop between the easy and hard settings. A lower Δ implies better generalization to incompleteness. The best and the second-best results are bold and underlined, respectively.

and NUS-WIDE-10K is demonstrated in Table II, and more comparison results on the IAPR TC-12 and NUS-WIDE datasets are summarized in Table III.

In general, our proposed CICH outperforms all the shallow methods as well as all deep hashing methods by a large margin on the MIRFLICKR-25K and NUS-WIDE-10K datasets *w.r.t.* different levels of incompleteness and the Δ metric. It is even superior most of the time to the state-of-the-art real-valued cross-modal retrieval methods, which can learn superior representations. This validates the superior learning ability and

TABLE III
MORE MAP ACCURACY COMPARISONS ON THE IAPR TC-12 AND NUS-WIDE DATASETS *W.R.T.* DIFFERENT LEVELS OF COMPLETENESS

dataset	methods	hard	medium	easy	Δ (easy - hard)(%)
IAPR TC-12	DCMH	0.612	0.610	0.619	0.76
	SSAH	0.542	0.579	0.579	3.68
	AGAH	0.457	0.465	0.482	2.59
	DADH	0.597	0.600	0.605	0.80
	DCHMT	0.585	0.600	0.603	1.80
	CICH(ours)	**0.643**	**0.644**	**0.645**	**0.20**
NUS-WIDE	DCMH	0.755	0.768	0.769	1.40
	SSAH	0.684	0.701	0.690	0.63
	AGAH	0.770	0.771	0.796	2.52
	DADH	0.735	0.756	0.770	3.56
	CICH(ours)	**0.798**	**0.799**	**0.801**	**0.30**

The bold numbers indicate the best results.

TABLE IV
THE AVERAGE $i \rightarrow t$ AND $t \rightarrow i$ MAP COMPARISONS ON THE MIRFLICKR-25K DATASETS AT 16-BIT AND 64-BIT HASH CODE LENGTHS *W.R.T.* DIFFERENT LEVELS OF COMPLETENESS

Methods	16 bits			64 bits		
	hard	medium	easy	hard	medium	easy
DCMH	0.850	0.852	0.852	0.848	0.853	0.860
SSAH	0.860	0.861	0.866	0.872	0.884	0.890
AGAH	0.687	0.690	0.763	0.821	0.830	0.849
DADH	0.839	0.843	0.852	0.854	0.865	0.868
DCHMT	0.828	0.829	0.831	0.840	0.838	0.842
CICH(ours)	**0.862**	**0.865**	**0.866**	**0.889**	**0.890**	**0.891**

The best performances are labeled in boldface.

incompleteness resistance of our method, while also verifying the insufficiency of current ICMH methods. On the MS COCO dataset, although the Δ metric is moderate, our proposed CICH significantly outperforms all compared methods in mAP values, which verifies our hamming ranking ability. By comparing non-deep and deep methods, we observe a clear mAP and Δs gap between traditional CMH and deep hashing methods, indicating the vulnerability of shallow methods with low-paired data. By comparison with the representative deep CMH methods DCMH, SSAH, and DCHMT, our method generally shows better mAP values, confirming our explicit learning ability on the incomplete data to preserve reconstructed semantic similarity, while jointly utilizing the recovered samples to eliminate the modal discrepancy. Compared with the modal-adversarial methods ADAH and DADH, our method generally achieves both better mAP and Δ values, verifying the merits of sample-level correspondence alignment. In contrast with the real-valued incomplete cross-modal retrieval methods PAN and MCCN, our method can achieve superior mAP results on the three datasets. This demonstrates the superior representation learning of our CICH owing to our semantic-balanced contrast of complete and incomplete samples as well as our correspondence recovery strategy. Generally, on three benchmark datasets, our method can achieve remarkable performance most of the time, especially on the two multi-label datasets MIRFLICKR-25K and MS COCO, which validate our high-quality semantic similarity reconstruction in the multi-label context. Therefore, these experiments verify that our proposed CICH can rescue the collapsed similarity and recover the uncertain correspondence in ICMH.

