Semi-Supervised Online Cross-Modal Hashing

Xiao Kang [1], Xingbo Liu[2 *], Xuening Zhang [3], Wen Xue [1], Xiushan Nie[2, 4], Yilong Yin [1]

[1]School of Software, Shandong University, Jinan 250101, China
[2]School of Computer Science and Technology, Shandong Jianzhu University, Jinan 250101, China
[3]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen 518055, China
[4]Shandong Yunhai Guochuang Cloud Computing Equipment Industry Innovation Co., Ltd, Jinan, China
{sckx, sclxb, ctxw}@mail.sdu.edu.cn, yukizhang0527@outlook.com, niexiushan@163.com, ylyin@sdu.edu.cn

Abstract

Online cross-modal hashing has gained increasing interest due to its ability to encode streaming data and update hash functions simultaneously. Existing online methods often assume either fully supervised or completely unsupervised settings. However, they overlook the prevalent and challenging scenario of semi-supervised cross-modal streaming data, where diverse data types, including labeled/unlabeled, paired/unpaired, and multi-modal, are intertwined. To address this issue, we propose Semi-Supervised Online Cross-modal Hashing (SSOCH). It presents an alignment-free pseudo-labeling strategy that extracts semantic information from unlabeled streaming data without relying on pairing relations. Furthermore, we design an online tri-consistent preserving scheme, integrating pseudo-labeled data regularization, discriminative label embedding, and fine-grained similarity preservation. This scheme fully explores consistency across data annotation, modalities, and streaming chunks, improving the model's adaptiveness in these challenging scenarios. Extensive experiments on benchmark datasets demonstrate the superiority of SSOCH under various scenarios, highlighting the importance of semi-supervised learning for online cross-modal hashing.

Introduction

With the continuous accumulation and explosive growth of multimedia data, cross-modal hashing has received extensive attention in various domains, such as image-text retrieval, video-audio synchronization, and personalized multimedia recommendation systems. Despite the inspiring performance, traditional batch-based cross-modal hashing methods have struggled to adapt to the dynamic and timely retrieval of streaming data (Sun et al. 2023; Sun, Peng, and Ren 2024). In response to this challenge, online cross-modal hashing has emerged (Liu et al. 2024). It aims to generate hash codes on time for the continuously growing streaming data and dynamically update the hash functions to accommodate the evolving data characteristics (Zhu et al. 2020).

Existing online cross-modal hashing methods can be broadly categorized into the supervised (Wang, Luo, and Xu 2020; Zhan et al. 2021; Kang et al. 2024b) and the unsupervised methods (Xie, Shen, and Zhu 2016; Liu et al. 2022a; Li et al. 2022). Supervised methods focus on enhancing and leveraging labels to capture the cross-modal consistency in streaming data, such as Online Latent Semantic Hashing (OLSH) (Yao et al. 2019) and Online Discriminative Cross-modal Hashing (ODCH) (Kang et al. 2023b). They attempt to learn semantic-preserving embeddings under the guidance of designed semantic labels, measuring the similarity of multi-modal streaming data more meticulously. Unsupervised methods primarily lay emphasis on decomposing and reconstructing feature representations to explore the geometric structure of streaming data. For example, Online Manifold-Guided Hashing (OMGH) (Liu et al. 2022a) designs a matrix tri-factorization framework to decompose high-dimensional features into modality-specific representations and hash codes. Dynamic Prototype Online Cross-modal Hashing (DPOCH) (Kang et al. 2024a) generates prototypes incrementally as sketches of accumulated data and updates them dynamically for adapting streaming data.

Despite promising advancements, existing online cross-modal hashing methods typically assume either a fully supervised or completely unsupervised scenario. These methods have overlooked a crucial fact: Real-world scenarios are constantly semi-supervised streaming data scenarios, *i.e.* data arrives in a streaming fashion, and each data chunk contains a small number of labeled instance pairs as well as a large number of unlabeled and unpaired instances (Zhang, Peng, and Yuan 2020). The complexity of such a scenario, including data distribution variations across chunks, heterogeneous modalities, and the lack of supervision and pairing information, poses severe challenges for learning hash codes and updating hash functions online. To the best of our knowledge, there are no hashing methods specifically designed to handle semi-supervised online cross-modal retrieval scenarios.

There are several batch-based semi-supervised cross-modal hashing methods, such as Weakly-supervised Cross-modal Hashing (WCHash) (Liu et al. 2022b), Semi-Supervised Semantic-Preserving Hashing (S3PH) (Wang et al. 2019), and Three-Stage Semi-Supervised Hashing (TS3H) (Fan et al. 2023). A convincing strategy of these methods is to complete the semantic label and utilize the pseudo-labels to discover the similar relationship among multi-modal data. Despite encouraging retrieval performance, it is not well-suited for adapting to more complex streaming data scenarios due to the following reasons: 1) The pseudo-

[*]Corresponding author

labeling strategy in these methods naturally relies on the paired information between different modalities to complement missing labels, which cannot adapt to the scenarios where the pairing information is unavailable (Zhang et al. 2024; Yu et al. 2023). 2) The supervised labels and generated pseudo-labels are generally in the form of one-hot or multi-hot encoding, which holds coarse separability and similarity relationships. Such labels may lead to the imbalance updating problem (Lin et al. 2019; Lu et al. 2019; Zhu et al. 2023) during online updates. 3) These methods typically rely on the complete dataset to explore the sample distribution patterns, lacking the ability to adapt to the distribution changes among data chunks in the streaming scenarios.

To address the above problems, we propose a Semi-Supervised Online Cross-modal Hashing (SSOCH) method. It integrates the semantic information from complex and diverse data into a unified framework. Specifically, we first design an alignment-free pseudo-labeling strategy. It exploits the feature-label associations from the intra-modality labeled instances and transfers them to the unlabeled data, thereby generating pseudo-labels without pair information. The generated pseudo-labels and the given supervised labels are then enhanced by Hadamard embedding to amplify their discriminability and separability. Furthermore, we seamlessly integrate the discriminative label/pseudo-label embedding and fine-grained similarity-preserving, achieving tri-consistent semantic preservation. Besides, we present theoretical analyses of the proposed method, including stability, generalization, convergence, and time complexity analysis. The main contribution of this study can be concluded as follows,

- We propose an online tri-consistent preserving framework that integrates multiple constraints to explore consistency across labeled/unlabeled data, modalities, and data chunks. To the best of our knowledge, it is the first attempt to address the complex semi-labeled semi-paired online cross-modal retrieval scenarios.
- An alignment-free pseudo-labeling strategy is presented to extract semantic information from unlabeled streaming data without relying on pairing relations. Besides, label discrimination enhancement is designed to alleviate the imbalance updating problem.
- Meticulous theoretical analyses and extensive experiments conducted on three widely used datasets verify the superiority of the proposed method.

The Proposed Method

In this section, we introduce notations and elaborate SSOCH from three aspects, *i.e.*, model formulation, optimization process, and theoretical analyses.

Notations and Problem Statement

Assume that we have a training set $\mathbf{X}$ composed of labeled multi-modality instance pairs $\mathbf{X}_{s,m} \in \mathbb{R}^{d_m \times n_s}$ with corresponding semantic labels $\mathbf{Y}_s \in \mathbb{R}^{c \times n_s}$ and unlabeled multi-modality instance $\mathbf{X}_{u,m} \in \mathbb{R}^{d_m \times n_{u,m}}$. d_m means dimension of m_{th} modality features. n_s and $n_{u,m}$ represent the number of labeled instance pairs and m_{th} modality unlabeled instances, respectively. Moreover, $sgn(\cdot)$ denotes sign function, $||\mathbf{X}||$ and $\mathbf{X}^T$ denote the ℓ_2-norm and transposition of matrix $\mathbf{X}$, respectively.

