

Article

Feature Fusion-Based Cross-Modal Proxy Hashing Retrieval

Yan Zhao  and Huaiying Li *

Faculty of Artificial Intelligence, Shanghai University of Electric Power, Shanghai 201300, China;
yanzhao@shiep.edu.cn

* Correspondence: lihuaiying@mail.shiep.edu.cn; Tel.: +86-15862707132

Abstract

Due to its cost-effective and high-efficiency retrieval advantages, deep hashing has attracted extensive attention in the field of cross-modal retrieval. However, despite significant progress in existing deep cross-modal hashing methods, several limitations persist: they struggle to establish consistent mapping relationships across different modalities, fail to effectively bridge the semantic gap between heterogeneous data, and consequently suffer from semantic information loss and incomplete semantic understanding during cross-modal learning. To address these challenges, this paper proposes a Feature Fusion-based Cross-modal Proxy Hashing (FFCPH) retrieval method. This approach integrates multi-modal semantic information through a feature fusion module to generate discriminative and robust fused features. Furthermore, a novel joint loss function, which comprises cross-modal proxy loss, cross-modal irrelevant loss, and cross-modal consistency loss, is designed to preserve inter-sample similarity ranking accuracy and mitigate the semantic gap across modalities. Experimental results on three widely used benchmark datasets demonstrate that the proposed method significantly outperforms state-of-the-art approaches in retrieval performance.

Keywords: deep hashing; cross-modal retrieval; feature fusion; deep learning

1. Introduction

With the increasing sophistication of Internet technologies, users generate vast amounts of multimodal data, including images, text, audio, and video, on a daily basis. This proliferation has led to exponential growth in multimedia repositories, which exhibit inherent heterogeneity and complex semantic relationships across modalities [1]. To enable effective cross-modal information fusion, it is essential to bridge the representational gaps between diverse data types, mitigating discrepancies in feature structure and distribution [2]. In this context, cross-modal hashing has emerged as a critical technique that learns unified binary hash codes to establish semantic correspondences across modalities. By mapping heterogeneous data into a common Hamming space, it significantly improves retrieval efficiency and scalability while preserving cross-modal semantic consistency, thereby supporting large-scale multimedia search applications [3].

In the past few years, the rapid evolution of deep learning methodologies has profoundly accelerated developments in cross-modal hashing. The incorporation of deep neural networks, particularly through architectures such as convolutional and recurrent networks, has established itself as a predominant research avenue. These models facilitate more effective representation learning by capturing high-level semantic correlations across heterogeneous data modalities, thereby substantially enhancing retrieval accuracy and



Received: 7 January 2026

Revised: 27 January 2026

Accepted: 29 January 2026

Published: 3 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

computational efficiency. Furthermore, the synergy between feature abstraction capabilities of deep learning and hashing techniques offers promising potential for scalable multimedia search in real-world applications [4]. Despite significant progress in deep cross-modal hashing, critical challenges persist. A primary issue is the difficulty of constructing semantically coherent and structure-preserving mappings across heterogeneous modalities. Moreover, existing methods struggle to reconcile inherent semantic discrepancies between diverse data types (e.g., images and text) due to their distinct feature distributions and structural properties. Consequently, semantic information is often lost during cross-modal alignment, leading to incomplete or distorted representations. These limitations collectively hinder semantic understanding and constrain the performance and generalizability of cross-modal retrieval systems.

To overcome the challenges identified in existing approaches, we introduce a feature fusion-based proxy hashing framework designed for cross-modal retrieval tasks. Raw image and text data are fed into a feature extraction network to obtain image and text features, respectively. These features are then integrated through a feature fusion module to combine multimodal semantic information, generating fused features that are both discriminative and robust. Finally, the hash function learning process is optimized through a carefully designed joint loss function that integrates three complementary components: cross-modal proxy loss, cross-modal irrelevant loss, and cross-modal consistency loss. In the feature extraction stage, the proposed method employs the CLIP model as its backbone architecture, replacing conventional convolutional neural networks (CNNs). CLIP incorporates a dual-encoder architecture comprising a visual encoder for image processing and a transformer-based encoder for textual input. On the image side, the ViT-B/16 pre-trained model of CLIP is used for feature extraction. On the text side, Byte-Pair Encoding (BPE) is employed for tokenization, and the processed text vectors are input into the text transformer encoder. For the feature fusion module, inspired by the multi-head attention mechanism in the Transformer architecture, a cross-modal feature fusion mechanism is designed. It integrates semantic information from images and text via an attention-weighted strategy to produce a unified feature representation with both discriminability and robustness. For the hash learning module, we propose a joint loss function comprising cross-modal proxy loss, cross-modal irrelevant loss, and cross-modal consistency loss, which replaces traditional pairwise or triplet losses. This design effectively preserves inter-sample similarity rankings and mitigates inter-modal semantic gaps. Specifically, the cross-modal proxy loss aligns heterogeneous features via proxy representations to bridge semantic gaps; the cross-modal irrelevant loss reduces modality-specific noise by penalizing irrelevant correlations between irrelevant pairs; and the cross-modal consistency loss ensures the compatibility and effectiveness of fused features, providing stable and semantically rich inputs for hash code generation. This multi-objective optimization framework enhances the discriminability of hash codes while improving the model's robustness and generalization capability in large-scale cross-modal retrieval.

In summary, the contributions of this paper are as follows:

- We propose a cross-modal feature fusion module incorporating attention mechanisms to integrate semantic information from images and text. This module effectively eliminates heterogeneity between modalities, producing discriminative and robust fused features that mitigate the semantic gap.
- Based on the feature fusion module, a cross-modal consistency loss is proposed and proxy hashing is introduced, thereby constructing a joint loss function composed of cross-modal proxy loss, cross-modal irrelevance loss, and cross-modal consistency loss. Compared to traditional pairwise and triplet losses, this loss function effectively

preserves the accuracy of inter-sample similarity ranking and mitigates the semantic gap between modalities.

- Extensive experiments conducted on three widely used benchmark datasets demonstrate that the proposed method significantly outperforms existing state-of-the-art approaches in cross-modal retrieval tasks.

2. Related Work

The fundamental objective of cross-modal hashing is to map high-dimensional heterogeneous data into a unified low-dimensional Hamming space, where semantic similarities between modalities such as images and text can be efficiently measured using compact binary hash codes. By performing Hamming distance computations on these binary representations, the method enables fast and scalable similarity retrieval across modalities, even in large-scale scenarios.

However, data from different modalities inherently exhibit heterogeneous characteristics, such as divergent statistical properties and structural representations. To achieve effective cross-modal retrieval, it is essential not only to bridge the modality gap by learning a common subspace but also to preserve both intra-modal and inter-modal semantic relationships within the shared Hamming space. This entails aligning high-level semantic concepts across modalities while maintaining the neighborhood structures of embedded data points, thereby facilitating semantically consistent and efficient retrieval across diverse data types [5].

Based on the presence or absence of supervisory signals—such as semantic labels or pairwise similarity information—during training, existing cross-modal hashing methods can be broadly categorized into two groups. Supervised methods explicitly incorporate such annotations to enhance semantic alignment across modalities, typically yielding more discriminative hash codes that better preserve label consistency. In contrast, unsupervised methods rely solely on the intrinsic structure and distribution of the data, such as feature correlations or proximity relations, without external supervision. While unsupervised approaches are more widely applicable in label-scarce settings, they often struggle to capture high-level semantic nuances compared to supervised methods [6,7].