Additionally, we observe that all methods suffer from a drop in results (see Δ results) when increasing the difficulty to harder incomplete levels. In general, the mAP results of incomplete-agnostic methods (SSAH, AGAH, etc.) drop severely in some cases when the missing ratio rises, while the values of incomplete cross-modal retrieval methods MCCN and PAN decrease moderately. In contrast, our method sees only a marginal drop on MIRFLICKR-25K and NUS-WIDE-10K and a moderate drop on MS COCO and is thus impressively more consistent across all levels. High Δ results of compared methods demonstrate that current CMH methods may collapse in similarity and correspondence learning because they are prone to missing samples, but our CICH can remedy these deficiencies consistently.

In addition, we also evaluate the stability of the proposed method of using different lengths of hash codes. Table IV summarizes the comparison results under 16-bit and 64-bit settings to give thorough validations across various bit numbers.

Moreover, we also perform evaluations with other metrics to verify the effective hash lookup ability of our method. We adopt PR-curves *w.r.t.* different incomplete levels on MIRFLICKR-25K and MS COCO. All results of deep cross-modal retrieval methods are compared in different sub-figures of Fig. 3. It can be observed that our method performs the best in most cases. Furthermore, we plot the NDCG@top100 histograms of all deep CMH methods in Fig. 4. All the best results on all missing levels of the two datasets are achieved by our proposed CICH. These results once more exhibit the superior semantic similarity learning and discriminative hash code learning of CICH.

2) Incomplete Evaluation Settings: To better validate the enhancement ability (with more incomplete samples) of our method and explore different incomplete settings considering training and testing phases, we propose to evaluate deep CMH methods with: *Discard* which discards unpaired samples, *Enhance* which utilizes the excess data for enhancement during training, and *Extend* which includes incomplete samples during both training and evaluation. The results on the three levels of incompleteness of MIRFLICKR-25K are presented in Table VI. From the results, we can observe that: (1) Compared between the two settings, the general DCMH method gains fewer improvements and even encounters drops in the 'Enhance' setting, showing that they are ineffective in utilizing the unpaired samples in a progressive manner; (2) Compared to others in the 'Enhance' setting, our method consistently achieves superior results, validating the effectiveness of our CICH with incrementally introduced incomplete samples.

3) Ablation Study: To assess the effectiveness of each module, we conducted an ablation study on variants of CICH by considering different available components. The variant settings and ablation results are presented in Table V. The 3 modules are expected to incrementally boost the performance. From the results, notable findings include that: (1) the PSSC demonstrates significantly better performance than common pairwise hashing, validating its global similarity preservation ability; (2) the PSSC+CCA variant further enhances the model performance, showing the CCA's information transfer ability between modalities to remedy the unclear pair correspondence; and (3) the proposed CICH performs the best among variants, confirming the effectiveness of SaCH in building robust contrastive pairs for modal similarity learning with both present and recovered

TABLE V
THE ABLATION STUDY OF OUR METHOD ON THREE DATASETS

Datasets	Variants	hard		medium		easy	
		i→ t	t→ i	i→ t	t→ i	i→ t	t→ i
MIRFLICKR-25K	Pair	0.743	0.769	0.818	0.807	0.862	0.824
	PSSC	0.896	0.831	0.899	0.833	0.901	0.837
	PSSC+CCA	0.907	0.842	0.906	0.843	0.915	0.845
	CICH	**0.917**	**0.845**	**0.918**	**0.845**	**0.924**	**0.847**
MS COCO	Pair	0.552	0.562	0.573	0.582	0.604	0.587
	PSSC	0.704	0.737	0.707	0.742	0.703	0.744
	PSSC+CCA	0.706	0.736	0.714	0.742	0.712	0.744
	CICH	**0.708**	**0.739**	**0.723**	**0.744**	**0.739**	**0.764**
NUS-10K	Pair	0.321	0.288	0.459	0.424	0.557	0.482
	PSSC	0.620	0.597	0.638	0.601	0.635	0.603
	PSSC+CCA	0.627	0.600	0.642	0.601	0.639	0.604
	CICH	**0.628**	**0.602**	**0.644**	**0.604**	**0.639**	**0.607**

"pair": pairwise hashing loss. "PSSC": prototypical semantic similarity coordination. "SaCH": semanticaware contrastive hashing. "CCA": contextual correspondence alignment. "CICH"="PSSC"+"SaCH"+"CCA" is the proposed method.