In the semi-supervised online retrieval scenario, data come in a streaming fashion, represented as $\mathbf{X}_{s,m}^{(t)} \in \mathbb{R}^{d_m \times n_{s,t}}$ with corresponding semantic labels $\mathbf{Y}_s^{(t)} \in \mathbb{R}^{c \times n_{s,t}}$ and $\mathbf{X}_{u,m}^{(t)} \in \mathbb{R}^{d_m \times n_{u,m,t}}$, where $n_{s,t}$ and $n_{u,m,t}$ denote the size of data at t-chunk. Correspondingly, $\mathbf{X}_{s,m}^{(t-1)} \in \mathbb{R}^{d_m \times N_{s,t}}$ and $\mathbf{X}_{u,m}^{(t-1)} \in \mathbb{R}^{d_m \times N_{u,m,t}}$ represents the feature of accumulated data before t-chunk, where $N_{s,t}$ and $N_{u,m,t}$ denote the size of accumulated data before t-chunk. The proposed SSOCH method aims to learn compact hash codes $\mathbf{B}^{(t)} \in \{-1, +1\}^{r \times n_{s,t}}$ for the semi-supervised instances of each chunk.

Model Formulation

As illustrated in Figure 1, we propose a semi-supervised online cross-modal hashing method. It comprises three key components: 1) Alignment-free pseudo-labeling, which enriches the label information for each modality separately without depending on cross-modal correspondence; 2) Discriminative Hadamard embedding, which enhances the discriminability and separability of both supervised labels and pseudo-labels; and 3) Online tri-consistent preserving scheme, which integrates multiple semantic information to guide the learning of hash codes and the updating of hash functions.

Alignment-free Pseudo-labeling Most existing pseudo-labeling methods (Liu et al. 2022b; Fan et al. 2023) rely on fully paired data to capture semantic relations. Such a training mechanism lacks flexibility under the complex streaming data scenario where different modality instances arrive dynamically and the paired relations are incomplete. To this end, we propose an alignment-free pseudo-labeling strategy. Its core idea is to narrow the semantic gap between the common label and each modality feature individually, thereby achieving the goal of pulling features of different modalities to the shared semantic space. Specifically, for the m_{th} modality, we first learn the feature-label association $\mathbf{M}_m$ from the labeled instances $\mathbf{X}_{s,m}^{(t)}$, and then transfer this association to the unlabeled instances. Considering the joint optimization between the newly arriving and the accumulated data, the learning process of the projection $\mathbf{M}_m$ can be formulated as,

$$\begin{aligned}\min_{\mathbf{M}_m} & \left\|\mathbf{Y}_s^{(t)} - \mathbf{M}_m \mathbf{X}_{s,m}^{(t)}\right\|^2 + \left\|\mathbf{Y}_s^{(t-1)} - \mathbf{M}_m \mathbf{X}_{s,m}^{(t-1)}\right\|^2 \\ & + \delta \left\|\mathbf{M}_m\right\|^2 ,\end{aligned} \tag{1}$$

where δ is a hyperparameter. By optimizing the Eq. (1), $\mathbf{M}_m$ can be computed as,

$$\mathbf{M}_m = \mathbf{C1}^{(t)} (\mathbf{C2}^{(t)} + \delta \mathbf{I})^{-1}, \tag{2}$$

where

$$\begin{aligned}\mathbf{C1}^{(t)} = \mathbf{C1}^{(t-1)} + \mathbf{Y}_s^{(t)} \mathbf{X}_{s,m}^{(t)T}, \mathbf{C1}^{(t-1)} = \mathbf{Y}_s^{(t-1)} \mathbf{X}_{s,m}^{(t-1)T}, \\ \mathbf{C2}^{(t)} = \mathbf{C2}^{(t-1)} + \mathbf{X}_{s,m}^{(t)} \mathbf{X}_{s,m}^{(t)T}, \mathbf{C2}^{(t-1)} = \mathbf{X}_{s,m}^{(t-1)} \mathbf{X}_{s,m}^{(t-1)T}.\end{aligned} \tag{3}$$

After obtaining the projections $\mathbf{M}_m$, we construct the k-hot pseudo-labels of unlabeled instances using $\mathbf{Y}_{u,m}^{(t)} =$

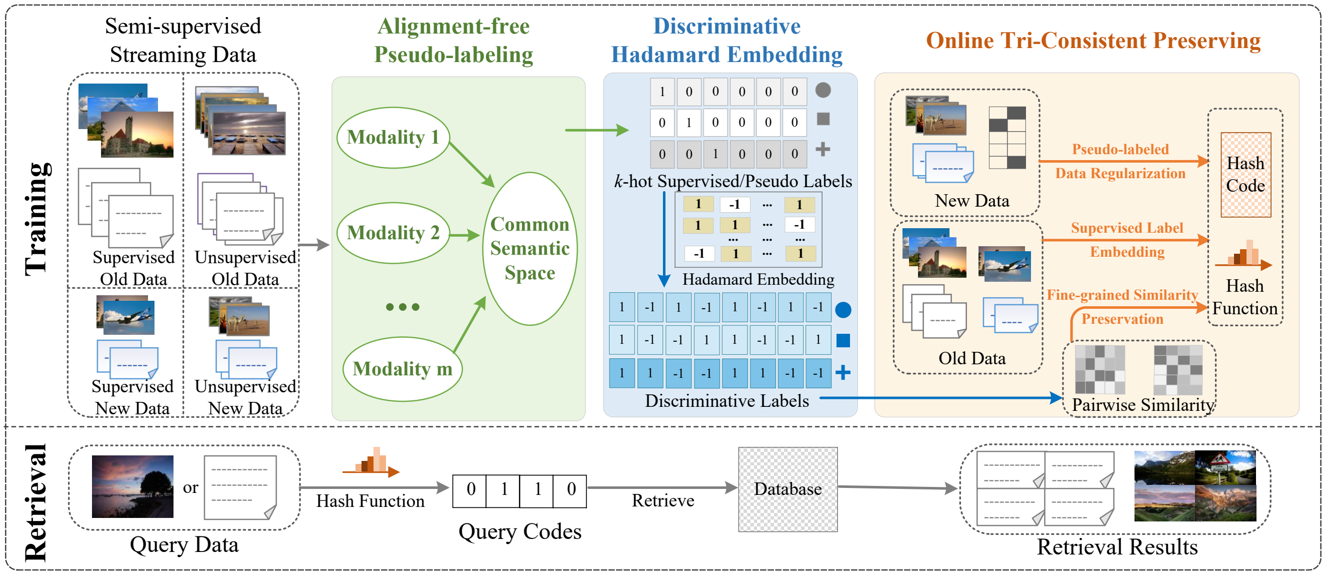


Figure 1: Framework of SSOCH, illustrated with toy data. When the new semi-supervised streaming data comes, we first learn the k-hot labels using the alignment-free pseudo-labeling strategy. Then Hadamard embedding is introduced to enhance the discrimination of labels and construct a fine-grained similarity graph. Finally, we design an online tri-consistent preserving framework to explore the semantic consistency across labeled/unlabeled data, modalities, and streaming chunks.

$top_k(\mathbf{M}_m\mathbf{X}^{(t)}_{u,m})$, where $top_k(\cdot)$ means setting the k largest values in the label vector to 1, while the remaining to 0.