2.1. Supervised Cross-Modal Hashing

Supervised approaches utilize human-annotated semantic labels to direct the training of hash functions, aiming to reduce the Hamming distance among analogous instances and simultaneously increase it among disparate ones. Due to the incorporation of strong semantic supervision, these methods typically learn more discriminative hash representations, thereby achieving higher retrieval accuracy in cross-modal tasks. In recent years, the development of deep learning has facilitated notable advancements in supervised deep cross-modal hashing techniques. For example, SCM employs supervised learning to train large-scale hash models and uses semantic correlation maximization constraints to ensure multi-label semantic consistency of hash codes [8]. SePH converts semantic labels into continuous probability distributions and employs sampling strategies to learn hash functions, reducing inter-modal discrepancies while preserving semantic consistency in the Hamming space [7]. DCH utilizes a pairwise affinity matrix along with a negative log-likelihood-based loss function to facilitate model optimization, thereby enabling direct learning of discriminative binary hash codes under discrete constraints [9]. EGDH balances the hash learning process across image and text modalities through isometric constraints to prevent dominance by a single modality, while leveraging discriminative loss functions to enhance semantic differentiation [10]. SDCH employs semantic-driven guidance to learn semantic features of cross-modal data and generates semantically consistent hash

codes, demonstrating strong semantic preservation capabilities in cross-modal retrieval [11]. VLKD adopts hierarchical knowledge distillation techniques to fuse and align visual and textual semantic features, progressively acquiring semantic knowledge to improve semantic preservation and retrieval accuracy in cross-modal tasks [12]. DCHMT learns joint representations of images and text and generates continuous hash codes through differentiable hashing techniques, later quantized into binary hash codes [13]. MSSPQ maps data from different modalities (e.g., images and text) into a unified semantic space to generate compact binary hash codes while ensuring semantic consistency in retrieval results [14]. MCCMR coordinates the feature fusion process of image and text modalities through a mutual compromise mechanism to enhance semantic consistency across modalities, enabling precise matching between images and text [15].

2.2. Unsupervised Cross-Modal Hashing

Unsupervised methods do not rely on any label information but instead learn hash functions from the inherent statistical associations and topological structures of the data. Although these approaches eliminate the need for manual annotations, their retrieval accuracy is generally lower than that of supervised counterparts, primarily owing to the absence of explicit semantic supervision. During the initial phase of unsupervised cross-modal hashing, methods predominantly employed classical machine learning frameworks, including matrix factorization and spectral clustering, to achieve shallow hash learning. For example, SH generates hash codes based on spectral graph theory by constructing a similarity matrix of the data and performing eigen decomposition [16]. CVH employs view-specific hash functions to project semantically related instances into analogous binary representations, thereby facilitating efficient similarity retrieval across different modalities [17]. IMH employs a linear regression-based hashing framework to project heterogeneous features into a common Hamming space, preserving both intra-modal structural consistency and inter-modal semantic correlations. This approach ensures robust out-of-sample generalization, reduces computational overhead, and maintains discriminative power in the generated hash codes, making it suitable for large-scale cross-modal retrieval applications [18]. LSSH utilizes sparse representation learning and matrix factorization to extract salient features from different modalities, aligns them into a unified latent space via joint semantic projection, and optimizes them iteratively to generate compact hash codes. This approach preserves intrinsic modality structures while enhancing discriminability and semantic coherence, thereby improving retrieval performance in large-scale cross-modal search [19].

In the past few years, the rapid advancement of deep learning has significantly accelerated the development of unsupervised cross-modal hashing learning methods. These methods can effectively model complex nonlinear relationships across heterogeneous modalities (e.g., images, text, audio) without requiring extensive labeled datasets. By leveraging deep architectures like autoencoders, generative adversarial networks (GANs), and deep metric learning frameworks, these methods effectively learn shared latent representations that preserve semantic consistency and structural dependencies across modalities. Such capability not only enhances the robustness and discriminative power of the learned features but also facilitates more effective downstream tasks, including cross-modal retrieval, zero-shot learning, and multimodal fusion. For instance, DJSRH incorporates intrinsic neighborhood structures across multiple modalities to uncover the underlying semantic information within cross-modal data [20]. UKD employs the principle of knowledge distillation by deriving relational semantics from a teacher network's outputs to formulate a semantic affinity matrix, thereby facilitating hash code learning in the student network [21]. AGCH leverages graph convolutional networks to perform unsupervised cross-modal hashing, integrating multi-metric approaches to synthesize inter-modal structural correlations and

construct a similarity matrix for hash function learning [22]. DMFH utilizes deep learning models to extract multi-scale features of cross-modal data and employs multimodal fusion to learn joint representations for hash code generation [23]. CIRH constructs a collaborative multimodal graph to model heterogeneous inter-modal dependencies and conducts semantic fusion over the graph to produce enriched multimodal embeddings [24]. To learn modality-invariant features, MIAN employs an asymmetric network architecture to separately process image and text modalities while introducing constraints to align their semantic representations in a shared space [25].

3. Our Method

3.1. Notation Definition

This paper primarily focuses on the cross-modal retrieval task between image and text modalities. Given a dataset containing N image-text pairs, it can be represented as a set $D = \{d_i\}_{i=1}^N$, where $d_i = \{x_i, y_i\}$, with x_i denoting the i -th image sample and y_i representing the y -th text sample. $l_i = [l_1, l_2, \dots, l_k]$ is the label corresponding to l_i , where C indicates the number of categories. The core objective of deep cross-modal hashing is to learn hash functions H^x and H^y for images and text, respectively, and then map the image and text data into corresponding binary hash codes $B^x \in \{-1, 1\}^K$ and $B^y \in \{-1, 1\}^K$, where K represents the length of the binary hash codes.

3.2. Overall Framework

The overall framework of FFCPH is illustrated in Figure 1. First, raw image and text data are input into the feature extraction network to obtain image features F_x and text features F_y . Subsequently, a feature fusion module integrates multimodal semantic information to generate fused features F_f that exhibit both discriminative and robust properties. Finally, hash function learning is accomplished under the guidance of a joint loss function composed of cross-modal proxy loss, cross-modal Irrelevant loss, and cross-modal consistency loss. Implementation details of each component will be elaborated in subsequent sections.

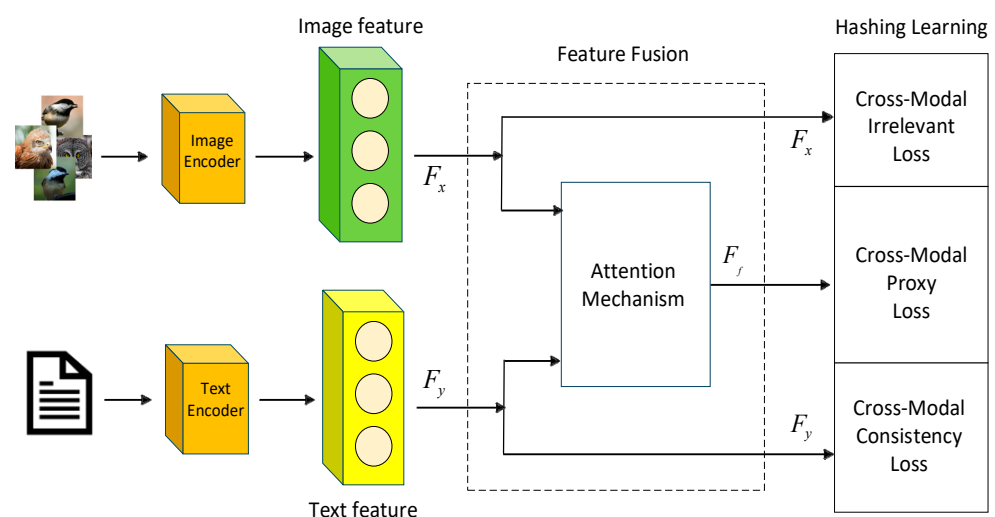


Figure 1. Overall framework.

3.3. Feature Extraction Module

This paper employs the CLIP pre-trained network for feature extraction of images and text [26]. The CLIP model utilizes contrastive learning to train on a large-scale dataset of image-text pairs, enabling effective alignment of image and text features. The architecture of CLIP incorporates two primary encoding branches: a visual encoder and a linguistic en-

coder. The visual component commonly utilizes a Vision Transformer (ViT) as its backbone, whereas the linguistic encoder is built upon a standard Transformer-based framework. For the image modality, CLIP employs a pre-trained ViT-B/16 model to obtain discriminative visual representations. For the text modality, Byte Pair Encoding (BPE) is applied for tokenization, and the processed token sequences are fed into the text encoder.