TABLE VI
THE MAP COMPARISON ON MIRFLICKR-25K *W.R.T.* OUR PROPOSED EVALUATION SETTINGS REGARDING INCOMPLETENESS

Methods	hard		medium		easy	
	i→ t	t→ i	i→ t	t→ i	i→ t	t→ i
DCMH(1)	0.903	0.799	0.905	0.818	0.900	0.821
DCMH(2)	0.889	0.823	0.897	0.820	0.891	0.823
SSAH(2)	0.906	0.836	0.910	0.836	0.913	0.844
DADH(2)	0.869	0.829	0.882	0.833	0.878	0.833
AGAH(2)	0.550	0.627	0.861	0.830	0.879	0.832
DCHMT(2)	0.855	0.798	0.854	0.800	0.861	0.805
CICH (ours)(1)	0.917	0.813	0.917	0.840	0.918	0.836
CICH (ours)(2)	**0.918**	**0.840**	**0.921**	**0.845**	**0.919**	**0.847**

(1): 'Discard'; (2): 'Enhance'.

TABLE VII
THE AVERAGE OF I → T AND T → I MAP VALUES *W.R.T.* DIFFERENT VALUES OF α AND β ON THE MIRFLICKR-25K DATASET

Parameters		β				
		0.0001	0.001	0.01	0.10	1.00
α	0.1	0.874	0.873	0.872	0.874	0.873
	0.5	0.881	0.881	0.883	0.880	0.879
	1.0	0.885	0.884	0.884	0.881	0.881
	5.0	**0.889**	0.888	0.888	0.887	0.886
	10.0	0.888	0.888	0.888	0.887	0.886

samples. The PSSC and the CCA serve as components to rebuild an initial feasible incomplete model through semantic similarity learning and sample-level correspondences, while the SaCH provides the model with an augmented contrastive relationship and learns discriminative hash codes. Together, these components work cooperatively to rescue the collapsed similarity learning and yield correspondence-consistent binary codes.

4) Parameter Analysis and Time Efficiency: To demonstrate the sensitivity of parameters, we conduct performance tests for some parameters. There are five parameters to be searched and analyzed, namely α, β, τ, K, and δ. Specifically, we evaluate the performance with the values of α and β combined in Table VII, and we search τ, K, and δ independently in Fig. 5. The mAP results on the easy setting are illustrated in Fig. 5. Based on the results, we can observe that the appropriate range for τ lies in $[0.25, 5.0]$, while it needs K in the range of $[5, 15]$ to reach higher mAPs. In a range of relatively greater values of about $[20, 500]$, δ yields the best performance. We also observe that a greater

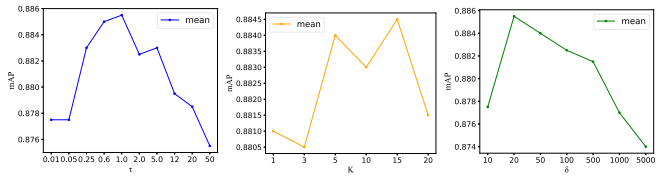


Fig. 5. The average of $i \rightarrow t$ and $t \rightarrow i$ mAP values *w.r.t.*. different values of parameters on the MIRFLICKR-25K dataset.

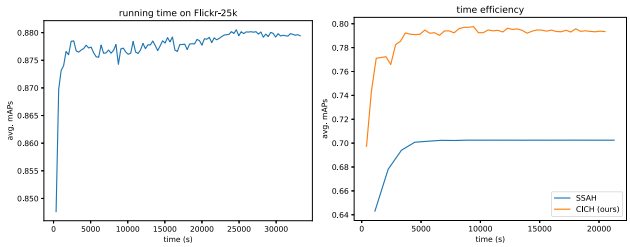


Fig. 6. The running time (left) and time efficiency (right) analysis of CICH.