Discriminative Hadamard Embedding Most existing methods (Li et al. 2022; Zhang, Wu, and Chen 2023; Yao et al. 2023) adopt one-hot labels to guide the learning of hash codes and hash functions. However, these labels hold a fixed Hamming distance of 2 among different classes. Guided by such labels, the generated hash codes are susceptible to misclassification (Kang et al. 2023b). To construct more discriminative labels, we introduce the Hadamard matrix $\mathbf{H} \in \{-1, 1\}^{C\times C}$ (Lin et al. 2020). It possesses two desirable properties that make it suitable for guiding the learning of hash codes. 1) *Balance*: Half of the elements are $+1$s, and the other half are -1s (except for the first row or column whose elements are all), thereby expanding the Hamming distance among different categories to $C/2$. 2) *Independence*: the row and column vectors of $\mathbf{H}$ are pairwisely orthogonal, thereby improving the separability among instances. Resorting to $\mathbf{H}$, we first build a codebook $\mathbf{H}' \in \{-1, 1\}^{C\times c}$ by randomly selecting c column vectors from $\mathbf{H}$ (excluding the first column). Then, the discriminative label $\mathbf{L}^{(t)} \in \mathbb{R}^{n_t\times C}$ can be constructed based on this codebook $\mathbf{H}'$. The process can be formulated as follows,

$$\mathbf{l}_i^{(t)} = \sum \mathbf{H}' \cdot diag(\mathbf{y}_i^{(t)}), \tag{4}$$

where $\mathbf{y}_i^{(t)}$ and $\mathbf{l}_i^{(t)}$ mean the original one-hot label and the discriminative label of the i-th instance, respectively. $diag(\cdot)$ represents vector diagonalization, which serves as a selector for codebook $\mathbf{H}'$. Note that 0 element appears in $\mathbf{L}^{(t)}$ in rare cases and it will be re-assigned with +1 or -1 with the balancedness principle (Lin et al. 2020). Through the proposed discriminative Hadamard embedding, the given one-hot labels of supervised instances $\mathbf{Y}_s^{(t)}$ and the learned pseudo-label of unsupervised instances $\mathbf{Y}_{u,m}^{(t)}$ can be transferred to the discriminative label $\mathbf{L}_s^{(t)}$ and $\mathbf{L}_{u,m}^{(t)}$.

Online Tri-consistent Preserving To explore semantic relations under a complex semi-supervised streaming scene, SSOCH proposes an online tri-consistent preserving scheme. The following context will be presented in three parts.

Pseudo-labeled Data Regularization aims to exploit the semantic consistency in unlabeled data. Specifically, two sets of projections are introduced: $\mathbf{G}$ is designed to bridge the gap between the pseudo-labels $\mathbf{L}_{u,m}^{(t)}$ and the hash codes $\mathbf{B}_u^{(t)}$. $\mathbf{W}_m$ is devised to connect the heterogeneous modalities $\mathbf{X}_{u,m}^{(t)}$ and $\mathbf{B}_u^{(t)}$. The process can be formulated as,

$$\begin{aligned} \min_{\mathbf{G},\mathbf{W}_m,\mathbf{B}_u^{(t)}} & \left\|\mathbf{B}_u^{(t)} - \mathbf{G}\mathbf{L}_{u,m}^{(t)}\right\|^2 + \left\|\mathbf{B}_u^{(t)} - \mathbf{W}_m\mathbf{X}_{u,m}^{(t)}\right\|^2, \\ \text{s.t. } & \mathbf{B}_u^{(t)} \in \{-1, +1\}^{r\times n_{u,m,t}}. \end{aligned} \tag{5}$$

To handle the situation of alignment missing, *i.e.* paired relations incomplete, we explore the semantic information of each modality separately instead of learning the shared codes $\mathbf{B}_u^{(t)}$. Eq. (5) can be rewritten as,

$$\min_{\mathbf{G},\mathbf{W}_m} \sum_{m=1}^{M} \theta_m \left\|\mathbf{G}\mathbf{L}_{u,m}^{(t)} - \mathbf{W}_m\mathbf{X}_{u,m}^{(t)}\right\|^2. \tag{6}$$

Note that, only the hash codes corresponding to the supervised data are stored in the retrieval library, thus improving the quality of hash codes and enhancing retrieval accuracy.

Discriminative Label Embedding aims to exploit the semantic consistency in labeled data. It adopts a simple yet effective linear regression to bridge the gap between the discriminative supervised label $\mathbf{L}_s^{(t)}$ and the hash code $\mathbf{B}^{(t)}$. Moreover, to smooth the solution and improve the stability, we implement ℓ_2-norm regularization on $\mathbf{G}$. The process can be formulated as follows,

$$\min_{\mathbf{G},\mathbf{B}^{(t)}} (\left\|\mathbf{B}^{(t-1)} - \mathbf{G}\mathbf{L}_s^{(t-1)}\right\|^2 + \left\|\mathbf{B}^{(t)} - \mathbf{G}\mathbf{L}_s^{(t)}\right\|^2) \\ + \gamma\|\mathbf{G}\|^2, \quad \text{s.t. } \mathbf{B}^{(t)} \in \{-1,+1\}^{r\times n_{s,t}}. \tag{7}$$

Fine-grained Similarity Preservation focuses on exploiting the consistency among newly arriving data and accumulated data. To avoid the imbalanced updating problem (Lin et al. 2019), we design a fine-grained asymmetric similarity matrix $\mathbf{S}_{oc}$, which is constructed based on the discriminative Hadamard labels $\mathbf{L}_s^{(t)}$. $\mathbf{S}_{oc}$ can be defined as follows,

$$\mathbf{S}_{oc} = 2\mathbf{U}^{(t-1)T}\mathbf{U}^{(t)} - \mathbf{1}\mathbf{1}^T, \tag{8}$$

where $\mathbf{u}_j^{(t)} = \mathbf{l}_j^{(t)}/\|\mathbf{l}_j^{(t)}\|$, $\mathbf{l}_j^{(t)}$ is the discriminative label of j_{th} instances, $\mathbf{1}$ is an all-one column vector and $\mathbf{E}$ is an all-one matrix. Meanwhile, we define $\mathbf{S}_{cc} = 2\mathbf{U}^{(t)T}\mathbf{U}^{(t)} - \mathbf{1}\mathbf{1}^T$ to exploit the similarity in each data chunk.

To preserve the pairwise similarity, we employ the inner product of hash codes to approximate the fine-grained asymmetric similarity matrix $\mathbf{S}_{oc}$. To avoid the challenging binary quadratic problem (Yang 2013), we replace the binary hash codes $\mathbf{B}^{(t)}$ with real-valued representations $\mathbf{V}^{(t)}$ and learn an orthogonal projection $\mathbf{R}$ to map $\mathbf{V}^{(t)}$ into $\mathbf{B}^{(t)}$. Additionally, bit balance and uncorrelation constraints on $\mathbf{V}^{(t)}$ are implemented to promote the accuracy of hash codes. The process can be formulated as,

$$\min_{\mathbf{B}^{(t)},\mathbf{R},\mathbf{V}^{(t)}} \left\|\mathbf{B}^{(t)} - \mathbf{R}\mathbf{V}^{(t)}\right\|^2 + \beta(\left\|\mathbf{V}^{(t)T}\mathbf{B}^{(t)} - r\cdot\mathbf{S}_{cc}\right\|^2 \\ + \left\|\mathbf{V}^{(t)T}\mathbf{B}^{(t-1)} - r\cdot\mathbf{S}_{oc}\right\|^2), \text{ s.t. } \mathbf{V}^{(t)T}\mathbf{V}^{(t)} = n_t\mathbf{I}, \\ \mathbf{V}^{(t)T}\mathbf{1} = \mathbf{0}, \mathbf{R}\mathbf{R}^T = \mathbf{I}, \mathbf{B}^{(t)} \in \{-1,+1\}^{r\times n_{s,t}}. \tag{9}$$

For the out-of-sample extension, we optimize the projections $\mathbf{W}_m$ to mapping the multi-modality features $\mathbf{X}_{s,m}$ into the common hash codes $\mathbf{B}$,

$$\min_{\mathbf{W}_m} \left\|\mathbf{B}^{(t)} - \mathbf{W}_m\mathbf{X}_{s,m}^{(t)}\right\|^2 + \left\|\mathbf{B}^{(t-1)} - \mathbf{W}_m\mathbf{X}_{s,m}^{(t-1)}\right\|^2 + \gamma\|\mathbf{W}_m\|^2. \tag{10}$$