3.3.1. Image-Side Network

Feature extraction for images is performed by the image encoder of CLIP. First, the input image undergoes preliminary feature extraction through convolutional layers. The feature maps output by the convolutional layers are then flattened into one-dimensional vectors and combined with trainable positional encodings to preserve spatial information. Finally, the processed feature vectors are fed into the image encoder composed of 12 encoding blocks. After processing through these 12 blocks, the final image features are obtained as F_x .

The equations are as follows:

$$F_x = ViT\left(\left[Conv\left(x_i \oplus e_{px}^x\right)\right]\right), \quad (1)$$

where e_{px}^x denotes the learnable positional encoding, and $\oplus$ represents the concatenation operation.

3.3.2. Text-Side Network

Text feature extraction is performed using the CLIP text transformer encoder. Initially, the input text is tokenized via Byte Pair Encoding (BPE), resulting in a sequence of 512 tokens. Each token is then augmented with a trainable positional encoding. The resulting text vectors are then fed into the transformer-based text encoder, which produces high-dimensional semantic embeddings through multi-layer self-attention and feed-forward computation, ultimately yielding the final distributed representations of the input text. The final extracted text feature is F_y .

3.4. Feature Fusion Module

Traditional cross-modal methods often perform interaction between image and text features by simply concatenating their feature vectors or merging them directly through fully connected layers. While such approaches combine information from both modalities, they essentially represent a linear and semantically agnostic fusion strategy. In complex semantic scenarios, the fused representations often lack the capacity to adequately model high-level semantic interactions across modalities, which consequently results in the omission of pivotal semantic cues during the integration process.

To address this issue, this paper draws inspiration from the multi-head attention mechanism in the Transformer and designs a feature fusion module incorporating an attention mechanism [27]. The core idea is to treat text features as “sequential semantic descriptions” (serving as key vectors K and value vectors V), while image features are regarded as objects to be “queried” and “enhanced” by contextual semantics (serving as query vectors Q). The attention mechanism empowers the feature fusion module to adaptively quantify the correlation strength, represented as attention weights, among individual elements of both visual and textual features. Based on these weights, the textual semantic information is weighted and aggregated, ultimately producing fused features that retain the original image information while incorporating relevant textual semantics.

3.4.1. The Standard Multi-Head Attention Mechanism

To elucidate the underlying principle of the feature fusion module, it is imperative to first revisit the mathematical formulation of the canonical multi-head attention mechanism, a core component of Transformer-based architectures. In standard multi-head attention, the input is structured into three distinct vector representations: queries (Q), keys (K), and values (V), typically projected from the same feature space or modality. For each attention head, the mechanism computes a weighted sum of value vectors, where the weights are derived from a scaled dot-product affinity between queries and keys. The computation of a single attention head is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right)V, \quad (2)$$

where Q , K , and V represent the query matrix, key matrix, and value matrix, respectively; d_k denotes the dimensionality of the key vector K , and $\sqrt{d_k}$ is the scaling factor used to prevent excessively large dot product results that may lead to vanishing gradients in the softmax operation [27]. The attention weight distribution is obtained through the above formula. Subsequently, Q , K , and V are projected h times using different linear projection matrices W_i^Q , W_i^K , and W_i^V , and h attention heads are computed in parallel. The outputs from all attention heads are concatenated and subsequently transformed through a linear projection layer to form the final multi-head attention representation. This process can be formally expressed as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \cdot W^O, \quad (3)$$

where W^O represents the trainable output projection matrix that integrates information captured by different attention heads.

3.4.2. Module Architecture

The architecture of the feature fusion module is shown in Figure 2. The purpose of this module is to combine text and image features through a specific method, thereby eliminating the heterogeneity between images and text and obtaining fused features that incorporate semantic information from both modalities. Considering that image and text features in this context are global feature vectors rather than long sequences, we adopt the computational logic of single-head attention to meet the efficiency and discriminative requirements of the cross-modal hashing task. Additionally, in the following sections, we define the image features as Q (the query) and the text features as both K (the matched keys) and V (the aggregated values).

The previously obtained image and text features are denoted as $X \in \mathbb{R}^{B \times V}$ and $Y \in \mathbb{R}^{B \times V}$, respectively, where B represents the batch size and V is the length of the feature vectors, consistent with the output dimensionality of the CLIP model. The attention map is first computed as follows:

$$A = X \cdot Y^T. \quad (4)$$

Here, matrix multiplication is used to compute the attention map between image features and text features. The shape of X changes from $[B, V]$ to $[B, V, 1]$, while the shape of Y also changes from $[B, V]$ to $[B, 1, V]$. Therefore, the resulting attention map A has the shape $[B, V, V]$. The attention map A represents the global semantic correlation strength between image features (as queries) and text features (as keys).

To transform the raw scores into valid probability distribution weights, we apply the softmax function to each row of the attention map A along the column direction for normalization:

$$\sigma = \text{softmax}(A). \quad (5)$$

Subsequently, the normalized attention weights σ are employed to perform a weighted summation on the original text features Y , generating a semantically filtered text vector Z tailored to the current image:

$$Z = \sigma \cdot Y. \quad (6)$$

At this point, the text vector Z is no longer an independent text feature but a dynamically reorganized representation enriched with cross-modal correlation information guided by the image features.

Finally, by adding the text feature (which has been filtered and reconstructed through the attention mechanism) to the original image feature, the final fused feature O is obtained:

$$O = \gamma Z + X. \quad (7)$$

Among them, γ serves as a learnable parameter that modulates the contribution of the weighted textual features to the resulting fused representation.

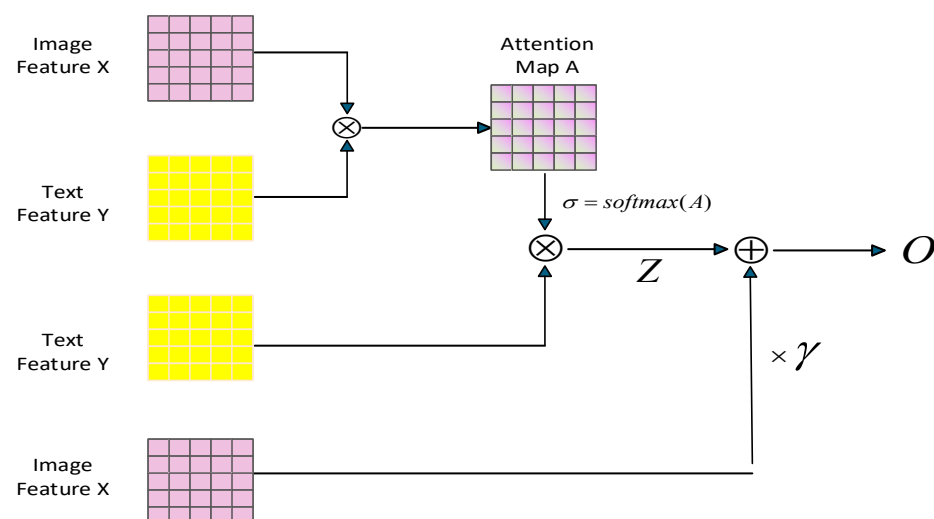


Figure 2. Feature fusion module block diagram.

Within the feature fusion module, the attention mechanism serves as a semantic filter that dynamically identifies and emphasizes the most relevant elements from the input representations. For a given image, it automatically assigns higher weights to text describing its core content while suppressing irrelevant or redundant textual information, thereby achieving effective information filtering. By allowing image features to actively ‘query’ text features and performing a weighted summation in the feature space, this module mathematically establishes a nonlinear mapping relationship from the visual space to the textual semantic space. Such an interactive mapping more effectively narrows the distance between different modalities compared to simply projecting features into a common subspace.

It is noteworthy that the standard Transformer architecture typically utilizes Multi-Head Attention (MHA) to jointly attend to information from different representation subspaces at different positions, thereby avoiding the averaging effect that may occur in a Single-Head mechanism. In our design, we adopted a Single-Head attention mechanism for the feature fusion module. This choice was primarily driven by considerations of

computational efficiency and model compactness, particularly given that the features extracted by the CLIP backbone are already high-level, semantically aligned representations. While the Single-Head mechanism offers a favorable trade-off between retrieval accuracy and computational overhead for our specific FFCPH framework, we acknowledge that it may have inherent limitations in capturing diverse semantic interactions compared to MHA. Exploring lightweight multi-head strategies to further enhance feature fusion capability remains a promising direction for future work.