$\alpha = 5.0$ with a smaller $\beta = 1 \times 10^{-4}$ can perform optimally. These results indicate that: (1) within the reasonable range of each parameter, our CICH can achieve stable and high mAP results, (2) the increased role of SaCH can facilitate similarity recovery, and (3) lower constraints on the latent $\boldsymbol{z}$ result in better semantic preservation. Generally, we can conclude that our CICH can perform preferably within a reasonable range of parameter settings.

For validating time efficiency, we plot the mAP performance dynamics *w.r.t.* training time in Fig. 6. Although CICH trained with a global PSSC module is expected to be more time-consuming, we found that CICH actually reaches the vicinity of the optimal value at early epochs. This validates that our method efficiently preserves the global similarities to recover cross-modal learning for ICMH.

5) Visualization: For further validation of our proposed CICH in discriminative hash code learning, we perform a t-SNE visualization on three baseline CMH methods and our CICH. The results are shown in Fig. 7, where the number of data points in each subfigure is the same. Image features are tagged by a dot ('.') and text features are labeled by a plus mark ('+'). Different colors represent different classes. We expect the model to learn class-discriminative and modality-aligned codes. From the results, we can observe that: (1) image and text features of the same class and the same modality often occur and overlap together in our result, while this is not true for compared methods (especially AGAH), which validates CICH's superiority in rebuilding modal similarity learning and fixing the corrupted correspondence, and (2) groups of the same class can form compact clusters into almost hot spots in our result, while those of the other methods (especially DCMH and DADH) are not concentrated and can only form dispersed areas (note their equal numbers of visualized data points), which verifies CICH's capability in semantic similarity learning.