Combining Eq. (6), Eq. (7), Eq. (9), and Eq. (10), the final objective function can be formulated as follows,

$$\min_{\Theta} \left\|\mathbf{B}^{(t)} - \mathbf{R}\mathbf{V}^{(t)}\right\|^2 + \sum_{m=1}^{M}(\theta_m\left\|\mathbf{G}\mathbf{L}_{u,m}^{(t)} - \mathbf{W}_m\mathbf{X}_{u,m}^{(t)}\right\|^2 \\ + \xi_m(\left\|\mathbf{B}^{(t)} - \mathbf{W}_m\mathbf{X}_{s,m}^{(t)}\right\|^2 + \left\|\mathbf{B}^{(t-1)} - \mathbf{W}_m\mathbf{X}_{s,m}^{(t-1)}\right\|^2)) \\ + \alpha(\left\|\mathbf{B}^{(t-1)} - \mathbf{G}\mathbf{L}_s^{(t-1)}\right\|^2 + \left\|\mathbf{B}^{(t)} - \mathbf{G}\mathbf{L}_s^{(t)}\right\|^2) \\ + \beta(\left\|\mathbf{V}^{(t-1)T}\mathbf{B}^{(t)} - r\cdot\mathbf{S}_{oc}\right\|^2 + \left\|\mathbf{V}^{(t)T}\mathbf{B}^{(t)} - r\cdot\mathbf{S}_{cc}\right\|^2), \\ + \gamma(\|\mathbf{G}\|^2 + \sum_{m=1}^{M}\|\mathbf{W}_m\|^2), \text{ s.t. } \mathbf{V}^{(t)T}\mathbf{V}^{(t)} = n_t\mathbf{I}, \\ \mathbf{V}^{(t)T}\mathbf{1} = \mathbf{0}, \mathbf{R}\mathbf{R}^T = \mathbf{I}, \mathbf{B}^{(t)} \in \{-1,+1\}^{r\times n_{s,t}}, \tag{11}$$

where $\Theta = \{\mathbf{B}^{(t)}, \mathbf{R}, \mathbf{V}^{(t)}, \mathbf{G}, \mathbf{W}_m\}$. α, β, ξ_m, θ_m and γ are hyperparameters.

Optimization Process

Directly optimizing the problem formulated in Eq. (11) is challenging, as it is nonconvex and noncontinuous. Fortunately, it can be optimized utilizing an iterative algorithm that solves sub-problems concerning individual variables. The optimization process for each variable is presented as follows.

G-Step: Learn the projection matrix $\mathbf{G}$, holding the other variables fixed. The optimization in Eq. (11) becomes,

$$\min_{\mathbf{G}} \alpha(\left\|\mathbf{B}^{(t)} - \mathbf{G}\mathbf{L}_s^{(t)}\right\|^2 + \left\|\mathbf{B}^{(t-1)} - \mathbf{G}\mathbf{L}_s^{(t-1)}\right\|^2) \\ + \sum_{m=1}^{M}\theta_m\left\|\mathbf{G}\mathbf{L}_{u,m}^{(t)} - \mathbf{W}_m\mathbf{X}_{u,m}^{(t)}\right\|^2 + \gamma\|\mathbf{G}\|^2. \tag{12}$$

The problem in Eq. (12) is a simple linear least squares question (Lawson and Hanson 1995). By setting the derivative with respect to $\mathbf{G}$ as $\mathbf{0}$, it can be computed as,

$$\mathbf{G} = (\sum_{m=1}^{2}\theta_m\mathbf{W}_m\mathbf{X}_{u,m}^{(t)}\mathbf{L}_{u,m}^{(t)T} + \alpha\mathbf{C}_4^{(t)}) \\ \cdot(\sum_{m=1}^{2}\theta_m\mathbf{L}_{u,m}^{(t)}\mathbf{L}_{u,m}^{(t)T} + \alpha\mathbf{C}_5^{(t)} + \gamma\mathbf{I})^{-1}, \tag{13}$$

where

$$\mathbf{C}_4^{(t)} = \mathbf{C}_4^{(t-1)} + \mathbf{B}^{(t)}\mathbf{L}_s^{(t)T}, \mathbf{C}_4^{(t-1)} = \mathbf{B}^{(t-1)}\mathbf{L}_s^{(t-1)T}, \\ \mathbf{C}_5^{(t)} = \mathbf{C}_5^{(t-1)} + \mathbf{L}_s^{(t)}\mathbf{L}_s^{(t)T}, \mathbf{C}_5^{(t-1)} = \mathbf{L}_s^{(t)}\mathbf{L}_s^{(t)T}. \tag{14}$$

V-Step: With all variables but common representation $\mathbf{V}^{(t)}$ fixed, the optimization problem in Eq. (11) becomes,

$$\min_{\mathbf{V}^{(t)}} \left\|\mathbf{B}^{(t)} - \mathbf{R}\mathbf{V}^{(t)}\right\|^2 + \beta\left\|\mathbf{V}^{(t)T}\mathbf{B}^{(t)} - r\cdot\mathbf{S}_{cc}\right\|^2, \\ \text{s.t. } \mathbf{V}^{(t)T}\mathbf{V}^{(t)} = n_t\mathbf{I}, \mathbf{V}^{(t)T}\mathbf{1} = \mathbf{0}. \tag{15}$$

Because of the orthogonal constraint on $\mathbf{V}^{(t)}$, the problem in Eq. (15) is a classical orthogonal Procrustes problem (Sun, Peng, and Ren 2024). Based on several algebraic transformations, Eq. (15) can be reformulated as,

$$\max_{\mathbf{V}^{(t)}} \text{Tr}(\mathbf{O}\mathbf{V}^{(t)T}), \\ \text{s.t. } \mathbf{V}^{(t)T}\mathbf{V}^{(t)} = n_t\mathbf{I}, \mathbf{V}^{(t)T}\mathbf{1} = \mathbf{0}. \tag{16}$$

Given $\mathbf{S}_{cc} = 2\mathbf{U}^{(t)T}\mathbf{U}^{(t)} - \mathbf{1}\mathbf{1}^T$, $\mathbf{O}$ can be defined as,

$$\mathbf{O} = \mathbf{R}^T\mathbf{B}^{(t)} + 2r\beta\mathbf{B}^{(t)}\mathbf{U}^{(t)T}\mathbf{U}^{(t)} - r\beta\mathbf{B}^{(t)}\mathbf{1}\mathbf{1}^T. \tag{17}$$

Let $\mathbf{J} = \mathbf{I} - \frac{1}{n_t}\mathbf{1}\mathbf{1}^T$, Eq. (16) can be solved by conducting the Singular Value Decomposition(SVD) as follows,

$$\mathbf{O}\mathbf{J}\mathbf{O}^T = \begin{bmatrix} \mathbf{Q} & \hat{\mathbf{Q}} \end{bmatrix} \begin{bmatrix} \mathbf{\Omega} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{Q} & \hat{\mathbf{Q}} \end{bmatrix}^T, \tag{18}$$

where $\mathbf{\Omega}$ denotes the diagonal matrix of positive eigenvalues. $\mathbf{Q}$ and $\hat{\mathbf{Q}}$ mean the eigenvectors. We execute the Gram-Schmidt process on $\hat{\mathbf{Q}}$ to obtain an orthogonal matrix $\overline{\mathbf{Q}}$. We further represent $\mathbf{P} = \mathbf{J}\mathbf{O}^T\mathbf{Q}\mathbf{\Omega}^{-1/2}$ and yield a random orthogonal matrix $\overline{\mathbf{P}}$. The closed-form solution of $\mathbf{V}^{(t)}$ can be calculated as,

$$\mathbf{V}^{(t)} = \sqrt{n_t}\begin{bmatrix} \mathbf{Q} & \overline{\mathbf{Q}} \end{bmatrix}\begin{bmatrix} \mathbf{P} & \overline{\mathbf{P}} \end{bmatrix}^T. \tag{19}$$