3.5. Proxy Hash

Proxy hashing is a method for learning hash codes, often applied to problems involving multimodal data, such as images and text. Compared to conventional pairwise and triplet losses, the proxy hashing approach introduces a set of learnable proxy vectors to simulate the semantic centers of different classes, enabling direct optimization of the relationships between proxies and samples. This reduces computational complexity while achieving competitive performance.

The core idea of proxy hashing is to introduce a set of learnable proxy vectors, which represent different semantic categories in a shared embedding space. The similarity relationships between these vectors and sample points serve as constraints for optimization. Specifically, as illustrated in Figure 3, four proxies P_1 , P_2 , P_3 and P_4 are embedded in the common space, while S_1 and S_2 are two sample points. If sample S_1 is associated with proxy P_2 and unrelated to other proxies, the proxy-based method will minimize the distance between S_1 and P_2 while maximizing the distances between S_1 and other semantically irrelevant proxies. Similarly, if sample S_2 is associated with proxy P_1 and unrelated to other proxies, the method will likewise minimize the distance between S_2 and P_1 while pushing S_2 away from unrelated proxies. This mechanism of attracting similar points and repelling dissimilar proxies effectively promotes clustering of cross-modal data in the shared semantic space, thereby laying the foundation for generating well-structured binary hash codes.

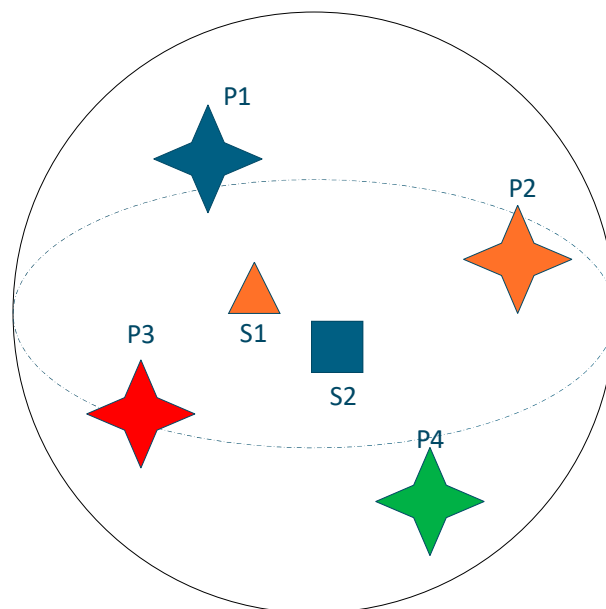


Figure 3. Diagram of proxy hash.

The initialization of proxy vectors is critical to ensuring training stability and convergence efficiency. In this work, the Kaiming initialization method is employed to randomly generate the initial values of each proxy vector. The specific formulation is as follows:

$$P_{c,v} \sim N(0, \sqrt{2/V}). \quad (8)$$

Here, $c \in [1, C]$, C denotes the total number of categories, $v \in [1, V]$, V represents the dimensionality of the proxy vectors for each category, and $P_{c,d}$ stands for the feature value of the v -th dimension of the c -th category. For a given tensor P , the Kaiming initialization method samples values from a normal distribution $N(0, \sqrt{2/V})$.

One advantage of this initialization method is that, in neural networks with activation functions, it maintains consistent variance distributions of inputs and outputs across layers, thereby helping to alleviate problems like gradient explosion and facilitating faster convergence during initial training. Furthermore, the proxy vectors can be jointly optimized with the encoding network during training and adaptively adjusted according to changes in the semantic distribution of multimodal samples.

3.6. Loss Function Design

To generate discriminative binary hash codes, this paper constructs a joint loss function composed of cross-modal proxy loss, cross-modal Irrelevant loss, and cross-modal consistency loss. This function aims to explicitly model multi-label semantic relationships through learnable proxy vectors, while leveraging consistency constraints to ensure the effectiveness of multimodal information fusion. Compared to traditional pairwise loss and triplet loss, the proposed loss function more effectively preserves similarity ranking among samples and alleviates semantic discrepancies across modalities. The subsequent sections will introduce these three loss functions separately. Additionally, the previously extracted image feature F_x , text feature F_y , and fused feature F_f will be denoted as x, y, f , respectively, for notational convenience.

3.6.1. Cross-Modal Proxy Loss

The fundamental principle of cross-modal proxy loss is to guide hash code learning through a set of learnable proxy vectors. For multi-label samples, which may be associated with multiple proxies, the loss function aims to minimize the distance between image features, text features, fused features, and semantically relevant proxies while maximizing their distance from semantically irrelevant proxies. Specifically, let $P = \{p_1, p_2, \dots, p_c\}$ denote the learnable proxies and C represent the number of categories. The cosine similarity between image features, text features, fused features, and the proxies is calculated as follows:

$$\cos_I = \frac{x \cdot p^T}{\|x\|_2 \cdot \|p\|_2}, \quad (9)$$

$$\cos_T = \frac{y \cdot p^T}{\|y\|_2 \cdot \|p\|_2}, \quad (10)$$

$$\cos_F = \frac{f \cdot p^T}{\|f\|_2 \cdot \|p\|_2}. \quad (11)$$

After obtaining the cosine similarity scores among the image features, text features, fused features, and the corresponding proxy vectors, the positive sample proxy loss for each feature type can be calculated as follows:

$$L_P^I = \frac{\sum_{i=1}^N \sum_{j=1}^C (1 - \cos_I(i, j)) \cdot l(i, j)}{P_{num}}, \quad (12)$$

$$L_P^T = \frac{\sum_{i=1}^N \sum_{j=1}^C (1 - \cos_T(i, j)) \cdot l(i, j)}{P_{num}}, \quad (13)$$

$$L_p^F = \frac{\sum_{i=1}^N \sum_{j=1}^C (1 - \cos_F(i, j)) \cdot l(i, j)}{P_{num}}. \quad (14)$$

where $\cos_I(i, j)$, $\cos_T(i, j)$, and $\cos_F(i, j)$ denote the cosine similarity between the i -th image, text, and fused features and the j -th proxy, respectively; $l(i, j)$ is the one-hot encoded label matrix indicating whether the i -th sample belongs to the j -th category; and P_{num} is the number of positive samples, i.e., the count of elements with a value of 1 in the label matrix. The positive sample proxy loss is designed to reduce the distance between image features, text features, fused features, and semantically relevant proxies.

Similarly, to increase the distance between samples and irrelevant proxies, the negative sample proxy loss is formulated as follows:

$$L_n^I = \frac{\sum_{i=1}^N \sum_{j=1}^C \max(\cos_I(i, j) - \theta, 0) \cdot (1 - l(i, j))}{N_{num}}, \quad (15)$$

$$L_n^T = \frac{\sum_{i=1}^N \sum_{j=1}^C \max(\cos_T(i, j) - \theta, 0) \cdot (1 - l(i, j))}{N_{num}}, \quad (16)$$

$$L_n^F = \frac{\sum_{i=1}^N \sum_{j=1}^C \max(\cos_F(i, j) - \theta, 0) \cdot (1 - l(i, j))}{N_{num}}, \quad (17)$$

where N_{num} is the number of negative samples, i.e., the count of elements with a value of 0 in the label matrix; θ is a threshold parameter used to control the boundary of similarity, where only cosine similarities exceeding θ will incur penalties. The negative sample proxy loss serves to maximize the distance between image features, text features, fused features, and semantically irrelevant proxies.

In summary, the total cross-modal proxy loss is obtained by summing the positive sample proxy loss and the negative sample proxy loss:

$$\mathcal{L}_P = \mathcal{L}_P^I + \mathcal{L}_P^T + \mathcal{L}_P^F + \mathcal{L}_n^I + \mathcal{L}_n^T + \mathcal{L}_n^F. \quad (18)$$

3.6.2. Cross-Modal Irrelevant Loss

To further enhance the discriminative ability of hash codes, it is necessary to explicitly maximize the distance between irrelevant sample pairs. Therefore, a cross-modal Irrelevant loss is designed. This work refers to the method of defining irrelevant pairs in HyP² Loss [28]: if the dot product of the label vectors of two samples is zero and the magnitudes of their respective label vectors are both greater than 1, the two samples are considered an irrelevant pair.