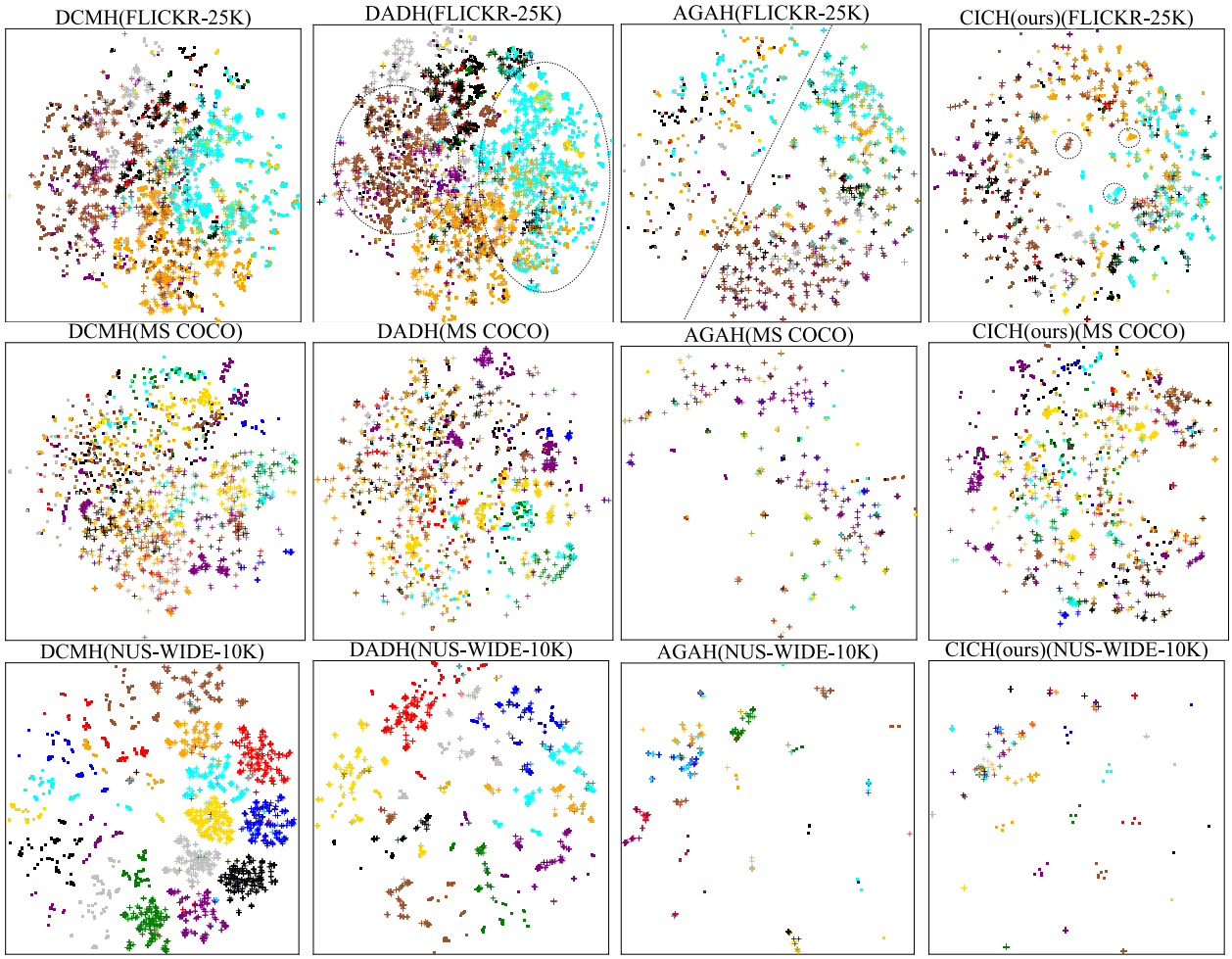


Fig. 7. The t-SNE visualization of three baseline methods and our proposed CICH on three datasets with the medium setting. The number of data points in each subfigure is the same.

V. Conclusion

In this work, we identified the challenge of semi-observed training data in deep cross-modal hashing, dubbed incomplete cross-modal hashing, including the issues of collapsed semantic and modal similarity learning, as well as uncertain cross-modal correspondence. To overcome these challenges, we proposed a novel Contrastive Incomplete Cross-modal Hashing (CICH) method that integrates cross-modal semantic coordination, imbalanced similarity calibration, and contextual correspondence alignment into one framework. Notably, this is the first work that enables contrastive learning for incomplete CMH. Specifically, the prototypical semantic similarity coordination module and semantic-aware contrastive hashing module enable CICH to capture global and explicit semantic relationships between modalities to generate discriminative hash codes, respectively. Moreover, the proposed contextual correspondence alignment module learns optimal shared knowledge across modalities to perform feature recovery and eliminate correspondence uncertainty. Extensive experiments on multiple benchmark datasets demonstrated the superiority of CICH compared to state-of-the-art methods across various settings.

References

[1] Y. Wang, X. Luo, L. Nie, J. Song, W. Zhang, and X.-S. Xu, "Batch: A scalable asymmetric discrete cross-modal hashing," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 11, pp. 3507–3519, Nov. 2021.

[2] M. Wang, W. Fu, S. Hao, H. Liu, and X. Wu, "Learning on big graph: Label inference and regularization with anchor hierarchy," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 5, pp. 1101–1114, May 2017.

[3] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.

[4] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3232–3240.

[5] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4242–4251.

[6] H. Wu et al., "Contrastive label correlation enhanced unified hashing encoder for cross-modal retrieval," in *Proc. 31st ACM Int. Conf. Inf. Knowl. Manage.*, 2022, pp. 2158–2168.

[7] M. Jing, J. Li, L. Zhu, K. Lu, Y. Yang, and Z. Huang, "Incomplete cross-modal retrieval with dual-aligned variational autoencoders," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 3283–3291.

[8] Z. Zeng, Y. Sun, and W. Mao, "MCCN: Multimodal coordinated clustering network for large-scale cross-modal retrieval," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 5427–5435.

[9] Z. Qiu, Q. Su, Z. Ou, J. Yu, and C. Chen, "Unsupervised hashing with contrastive information bottleneck," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 959–965.