W-Step: Learn the projection $\mathbf{W}_m$ with other variables fixed. Analogous to the solution of $\mathbf{G}$, the closed-form solution of $\mathbf{W}_m$ can be calculated as,

$$\begin{aligned}\mathbf{W}_m = &(\xi_m\mathbf{H}_m^{(t)} + \theta_m\mathbf{G}\mathbf{L}_{u,m}^{(t)}\mathbf{X}_{u,m}^{(t)T}) \\ &\cdot(\xi_m\mathbf{F}_m^{(t)} + \theta_m\mathbf{X}_{u,m}^{(t)}\mathbf{X}_{u,m}^{(t)T} + \gamma\mathbf{I})^{-1},\end{aligned} \tag{20}$$

where

$$\begin{aligned}&\mathbf{H}_m^{(t)} = \mathbf{B}^{(t)}\mathbf{X}_{s,m}^{(t)T} + \mathbf{H}_m^{(t-1)}, \mathbf{F}_m^{(t)} = \mathbf{X}_{s,m}^{(t)}\mathbf{X}_{s,m}^{(t)T} + \mathbf{F}_m^{(t-1)}, \\ &\mathbf{H}_m^{(t-1)} = \mathbf{B}^{(t-1)}\mathbf{X}_{s,m}^{(t-1)}, \mathbf{F}_m^{(t-1)} = \mathbf{X}_{s,m}^{(t-1)}\mathbf{X}_{s,m}^{(t-1)T}.\end{aligned} \tag{21}$$

B-Step: Learn the hash code $\mathbf{B}^{(t)}$ with others variables unchanged. The optimization problem defined in Eq. (11) is deduced into the following subproblem,

$$\begin{aligned}&\min_{\mathbf{B}^{(t)}} \left\|\mathbf{B}^{(t)} - \mathbf{R}\mathbf{V}^{(t)}\right\|^2 + \alpha\left\|\mathbf{B}^{(t)} - \mathbf{G}\mathbf{L}_s^{(t)}\right\|^2 \\ &+ \beta(\left\|\mathbf{V}^{(t)T}\mathbf{B}^{(t)} - r\cdot\mathbf{S}_{cc}\right\|^2 + \left\|\mathbf{V}^{(t-1)T}\mathbf{B}^{(t)} - r\cdot\mathbf{S}_{oc}\right\|^2) \\ &+ \sum_{m=1}^{2}\xi_m\left\|\mathbf{B}^{(t)} - \mathbf{W}_m\mathbf{X}_{s,m}^{(t)}\right\|^2, \text{s.t. } \mathbf{B}^{(t)} \in \{-1,+1\}^{r\times n_{s,t}}.\end{aligned} \tag{22}$$

Considering $\left\|\mathbf{B}^{(t)}\right\|^2 = n_{s,t} \times r$, after several algebraic transformations, Eq. (22) can be rewritten as follows,

$$\begin{aligned}&\max_{\mathbf{B}^{(t)}} \mathrm{Tr}(\mathbf{B}^{(t)T}(\mathbf{R}\mathbf{V}^{(t)} + \alpha\mathbf{G}\mathbf{L}_s^{(t)} + \beta r(\mathbf{V}^{(t)}\mathbf{S}_{cc} \\ &+ \mathbf{V}^{(t-1)}\mathbf{S}_{oc}) + \sum_{m=1}^{2}\xi_m\mathbf{W}_m\mathbf{X}_{s,m}^{(t)}), \\ &\text{s.t. } \mathbf{B}^{(t)} \in \{-1,+1\}^{n_{s,t}\times r}.\end{aligned} \tag{23}$$

Given the definition of $\mathbf{S}_{cc}$ and $\mathbf{S}_M^{(t-1)}$, the solution of $\mathbf{B}^{(t)}$ can be derived as,

$$\begin{aligned}\mathbf{B}^{(t)} = &\mathrm{sgn}(\mathbf{R}\mathbf{V}^{(t)} + \alpha\mathbf{G}\mathbf{L}_s^{(t)} + 2\beta r\mathbf{C}_6^{(t)}\mathbf{U}^{(t)} \\ &- \beta r\mathbf{C}_7^{(t)}\mathbf{1}^T + \sum_{m=1}^{2}\xi_m\mathbf{W}_m\mathbf{X}_{s,m}^{(t)}),\end{aligned} \tag{24}$$

where

$$\begin{aligned}&\mathbf{C}_6^{(t)} = \mathbf{C}_6^{(t-1)} + \mathbf{V}^{(t)}\mathbf{U}^{(t)T},\ \mathbf{C}_6^{(t-1)} = \mathbf{V}^{(t-1)}\mathbf{U}^{(t-1)T}, \\ &\mathbf{C}_7^{(t)} = \mathbf{C}_7^{(t-1)} + \mathbf{V}^{(t)}\mathbf{1},\ \mathbf{C}_7^{(t-1)} = \mathbf{V}^{(t-1)}\mathbf{1}.\end{aligned} \tag{25}$$

R-Step: Learn the orthogonal matrix $\mathbf{R}$ with other variables unchanged. The optimization problem in Eq. (11) can be rewritten as,

$$\max_{\mathbf{R}} Tr(\mathbf{V}^{(t)}\mathbf{B}^{(t)T}\mathbf{R}), \quad \text{s.t.} \quad \mathbf{R}\mathbf{R}^T = \mathbf{I}. \tag{26}$$

To optimize $\mathbf{R}$, we first perform SVD, *i.e.*, $\mathbf{V}^{(t)}\mathbf{B}^{(t)T} = \mathbf{K}\sum\mathbf{O}^T$, where $\mathbf{K}$ is a $C \times C$ orthogonal matrix, $\sum$ is a $C \times r$ matrix and $\mathbf{O}$ is a $r \times r$ orthogonal matrix. The closed-form solution for $\mathbf{R}$ is

$$\mathbf{R} = \mathbf{O}\hat{\mathbf{K}}^T, \tag{27}$$

where $\hat{\mathbf{K}}$ contains first r columns of $\mathbf{K}$.

Theoretical Analyses

This subsection presents theoretical analyses of the proposed method, including stability, generalization, convergence, and time complexity analysis.

Stability Analyses The proposed hashing function is $\beta(n)$-stable for learning the projection $\mathbf{W}_m$. Specifically, given the dataset $\{D = [x_1, x_2, \cdots, x_n] : \|x_m\| \le g\}$, and D' be the sample with the i-th instance x_i in D replaced with an iid one x'_i. Considering the hashing function $F(W) \in [0, G]$ defined in Eq. (11), we have,

$$\forall \quad D, D', |F_D(W) - F_{D'}(W)| \le 24g^2G/n_t\gamma, \tag{28}$$

where $F_D(W)$ and $F_{D'}(W)$ mean $F(W)$ optimized by employing dataset D and D', respectively. And G denotes a universal constant.

Generalization Bound Based on stability analyses, we present a tight generalization bound. Given the sample feature space $\mathbf{X}_m = \{x_m \in \mathbb{R}^{d_m} : \|x_m\| \le g\}$, for any $\delta \in (0, 1)$, at least with probability $1 - \delta$, the hash function $F(W) \in [0, G]$ have,

$$R(W) \le \widehat{R}(W) + 24g^2G/n_t\gamma + (48g^2G/\gamma + G)\sqrt{\frac{\ln(1/\delta)}{2n_t}}, \tag{29}$$

where $R(W)$ and $\widehat{R}(W)$ denote generalization error and empirical error, respectively.