For cross-modal datasets, three types of irrelevant pairs may exist: irrelevant pairs between image samples, irrelevant pairs between text samples, and, correspondingly, irrelevant pairs between images and text. Thus, three types of Irrelevant losses are designed in this work: intra-modal Irrelevant loss (for image-image and text-text pairs) and inter-modal Irrelevant loss (for image-text pairs).

For image samples, the Irrelevant loss is formulated as follows:

$$L_I^x = \frac{\sum_{i=1}^{N'} \sum_{j=1}^{N'} \alpha \cdot \max(\cos(x_i, x_j) - \theta, 0) \cdot (1 - \cos(x_i, x_j))}{N_{zero}}, \quad (19)$$

where N is the number of irrelevant pairs, α is a hyperparameter, N_{zero} is the number of samples from different categories, and $\cos(x_i, x_j)$ is the cosine similarity between the i -th image feature and the j -th image feature.

Similarly, the Irrelevant loss for text samples is formulated as follows:

$$L_I^y = \frac{\sum_{i=1}^{N'} \sum_{j=1}^{N'} \alpha \cdot \max(\cos(y_i, y_j) - \theta, 0) \cdot (1 - \cos(y_i, y_j))}{N_{zero}}. \quad (20)$$

The Irrelevant loss between image and text samples is formulated as follows:

$$L_I^{xy} = \frac{\sum_{i=1}^{N'} \sum_{j=1}^{N'} \alpha \cdot \max(\cos_F(x_i, y_j) - \theta, 0) \cdot (1 - \cos(x_i, y_j))}{N_{zero}}. \quad (21)$$

In summary, the final cross-modal Irrelevant loss is obtained by summing the three Irrelevant losses:

$$L_I = L_I^x + L_I^y + L_I^{xy}. \quad (22)$$

3.6.3. Cross-Modal Consistency Loss

To guarantee that the fused feature $= f$ generated by the feature fusion module effectively preserves semantic information from both the original image and text modalities while promoting semantic alignment and matching between the two modalities, this paper designs a cross-modal consistency loss. This loss function constrains the distribution consistency between the fused feature and the original unimodal features in the semantic space, achieving complementary enhancement of multimodal information.

The specific formulation of the cross-modal consistency loss is as follows:

$$L_c = 1/2(MSE(f, x) + MSE(f, y)) \quad (23)$$

where MSE denotes the mean squared error, calculated as follows:

$$MSE(a, b) = \frac{1}{d} \sum_{j=1}^d (a_j - b_j)^2, \quad (24)$$

where d denotes the dimensionality of the feature vector.

By minimizing the cross-modal consistency loss, the fused features can maintain numerical distribution proximity to the original image and text features, ensuring no loss of semantic information. This provides stable and semantically rich multimodal input for subsequent hash code generation, thereby enhancing the accuracy and robustness of cross-modal retrieval.

3.6.4. Total Loss

By summing the cross-modal proxy loss, cross-modal Irrelevant loss, and cross-modal consistency loss, the final total loss can be obtained as:

$$L = L_P + \alpha L_I + L_C, \quad (25)$$

where L_P denotes the cross-modal proxy loss, L_I represents the cross-modal Irrelevant loss, and L_C corresponds to the cross-modal consistency loss. The hyperparameter α controls the influence of the cross-modal Irrelevant loss on the total loss.

The cross-modal proxy loss establishes semantic associations between samples and category proxies to enhance the discriminability of hash codes. It aligns heterogeneous samples with their corresponding category-level proxies in a shared space, improving both intra-class compactness and inter-class separability. Concurrently, the cross-modal irrelevant loss increases the distance between irrelevant pairs through margin constraints, effectively reducing false correlations. Lastly, the cross-modal consistency loss preserves semantic and structural compatibility between original and fused features, ensuring the

fused representations retain critical multimodal information while enabling effective hash space alignment. This loss provides a robust feature foundation that supports the performance of other objective functions, ultimately improving cross-modal retrieval accuracy and efficiency.

4. Experimental Results and Analysis

4.1. Experimental Setup

Experiments were performed on three widely adopted datasets: MIRFLICKR-25K, NUS-WIDE, and MS COCO. These datasets are commonly employed in a variety of vision-language research tasks, such as image classification, object recognition, and cross-modal retrieval between images and text.

The MIRFLICKR-25K dataset consists of 24,581 images, each associated with a corresponding textual description. The images are annotated with 24 distinct labels, corresponding to 24 semantic categories.

NUS-WIDE is a larger dataset comprising 269,648 image-text pairs from 81 categories. Owing to the severe class imbalance, with many categories containing only a limited number of samples, 21 common categories were selected in this study, which consists of 195,834 image-text pairs.

MS COCO is a dataset designed to advance research in various computer vision domains, particularly object detection and image captioning. It includes 40,504 validation images and 82,785 training images, each paired with textual descriptions. The dataset covers 80 object categories.

All experiments were performed on an NVIDIA RTX 3090Ti GPU using the Adam optimizer for parameter updates. The hyperparameter α was set to 0.8, the initial learning rate was 0.001, and the batch size was uniformly set to 128.

4.2. Baseline Methods and Evaluation Metrics

To evaluate the efficacy of the proposed FFCPH approach, ten deep cross-modal hashing techniques were employed as baseline methods for comparative analysis, including DCMH [29], SSAH [30], AGAH [31], DADH [32], MSSPQ [14], DCHMT [13], MIAN [25], DAPH [33], MCCMR [18], and DNPH [34]. All compared methods utilize deep neural networks as their backbone frameworks, and their source codes and data are mostly provided in the corresponding publications.

To ensure comprehensive and credible performance evaluation, three commonly used evaluation metrics from information retrieval were utilized: mean Average Precision (MAP) [35], Precision-Recall (PR) curves, and TopN-Precision (TopN-P) curves.

MAP, an abbreviation for mean Average Precision, serves as a standard metric for assessing comprehensive effectiveness. A higher MAP value reflects better retrieval accuracy of the deep hashing method.

The formula for MAP is as follows:

$$MAP = \frac{1}{Q} \sum_{i=1}^N A P_i. \quad (26)$$

The Precision-Recall (PR) curve visually represents the trade-off between precision and recall across different retrieval thresholds, such as Hamming distance thresholds or similarity thresholds. By traversing different thresholds, a series of data points is obtained, and connecting these points yields the PR curve.

The TopN curve serves as a graphical tool to assess the effectiveness of a ranking model by focusing on its top-N returned items. TopN precision quantifies the ratio of relevant instances within these top-N retrieved results. By plotting the precision curve as

N increases from 1, the model's performance at different retrieval depths can be intuitively assessed. A steeper initial slope and a consistently higher overall curve indicate better performance of the model in the top- N results.

4.3. Experimental Results Comparison

To further substantiate the thoroughness and efficacy of the proposed approach, cross-modal retrieval trials were performed across three datasets employing hash codes with lengths of 16, 32, and 64. As shown in Table 1, FFCPH demonstrates the best retrieval performance on all three widely used datasets—MIRFLICKR-25K, NUS-WIDE, and MS COCO—in cross-modal retrieval applications.

Table 1. The mAP results on three benchmark datasets at various code lengths.