[10] X. Luo et al., "Cimon: Towards high-quality hash codes," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 902–908.
[11] X. Xia, G. Dong, F. Li, L. Zhu, and X. Ying, "When clip meets cross-modal hashing retrieval: A new strong baseline," *Inf. Fusion*, vol. 100, 2023, Art. no. 101968.
[12] Z.-D. Chen, C.-X. Li, X. Luo, L. Nie, W. Zhang, and X.-S. Xu, "SCRATCH: A scalable discrete matrix factorization hashing framework for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 2262–2275, Jul. 2020.
[13] C. Bai, C. Zeng, Q. Ma, J. Zhang, and S. Chen, "Deep adversarial discrete hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retrieval*, 2020, pp. 525–531.
[14] E. Yu, J. Ma, J. Sun, X. Chang, H. Zhang, and A. G. Hauptmann, "Deep discrete cross-modal hashing with multiple supervision," *Neurocomputing*, vol. 486, pp. 215–224, 2022.
[15] H. T. Shen et al., "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 10, pp. 3351–3365, Oct. 2021.
[16] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3864–3872.
[17] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 1735–1742.
[18] R. D. Hjelm et al., "Learning deep representations by mutual information estimation and maximization," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–14.
[19] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9729–9738.
[20] J.-B. Grill et al., "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.
[21] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, PMLR, 2021, pp. 8748–8763.
[22] H. Lu, N. Fei, Y. Huo, Y. Gao, Z. Lu, and J.-R. Wen, "COTS: Collaborative two-stream vision-language pre-training model for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15692–15701.
[23] J. Guo and W. Zhu, "Collective affinity learning for partial cross-modal hashing," *IEEE Trans. Image Process.*, vol. 29, pp. 1344–1355, 2019.
[24] L. Wu, Y. Wang, and L. Shao, "Cycle-consistent deep generative hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1602–1612, Apr. 2019.
[25] Z. Zeng, S. Wang, N. Xu, and W. Mao, "PAN: Prototype-based adaptive network for robust cross-modal retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1125–1134.
[26] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun., Control, Comput.*, 1999, pp. 368–377.
[27] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, "Deep variational information bottleneck," in *Proc. Int. Conf. Learn. Representations*, 2017, pp. 1–16.
[28] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–9.
[29] M. J. Huiskes and M. S. Lew, "The MIR flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retrieval*, 2008, pp. 39–43.
[30] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2014, pp. 740–755.
[31] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from national university of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, Art. no. 48.
[32] H. J. Escalante et al., "The segmented and annotated IAPR TC-12 benchmark," *Comput. Vis. Image Understanding*, vol. 114, no. 4, pp. 419–428, 2010.
[33] D. Wang, Q. Wang, L. He, X. Gao, and Y. Tian, "Joint and individual matrix factorization hashing for large-scale cross-modal retrieval," *Pattern Recognit.*, vol. 107, 2020, Art. no. 107479.
[34] W. Gu, X. Gu, J. Gu, B. Li, Z. Xiong, and W. Wang, "Adversary guided asymmetric hashing for cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retrieval*, 2019, pp. 159–167.
[35] J. Tu, X. Liu, Z. Lin, R. Hong, and M. Wang, "Differentiable cross-modal hashing via multimodal transformers," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 453–461.

Haoyang Luo received the BS degree in computer science and technology from the Harbin Institute of Technology, Shenzhen, in 2021. He is currently working toward the master's degree with the Harbin Institute of Technology, Shenzhen, China. His current research interests include multi-modal learning and vision-language foundation models.

Zheng Zhang received the PhD degree from the Harbin Institute of Technology, China. He was a postdoctoral research fellow with The University of Queensland, Australia. He is currently with the Harbin Institute of Technology, Shenzhen, China. He has co-authored more than 100 technical papers in prestigious international journals and conferences. His research interests include multimedia content analysis and understanding. He is an AE of *IEEE Transactions on Affective Computing*, *IEEE Journal of Biomedical and Health Informatics*, and others and also serves as an area chair of ICML, NeurIPS, CVPR, ACM MM, etc.

Liqiang Nie (Senior Member, IEEE) received the BEng degree from Xi'an Jiaotong University and the PhD degree from the National University of Singapore (NUS). After the PhD degree, he continued his research with NUS, as a research fellow for three years. He is currently a professor with the Harbin Institute of Technology (Shenzhen). He has co-authored more than 200 articles and four books. His research interests include multimedia computing and information retrieval. He is an associate editor of *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Multimedia*, *ACM Transactions on Multimedia Computing, Communications, and Applications*, and *Information Sciences*.