Convergence Analyses For convenience, the overall loss function is defined as $\mathcal{J}(\mathbf{B}^{(t)}, \mathbf{R}, \mathbf{V}^{(t)}, \mathbf{G}, \mathbf{W}_m)$. To optimize this function, we propose an alternate optimization algorithm that updates each variable while keeping the other four variables fixed. This approach ensures that the objective function $\mathcal{J}$ monotonically decreases with the increase of iterations. Moreover, the function is lower bound by zero. Consequently, the proposed algorithm is guaranteed to converge to a local minimum.

Time Complexity Analysis The time complexities for optimizing G-Step, V-Step, and R-Step are $O(n_{u,m,t}d_mr + n_{u,m,t}C^2 + n_{s,t}Cr)$, $O(n_{s,t}rd_m + n_{s,t}r^2)$, and $O(n_{s,t}Cr + n_{s,t}r^2)$, respectively. The time complexity for solving hash codes $\mathbf{B}^{(t)}$ is $O(n_{s,t}d_mr + n_{s,t}Cr + n_{s,t}r^2)$. Besides, the time complexity for learning projection $\mathbf{W}_m$ is $O(n_{u,m,t}d_m^2 + n_{u,m,t}d_mr + n_{u,m,t}Cr)$. As $n_{u,m,t}$ is usually much larger than $n_{s,t}$, r, and C, the overall training time complexity for training SSOCH can be simplified as $T \cdot O(n_{u,m,t}d_m^2 + n_{u,m,t}d_mr + n_{u,m,t}Cr)$, where T is the number of iterations. Considering the time complexity is linear to $n_{u,m,t}$, the proposed SSOCH is scalable for large-scale streaming datasets.

Experiments

To validate the effectiveness of the proposed SSOCH method, we conduct experiments on three widely-used cross-modal retrieval datasets: IAPR TC-12 (Escalante et al. 2010), NUSWIDE (Chua et al. 2009) and MIRFLICKR (Huiskes and Lew 2008). This section comprehensively describes our experiments, including experiment settings, performance analyses, and ablation studies.

Method	IAPR TC-12				NUSWIDE				MIRFLICKR			
	8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits
OCMH	0.3039	0.3025	0.3012	0.3019	0.3419	0.3416	0.3398	0.3397	0.5553	0.5565	0.5551	0.5557
OCMFH	0.3007	0.3006	0.2989	0.2983	0.3676	0.3677	0.3741	0.3769	0.5641	0.5618	0.5605	0.5590
OUCDH	0.3030	0.3010	0.3005	0.3021	0.3392	0.3392	0.3393	0.3393	0.5558	0.5554	0.5561	0.5548
OMGH	0.3047	0.3046	0.3046	0.3044	0.3406	0.3403	0.3406	0.3408	0.5562	0.5559	0.5556	0.5560
SPOCH	0.3383	0.3484	0.3634	0.3817	0.3496	0.3417	0.3419	0.3887	0.5657	0.5661	0.5781	0.5848
DPOCH	0.3026	0.3099	0.3100	0.3104	0.3402	0.3400	0.3405	0.3409	0.5611	0.5616	0.5630	0.5624
OLSH	0.3531	0.3627	0.3543	0.3412	0.3742	0.3529	0.3743	0.3463	0.6505	0.6593	0.6249	0.6036
ROHLSE	0.3018	0.3206	0.3290	0.3342	0.5784	0.5816	0.5883	0.5958	0.6419	0.6398	0.6452	0.6494
OFDDH	0.3392	0.3360	0.3395	0.3443	0.3630	0.3624	0.3728	0.3780	0.5996	0.6038	0.6175	0.6339
OSAH	0.4040	0.4232	0.4299	0.4395	0.5943	0.6198	0.6412	0.6473	0.6964	0.6991	0.7073	0.7152
DOCH	0.3725	0.3749	0.3933	0.4002	0.6091	0.6288	0.6386	0.6457	0.6456	0.6597	0.6729	0.6884
DOCMH	0.4234	0.4483	0.4623	0.4775	0.6116	0.6257	0.6411	**0.6585**	**0.7067**	0.7207	0.7257	0.7283
ODCH	0.3561	0.3614	0.3743	0.3799	0.6053	0.6204	0.6315	0.6350	0.6706	0.6808	0.6850	0.6901
SSOCH	**0.4479**	**0.4599**	**0.4819**	**0.4992**	**0.6144**	**0.6331**	**0.6447**	0.6483	0.6968	**0.7225**	**0.7249**	**0.7305**
OCMH	0.3010	0.3031	0.3029	0.3007	0.3432	0.3438	0.3402	0.3399	0.5523	0.5552	0.5561	0.5555
OCMFH	0.3377	0.3369	0.3331	0.3303	0.3932	0.4036	0.4168	0.4223	0.5604	0.5598	0.5594	0.5590
OUCDH	0.3027	0.3009	0.3005	0.3021	0.3392	0.3391	0.3392	0.3392	0.5557	0.5558	0.5560	0.5548
OMGH	0.3047	0.3046	0.3046	0.3048	0.3410	0.3407	0.3406	0.3406	0.5565	0.5559	0.5564	0.5562
SPOCH	0.3468	0.3539	0.3781	0.4017	0.3525	0.3430	0.3457	0.3981	0.5666	0.5673	0.5776	0.5905
DPOCH	0.3032	0.3106	0.3124	0.3142	0.3405	0.3403	0.3410	0.3416	0.5602	0.5616	0.5625	0.5627
OLSH	0.3448	0.3649	0.3560	0.3420	0.3864	0.3602	0.4067	0.3587	0.6069	0.6191	0.5999	0.5870
ROHLSE	0.4861	0.5104	0.5306	0.5444	0.6725	0.6800	0.6863	0.6924	0.7696	0.7791	0.7888	0.7929
OFDDH	0.3442	0.3440	0.3496	0.3578	0.3770	0.3781	0.3972	0.4090	0.6097	0.6175	0.6388	0.6558
OSAH	0.4578	0.4974	0.5194	0.5392	0.7297	0.7567	0.7673	0.7622	0.7396	0.7405	0.7422	0.7414
DOCH	0.3713	0.3876	0.4165	0.4236	0.7163	0.7584	0.7782	0.7887	0.6324	0.6547	0.6734	0.6838
DOCMH	0.4887	0.5309	0.5707	0.5974	0.7425	0.7670	0.7761	0.7839	0.7504	0.7695	0.7805	0.7857
ODCH	0.5110	0.5611	0.6027	0.6289	0.7417	0.7658	**0.7909**	**0.7961**	**0.7698**	0.7855	0.7915	0.8019
SSOCH	**0.5172**	**0.5622**	**0.6050**	**0.6306**	**0.7446**	**0.7698**	0.7825	0.7893	0.7660	**0.7863**	**0.7946**	**0.8037**

Table 1: Performance about mAP@all score on three benchmark datasets. The top panel is the performance for the Image2Text task, while the bottom panel is for the Text2Image task. The best mAP@all values of each case are shown in boldface and the suboptimal results are shown in underlines.

Experiment Settings

We compare the proposed SSOCH with thirteen state-of-the-art online cross-modal hashing methods: OCMH (Xie, Shen, and Zhu 2016), OCMFH (Wang et al. 2020), OUCDH (Li et al. 2022), OMGH (Liu et al. 2022a), SPOCH (Kang et al. 2023a), DPOCH (Kang et al. 2024a), OLSH(Yao et al. 2019), ROHLSE(Li et al. 2023), OFDDH(Liu, Wang, and Cheung 2022), OSAH(Zhang, Wu, and Chen 2023), DOCH(Zhan et al. 2021), DOCMH(Kang et al. 2024b), ODCH(Kang et al. 2023b). Among them, OCMH, OCMFH, OUCDH, OMGH, SPOCH, and DPOCH are unsupervised methods, while others are supervised ones. We sample 10% instances in each chunk for the supervised baselines to simulate the semi-supervised training situation. Additionally, for fairness, all experimental results are averaged over five runs.