Task	Method	MIRFLICKR-25k			NUS-WIDE			MS COCO		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
I2T	DCMH	0.7687	0.7736	0.7797	0.5379	0.5513	0.5617	0.5399	0.5444	0.5627
	SSAH	0.8079	0.8129	0.8220	0.6032	0.6058	0.6095	0.5411	0.4855	0.5395
	AGAH	0.7284	0.7217	0.7195	0.3945	0.4107	0.4258	0.5501	0.5515	0.5518
	DADH	0.8098	0.8162	0.8193	0.6350	0.6568	0.6546	0.5952	0.6118	0.6237
	MSSPQ	0.7868	0.8011	0.8172	0.6346	0.6478	0.6615	0.5710	0.5881	0.5862
	DCHMT	0.8201	0.8253	0.8222	0.6596	0.6706	0.6863	0.6309	0.6216	0.6553
	MIAN	0.8123	0.8220	0.8355	0.6130	0.5894	0.6057	0.5350	0.5579	0.5404
	DAPH	0.7939	0.8071	0.8146	0.6642	0.6835	0.6925	0.6344	0.7030	0.7300
	MCCMR	0.8016	0.8156	0.8177	0.6535	0.6745	0.6753	0.6522	0.6876	0.7273
	DNPH	0.8108	0.8269	0.8374	0.6689	0.6811	0.6939	0.6678	0.7082	0.7335
	FFCPH	0.8185	0.8455	0.8658	0.6837	0.7062	0.7181	0.7128	0.7561	0.7782
T2I	DCMH	0.7857	0.7998	0.8029	0.5747	0.5810	0.5853	0.5271	0.5424	0.5450
	SSAH	0.8089	0.8127	0.8017	0.6011	0.6058	0.6167	0.4901	0.4798	0.5053
	AGAH	0.7082	0.7182	0.7344	0.4344	0.3980	0.4382	0.5012	0.5146	0.5191
	DADH	0.8019	0.8101	0.8137	0.6111	0.6182	0.6281	0.5649	0.5790	0.5870
	MSSPQ	0.7946	0.7885	0.8022	0.6312	0.6631	0.6882	0.5472	0.5630	0.5985
	DCHMT	0.7982	0.8048	0.8031	0.6761	0.6837	0.6943	0.6241	0.6212	0.6486
	MIAN	0.8058	0.8151	0.8182	0.6365	0.6312	0.6416	0.5489	0.5493	0.5365
	DAPH	0.7949	0.8134	0.8212	0.6611	0.6813	0.7016	0.6307	0.6870	0.7180
	MCCMR	0.8023	0.8204	0.8244	0.6598	0.6794	0.6842	0.6585	0.6944	0.7396
	DNPH	0.8015	0.8176	0.8236	0.6871	0.6994	0.7182	0.6648	0.7173	0.7459
	FFCPH	0.8034	0.8219	0.8413	0.7053	0.7184	0.7283	0.7112	0.7596	0.7811

For the I→T task on MIRFLICKR-25K, FFCPH slightly underperforms DCHMT at 16 bits but outperforms DNPH by 1.8% and 2.8% at 32 and 64 bits, respectively. In the T→I task, it also trails SSAH at 16 bits, while surpassing MCCMR by 0.2% and 1.7% at 32 and 64 bits, respectively.

For the I→T task on the NUS-WIDE dataset, FFCPH achieves improvements of 1.5%, 2.3%, and 2.4% over DNPH, DAPH, and DNPH at 16, 32, and 64 bits, respectively. In the T→I task, it outperforms DNPH by 1.8%, 1.9%, and 1.0% at the three bit lengths.

For the I→T task on the MS COCO dataset, FFCPH surpasses DNPH by 4.5%, 4.8%, and 4.5% at 16, 32, and 64 bits, respectively. In the T→I task, it outperforms DNPH by 4.6%, 4.2%, and 3.5% at the corresponding bit lengths.

These results show that the performance improvement of the proposed method is more pronounced on MS COCO than on NUS-WIDE and MIRFLICKR-25K. This is due to the fact that MS COCO contains 80 categories, significantly more than the 21 in NUS-WIDE and 24 in MIRFLICKR-25K. While previous methods are highly sensitive to the number of categories, the proposed approach demonstrates greater robustness and broader applicability.

Overall, FFCPH performs best on all three datasets, further confirming its stability and generalization ability under varying hash code dimensions.

Figures 4 and 5 present a comparison of top-N curves and PR curves between the FFCPH method and several mainstream approaches at a 32-bit encoding length on three datasets.

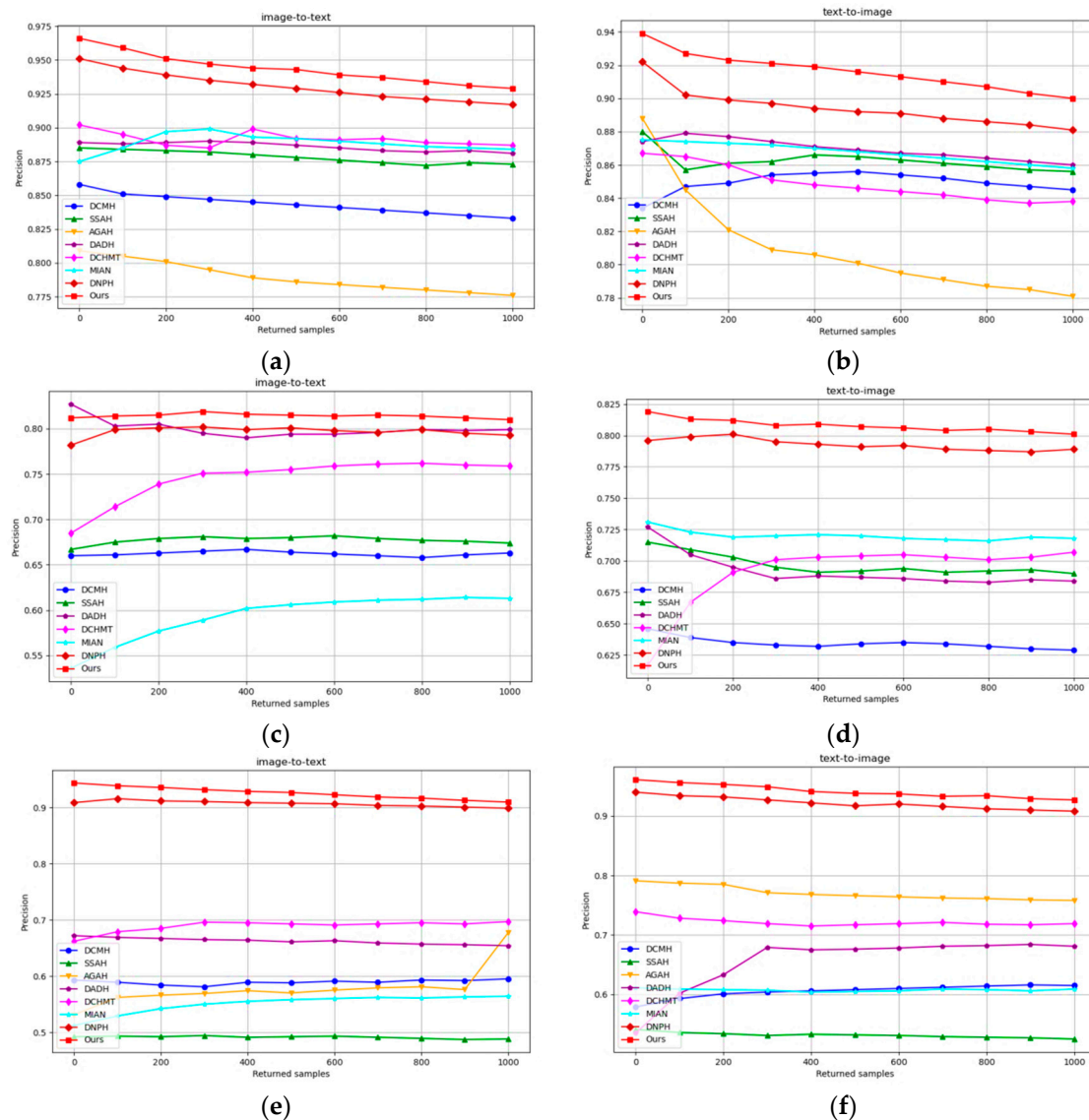


Figure 4. Results of TopN-Precision curves with 32 bits on three benchmark datasets: (a) MIRFLICKR-25k I2T; (b) MIRFLICKR-25k T2I; (c) NUS-WIDE I2T; (d) NUS-WIDE T2I; (e) MS COCO I2T; (f) MS COCO T2I.