In the implementation, we empirically set $\xi_m = 0.5$, $\delta = 10^{-3}$ and $\gamma = 10$. Moreover, we set $\alpha = 10^{-4}$, $\beta = 10$ and $\theta = 10^{-8}$ through cross-validation and grid search. The number of iterations T is set to 10. For simplicity, the number of categories in the pseudo-label is fixed as 3. The length of the Hadamard label C is set as 32 for the MIRFLICKR and NUSWIDE datasets, and 256 for the IAPR TC-12 dataset. All experiments are performed on a computer with an Intel(R) Core(TM) i9-10900K CPU@ 3.70GHz 64GB RAM.

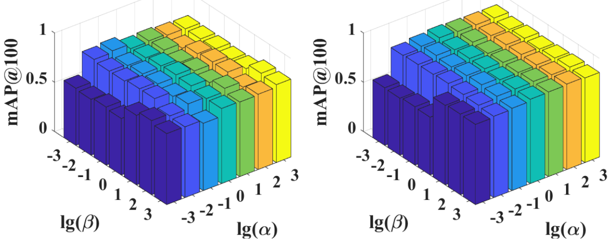
We evaluate the proposed approach on two standard cross-modal retrieval tasks: Text2Image, where text queries are used to search for similar images, and Image2Text, where image queries are used to search for similar text. We adopt the most commonly used evaluation metrics: mean Average Precision (mAP@all), mean Average Precision@K (mAP@K), and precision@K, with K set to 50 and 100. Additionally, we report the training time to assess the computational efficiency of the proposed method.

Performance Analyses

Accuracy Discussion: Table 1 presents the mAP@all scores of the proposed SSOCH method and the compared methods on three benchmark datasets. SSOCH exhibits significant performance superiority over the baselines, particularly on the IAPR TC-12 dataset. This demonstrates that the proposed online tri-consistent preserving scheme effectively integrates consistent information across different scenarios, facilitating the effectiveness of hash codes. Besides, when compared with the ODCH method, we observe a slightly inferior performance on the NUSWIDE dataset. A possible reason is that NUSWIDE is a huge-scale multi-label dataset that contains extremely abundant supervised semantic information, which hinders performance gains from exploring the structure infor-

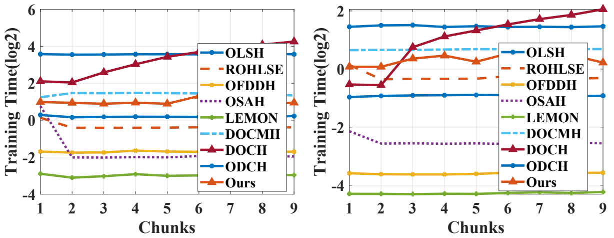
Method	IAPR TC-12				NUSWIDE				MIRFLICKR			
	8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits	8 bits	16 bits	32 bits	64 bits
SSOCH-h	0.3992	0.4225	0.4390	0.4463	0.5854	0.6227	0.6278	0.6401	0.6933	0.7067	0.7115	0.7235
SSOCH-u	0.4426	0.4566	0.4768	0.4929	0.6121	0.6289	0.6364	0.6397	0.6914	0.7187	0.7135	0.7202
SSOCH-l	0.4478	0.4567	0.4749	0.4942	0.5994	0.6355	0.6323	0.6389	0.6919	0.7121	0.7139	0.7197
SSOCH-s	0.3139	0.3244	0.3231	0.3283	0.3476	0.3495	0.3543	0.3640	0.5534	0.5548	0.5539	0.5581
SSOCH	0.4479	0.4599	0.4819	0.4992	0.6144	0.6331	0.6447	0.6483	0.6968	0.7225	0.7249	0.7305
SSOCH-h	0.4642	0.5001	0.5288	0.5529	0.7351	0.7645	0.7722	0.7831	0.7474	0.7719	0.7829	0.7936
SSOCH-u	0.5058	0.5564	0.6058	0.6219	0.7326	0.7646	0.7813	0.7855	0.7658	0.7803	0.7885	0.7967
SSOCH-l	0.5070	0.5547	0.6056	0.6225	0.7304	0.7727	0.7791	0.7879	0.7602	0.7796	0.7909	0.7956
SSOCH-s	0.3158	0.3270	0.3302	0.3380	0.3519	0.3565	0.3655	0.3808	0.5535	0.5553	0.5540	0.5575
SSOCH	0.5172	0.5622	0.6050	0.6306	0.7446	0.7698	0.7825	0.7893	0.7660	0.7863	0.7946	0.8037

Table 2: Ablation study in terms of mAP@all score on three benchmark datasets. The top panel is the performance for the Image2Text task, while the bottom panel is for the Text2Image task.



(a) Image2Text (b) Text2Image

Figure 2: The parameter sensitivity with mAP@100 score about α and β on the MIRFLICKR dataset.



(a) NUSWIDE (b) MIRFLICKR

Figure 3: The training time results vary with chunks based on two datasets.

mation of unsupervised data.

Parameter Sensitivity Analyses: We conduct experiments to analyze the parameter sensitivity, including the parameters α and β. Figure 2 depicts the mAP@100 scores of the proposed SSOCH when these two parameters range from 10^{-3} to 10^{3}. The proposed SSOCH exhibits satisfactory stability and sensitivity as parameters change.

Training Time Analyses: Figure 3 shows that SSOCH achieves comparable and superior time costs compared to the state-of-the-art baselines on different data chunks. These findings indicate that our method not only effectively enhances retrieval accuracy, but also maintains competitive efficiency.

Ablation Study

To further validate the contribution of each component, we design four variants. The comparison performance is summarized in Table 2.

SSOCH-h: To evaluate the discriminative Hadamard embedding, we design SSOCH-h. It adopts traditional one-hot label embedding. A remarkable dropping appears, indicating the superiority of the proposed discriminative labels.

SSOCH-u: To evaluate the effectiveness of the alignment-free pseudo-label, we design SSOCH-u. It discards the pseudo-label data regularization. A slight drop in SSOCH-e demonstrates the effectiveness of exploring the semantic information underlying the unlabeled data.

SSOCH-l: To assess the impact of label embedding, we design SSOCH-l. It discards the discriminative label embedding in Eq. (7). The performance of SSOCH is superior to SSOCH-l, verifying this label embedding can exploit elaborate semantic information.

SSOCH-s: The variant SSOCH-s discards the asymmetric similarity preserving in Eq. (9). The results on this variant illustrate a sharp drop, verifying the importance of fine-grained pairwise similarity preservation.

Conclusion

This study proposes a novel online hashing method to address the challenges of complex semi-labeled and semi-paired data in online cross-modal retrieval scenarios. Specifically, we integrate alignment-free pseudo-labeling and pseudo-label data regularization to explore semantic information in unlabeled streaming data, thereby assisting the hashing learning process. Hadamard embedding is introduced to enhance label discrimination and generate fine-grained similarity relationships, thus alleviating the imbalance updating problem. Furthermore, the method integrates various data information into the online tri-consistent preserving scheme, fully exploring the consistency across labeled/unlabeled data, modalities, and streaming chunks. Comprehensive theoretical analyses and extensive experiments demonstrate the superiority of SSOCH. In future work, we plan to explore techniques to enhance the quality of hash codes for unlabeled data, thereby enhancing retrieval performance over the entire dataset.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (U23A20389, 62206160, 62176141, 62176139), Young Talent of Lifting Engineering for Science and Technology in Shandong (SDAST2024QTA020), Major Basic Research Project of Shandong Provincial Natural Science Foundation (ZR2024ZD03), Natural Science Foundation of Shandong Province (ZR2022QF082), Taishan Scholar Project of Shandong Province (tsqn202103088), Shandong Provincial Natural Science Foundation for Distinguished Young Scholars (ZR2021JQ26), the Science and Technology Innovation Program for Distinguished Young Scholars of Shandong Province Higher Education Institutions (2024KJH084), and special funds for distinguished professors of Shandong Jianzhu University.

References

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. NUS-WIDE: a real-world web image database from National University of Singapore. In *Proceedings of the ACM international conference on image and video retrieval*, 48. ACM.

Escalante, H. J.; Hernández, C. A.; González, J. A.; López-López, A.; Montes-y-Gómez, M.; Morales, E. F.; Sucar, L. E.; Pineda, L. V.; and Grubinger, M. 2010. The segmented and annotated IAPR TC-12 benchmark. *Comput. Vis. Image Underst.*, 114(4): 419–428.

Fan, W.; Zhang, C.; Li, H.; Jia, X.; and Wang, G. 2023. Three-Stage Semisupervised Cross-Modal Hashing With Pairwise Relations Exploitation. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Huiskes, M. J.; and Lew, M. S. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM SIGMM International Conference on Multimedia Information Retrieval, 2008*, 39–43.

Kang, X.; Liu, X.; Lu, P.; Zhao, Z.; Nie, X.; Wang, S.; and Yin, Y. 2023a. Online Cross-Modal Hashing with Double Structure Preserving. *Journal of Computer Research and Development*, 1–13.

Kang, X.; Liu, X.; Xue, W.; Nie, X.; and Yin, Y. 2024a. Online Cross-modal Hashing With Dynamic Prototype. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(8).

Kang, X.; Liu, X.; Xue, W.; Zhang, X.; Nie, X.; and Yin, Y. 2024b. Discrete online cross-modal hashing with consistency preservation. *Pattern Recognition*, 110688.

Kang, X.; Liu, X.; Zhang, X.; Nie, X.; and Yin, Y. 2023b. Online Discriminative Cross-modal Hashing. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Lawson, C. L.; and Hanson, R. J. 1995. *Solving least squares problems*, volume 15 of *Classics in applied mathematics*. Siam.

Li, L.; Shu, Z.; Yu, Z.; and Wu, X.-J. 2023. Robust online hashing with label semantic enhancement for cross-modal retrieval. *Pattern Recognition*, 145: 109972.

Li, X.; Wu, W.; Yuan, Y.; Pan, S.; and Shen, X. 2022. Online unsupervised cross-view discrete hashing for large-scale retrieval. *Appl. Intell.*, 52(13): 14905–14917.

Lin, M.; Ji, R.; Liu, H.; Sun, X.; Chen, S.; and Tian, Q. 2020. Hadamard Matrix Guided Online Hashing. *Int. J. Comput. Vis.*, 128(8): 2279–2306.

Lin, M.; Ji, R.; Liu, H.; Sun, X.; Wu, Y.; and Wu, Y. 2019. Towards Optimal Discrete Online Hashing with Balanced Similarity. In *AAAI*, 8722–8729. AAAI Press.

Liu, X.; Li, J.; Nie, X.; Zhang, X.; and Yin, Y. 2024. Fast Unsupervised Cross-Modal Hashing with Robust Factorization and Dual Projection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(12).

Liu, X.; Wang, X.; and Cheung, Y.-M. 2022. FDDH: Fast Discriminative Discrete Hashing for Large-Scale Cross-Modal Retrieval. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11): 6306–6320.

Liu, X.; Yi, J.; Cheung, Y.-m.; Xu, X.; and Cui, Z. 2022a. OMGH: Online Manifold-Guided Hashing for Flexible Cross-modal Retrieval. *IEEE Transactions on Multimedia*, 1–1.

Liu, X.; Yu, G.; Domeniconi, C.; Wang, J.; Xiao, G.; and Guo, M. 2022b. Weakly Supervised Cross-Modal Hashing. *IEEE Transactions on Big Data*, 8(2): 552–563.

Lu, X.; Zhu, L.; Cheng, Z.; Li, J.; Nie, X.; and Zhang, H. 2019. Flexible Online Multi-Modal Hashing for Large-Scale Multimedia Retrieval. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*, 1129–1137.

Sun, Y.; Peng, D.; and Ren, Z. 2024. Discrete aggregation hashing for image set classification. *Expert Systems with Applications*, 237: 121615.

Sun, Y.; Wang, X.; Peng, D.; Ren, Z.; and Shen, X. 2023. Hierarchical Hashing Learning for Image Set Classification. *IEEE Transactions on Image Processing*, 32: 1732–1744.

Wang, D.; Wang, Q.; An, Y.; Gao, X.; and Tian, Y. 2020. Online Collective Matrix Factorization Hashing for Large-Scale Cross-Media Retrieval. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 1409–1418. ACM.

Wang, X.; Liu, X.; Hu, Z.; Wang, N.; Fan, W.; and Du, J.-X. 2019. Semi-Supervised Semantic-Preserving Hashing for Efficient Cross-Modal Retrieval. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 1006–1011.

Wang, Y.; Luo, X.; and Xu, X. 2020. Label Embedding Online Hashing for Cross-Modal Retrieval. In *The 28th ACM International Conference on Multimedia, 2020*, 871–879. ACM.

Xie, L.; Shen, J.; and Zhu, L. 2016. Online Cross-Modal Hashing for Web Image Retrieval. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence,*, 294–300. AAAI Press.

Yang, R. 2013. New results on some quadratic programming problems. *Dissertations & Theses - Gradworks.*

Yao, T.; Li, Y.; Guan, W.; Wang, G.; Li, Y.; Yan, L.; and Tian, Q. 2023. Discrete Robust Matrix Factorization Hashing for Large-Scale Cross-Media Retrieval. *IEEE Transactions on Knowledge and Data Engineering*, 35(2): 1391–1401.

Yao, T.; Wang, G.; Yan, L.; Kong, X.; Su, Q.; Zhang, C.; and Tian, Q. 2019. Online latent semantic hashing for cross-media retrieval. *Pattern Recognit.*, 89: 1–11.

Yu, J.; Ma, L.; Li, Z.; Peng, Y.; and Xie, S. 2023. Open-World Object Detection via Discriminative Class Prototype Learning. *CoRR*, abs/2302.11757.

Zhan, Y. W.; Wang, Y.; Sun, Y.; Wu, X. M.; Luo, X.; and Xu, X. S. 2021. Discrete Online Cross-Modal Hashing. *Pattern Recognition*.

Zhang, D.; Wu, X.-J.; and Chen, G. 2023. ONION: Online Semantic Autoencoder Hashing for Cross-Modal Retrieval. *ACM Transactions on Intelligent Systems and Technology*, 14(2): 1–18.

Zhang, J.; Peng, Y.; and Yuan, M. 2020. SCH-GAN: Semi-Supervised Cross-Modal Hashing by Generative Adversarial Network. *IEEE Transactions on Cybernetics*, 50(2): 489–502.

Zhang, X.; Liu, X.; Nie, X.; Kang, X.; and Yin, Y. 2024. Semi-Supervised Semi-Paired Cross-Modal Hashing. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(7): 6517–6529.

Zhu, L.; Lu, X.; Cheng, Z.; Li, J.; and Zhang, H. 2020. Deep Collaborative Multi-View Hashing for Large-Scale Image Search. *IEEE Trans. Image Process.*, 29: 4643–4655.

Zhu, L.; Zheng, C.; Guan, W.; Li, J.; Yang, Y.; and Shen, H. T. 2023. Multi-modal Hashing for Efficient Multimedia Retrieval: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 1–20.