To further observe the retrieval performance of FFCPH, Top-N Precision curves were generated for three datasets, as shown in Figure 4.

For the MIRFLICKR-25K and MS COCO datasets, FFCPH consistently demonstrates superior performance compared to other methods at all retrieval points in both I→T and T→I tasks, demonstrating its strong semantic modeling capability and retrieval stability. In the I→T task on NUS-WIDE, FFCPH initially exhibits lower Precision than DADH but surpasses DADH when N exceeds 100, maintaining its leading performance thereafter, indicating superior robustness compared to other methods.

Overall, FFCPH exhibits excellent performance across all tasks on the three mainstream datasets. This performance shows the model's exceptional adaptability and reliable effectiveness in learning and transferring semantic associations across different modalities.

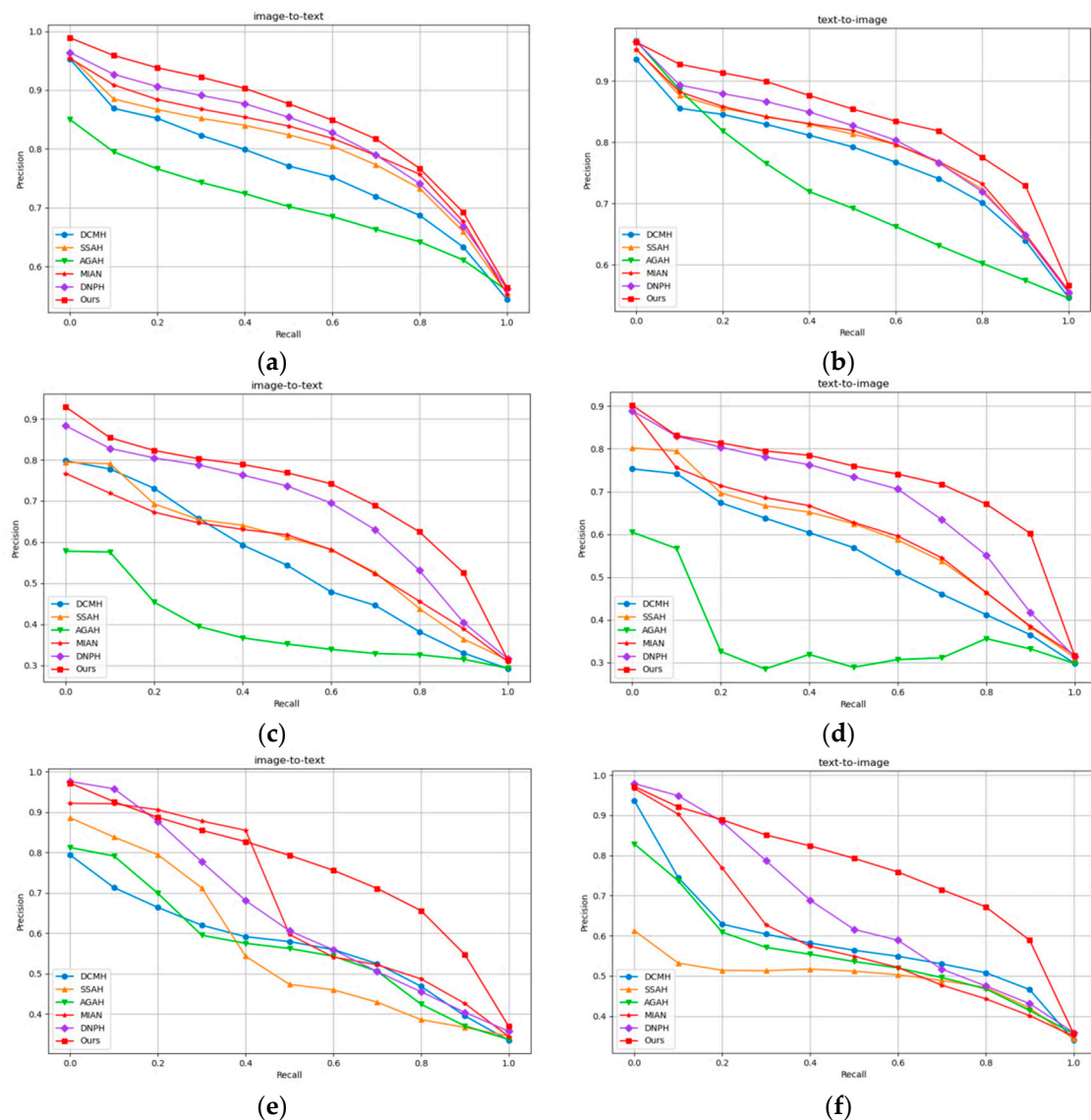


Figure 5. Results of Precision-Recall curves with 32 bits on three benchmark datasets: (a) MIRFLICKR-25k I2T; (b) MIRFLICKR-25k T2I; (c) NUS-WIDE I2T; (d) NUS-WIDE T2I; (e) MS COCO I2T; (f) MS COCO T2I.

To further observe the retrieval performance of FFCPH, PR curves were generated for three datasets, as shown in Figure 5.

On the MIRFLICKR-25K dataset, FFCPH achieved the best results in both I→T and T→I tasks, consistently outperforming other methods across all retrieval points. This demonstrates FFCPH's ability to stably retrieve more relevant samples.

On the NUS-WIDE dataset, the overall precision of FFCPH in both tasks was slightly lower than that on MIRFLICKR-25K. Nevertheless, it maintained optimal performance and stability across the entire recall range.

On the MS COCO dataset, FFCPH exhibited slightly lower precision than DNPH and MIAN at low recall rates (Recall < 0.4) in the I→T task but consistently outperformed other methods when Recall > 0.4, with significantly higher precision. In the T→I task, FFCPH showed marginally lower precision than DNPH when Recall < 0.2 but maintained

superiority thereafter. These results indicate the high robustness and strong generalization capacity of the FFCPH model.

In summary, FFCPH demonstrates high precision-recall balance across different datasets and tasks, effectively controlling false positives while ensuring sufficiently high recall rates.

4.4. Ablation Study

The FFCPH framework processes raw image and text data through a feature extraction network to obtain image features and text features separately. These features are then integrated via a feature fusion module to generate fused features that exhibit both discriminative power and robustness. Finally, hash function learning is accomplished under the guidance of a joint loss function composed of cross-modal proxy loss, cross-modal Irrelevant loss, and cross-modal consistency loss.

In this section, to validate the effectiveness of individual components in FFCPH, we design four variants of the model to evaluate the contribution of each component to overall performance. The variants are constructed as follows:

FFCPH-1: Replaces the CLIP-based image encoder with a ResNet-18 network.

FFCPH-2: Removes the feature fusion module and the cross-modal consistency loss. Only the cross-modal proxy loss and cross-modal Irrelevant loss are used for optimization. In the cross-modal proxy loss, only image and text features are considered—fused features are excluded, i.e., $L = L_P + \alpha L_I$.

FFCPH-3: Removes the cross-modal Irrelevant loss and optimizes parameters using only the cross-modal proxy loss and cross-modal consistency loss, i.e., $L = L_P + L_C$.

FFCPH-4: Removes the cross-modal proxy loss and optimizes the model using only the cross-modal Irrelevant loss and cross-modal consistency loss, i.e., $L = \alpha L_I + L_C$.

To evaluate the impact of individual components in the FFCPH model on cross-modal retrieval performance, experiments were carried out using four model variants across three datasets, focusing on the I→T task. The corresponding results are provided in Table 2.

Table 2. The mAP of ablation experiments on three datasets at various code lengths.

Task	Method	MIRFLICKR-25k			NUS-WIDE			MS COCO		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
I2T	FFCPH-1	0.7982	0.8195	0.8235	0.6689	0.6853	0.6914	0.6573	0.7012	0.7152
	FFCPH-2	0.8057	0.8397	0.8594	0.6791	0.6987	0.7128	0.6948	0.7433	0.7692
	FFCPH-3	0.8159	0.8373	0.8547	0.6781	0.7014	0.7158	0.7014	0.7470	0.7684
	FFCPH-4	0.8113	0.8331	0.8514	0.6720	0.6985	0.7087	0.6973	0.7329	0.7592
	FFCPH	0.8185	0.8455	0.8658	0.6837	0.7062	0.7181	0.7128	0.7561	0.7782

FFCPH-1, a variant that replaces the CLIP image encoder with ResNet18, exhibits a significant decline in mAP across all three datasets for various hash code lengths. Specifically, the average mAP decreases by 3.0% on MIRFLICKR-25K, 2.1% on NUS-WIDE, and 5.8% on MS COCO. This indicates that the CLIP image encoder, by modeling long-range visual dependencies, extracts higher-quality semantic features compared to ResNet18.

FFCPH-2, which removes the feature fusion module and cross-modal consistency loss and optimizes the model solely with the cross-modal proxy loss and cross-modal Irrelevant loss, shows a marked reduction in mAP across all datasets and hash code lengths. The average mAP drops by 0.8% on MIRFLICKR-25K, 0.6% on NUS-WIDE, and 1.3% on MS COCO. This demonstrates that the feature fusion module and cross-modal consistency loss ensure the effectiveness and modality compatibility of fused features, providing stable and

semantically rich multimodal inputs for subsequent hash code generation, and thereby enhancing retrieval accuracy and robustness.

FFCPH-3, which eliminates the cross-modal Irrelevant loss and optimizes parameters using only the cross-modal proxy loss and cross-modal consistency loss, also experiences a noticeable decline in mAP. The average reduction is 0.7% on MIRFLICKR-25K, 0.4% on NUS-WIDE, and 1.0% on MS COCO. This confirms that the cross-modal Irrelevant loss captures fine-grained semantic relationships between data samples and disentangles irrelevant pairs.

FFCPH-4, a variant that removes the cross-modal proxy loss and optimizes the model using only the cross-modal Irrelevant loss and cross-modal consistency loss, exhibits a significant decrease in mAP. The average reduction is 1.1% on MIRFLICKR-25K, 1.0% on NUS-WIDE, and 1.9% on MS COCO. This indicates that the cross-modal proxy loss ensures that relevant data-proxy pairs are embedded closely while pushing apart irrelevant pairs.

In summary, the ablation study validates the contributions of all modules in FFCPH to overall retrieval performance. The cross-modal proxy loss establishes semantic associations between samples and category proxies, ensuring the discriminability of hash codes. The cross-modal Irrelevant loss explicitly separates irrelevant sample pairs, enhancing the distinctiveness of hash codes. The feature fusion module and cross-modal consistency loss focus on guaranteeing the effectiveness and modality compatibility of fused features, providing a high-quality feature representation foundation for the aforementioned losses. These modules collaboratively contribute to the superiority of FFCPH in cross-modal retrieval.

4.5. Parameter Sensitivity Analysis

This section primarily conducts parameter analysis for α . Figure 6 illustrates the variation in mAP across different tasks and hash code lengths on the MIRFLICKR-25k dataset when α takes different values. The values of α range from 0.1 to 0.9 and are used to evaluate the impact of cross-modal uncorrelated loss on retrieval performance. Specific results are shown in Figure 6.

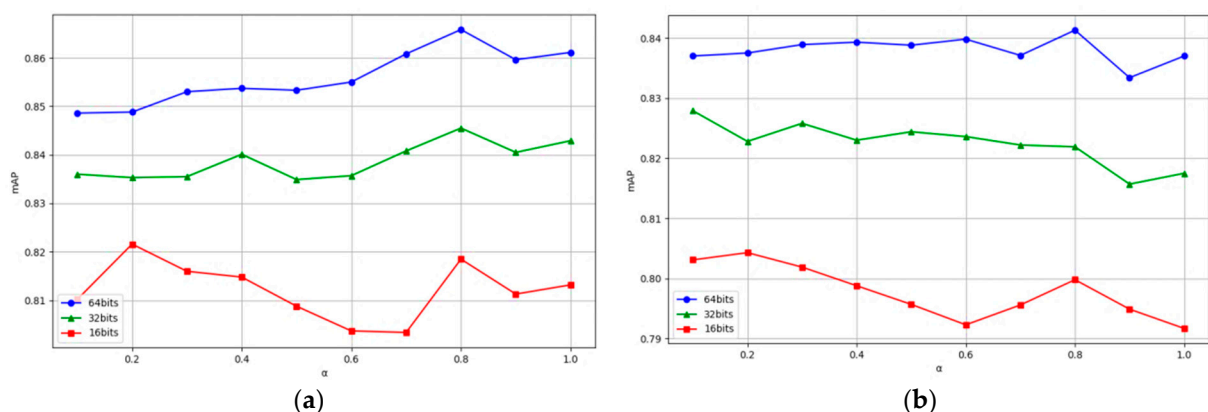


Figure 6. Parameter Analysis on MIRFLICKR-25k at various code lengths: (a) MIRFLICKR-25k I2T; (b) MIRFLICKR-25k T2I.

As shown in Figure 6, in the I→T task, the model achieves optimal performance at $\alpha = 0.2$ for 16-bit, $\alpha = 0.8$ for 32-bit, and $\alpha = 0.8$ for 64-bit.

In the T→I task, the model achieves optimal performance at $\alpha = 0.2$ for 16-bit, $\alpha = 0.1$ for 32-bit, and $\alpha = 0.8$ for 64-bit.

Based on comparative analysis, we set the hyperparameter $\alpha = 0.8$ in our experiments.

5. Threats to Validity and Reliability

In this section, we discuss the potential internal and external threats to the validity and reliability of the proposed FFCPH framework to provide a comprehensive perspective on its experimental findings.

5.1. Internal Threats

Limitations in Architectural Design. An important internal threat arises from the architectural constraints of the current feature fusion module. While effective in facilitating local inter-modal interactions, the module primarily focuses on local operations. This design may fall short in capturing long-range contextual information and modeling deeper semantic relationships, potentially limiting the model's ability to establish global associations across distant regions. Consequently, the retrieval performance might be compromised in scenarios requiring complex global reasoning.

Sensitivity to Dataset Scale. Another internal threat pertains to the influence of dataset size on the model's generalization. We observed that the performance gains of our method are less pronounced on smaller-scale datasets, such as MIRFLICKR-25K, compared to large-scale benchmarks like MS COCO. This suggests that the framework's ability to learn hash codes is sensitive to the diversity and volume of training data. In data-scarce environments, the model may risk overfitting or failing to capture the complete semantic distribution, which affects the stability of the results across different datasets.

5.2. External Threats

Dependency on Pre-trained Backbones. The reliability of our method is significantly influenced by the choice of the feature extraction backbone. Our approach relies heavily on the pre-trained knowledge within the CLIP model. As demonstrated in the ablation study, replacing the CLIP backbone with a simpler encoder (e.g., ResNet-18) leads to a significant degradation in retrieval performance. This indicates that the generalization capability of FFCPH is constrained by the quality and scope of the pre-trained features. If applied to domains where CLIP's features are suboptimal or where the domain shift is substantial, the model's performance may exhibit brittleness.

Hyperparameter Configuration. Like most deep learning approaches, the performance of FFCPH is sensitive to hyperparameter settings, including the learning rate, batch size, and the weights of different loss terms. Although we have conducted extensive grid searches to identify the optimal configuration for the reported datasets, these specific parameters may not be universally optimal for all potential applications. Variations in these settings could lead to fluctuations in retrieval accuracy, necessitating careful retuning when deploying the model in new scenarios.

6. Conclusions

To overcome the constraints associated with conventional deep cross-modal hashing approaches, such as their inability to establish consistent mappings across modalities and bridge the semantic gap in cross-modal data—resulting in semantic information loss and incomplete understanding—this paper proposes and validates a Feature Fusion-based Cross-modal Proxy Hashing (FFCPH) retrieval method. The approach inputs raw image and text data into feature extraction networks to extract image and text features separately, then integrates multimodal semantic information via a feature fusion module to generate discriminative cross-modal fused features. Finally, it learns to generate hash codes using a combined loss function including cross-modal proxy loss, cross-modal Irrelevant loss, and cross-modal consistency loss.

A series of comprehensive experiments was performed using three widely adopted benchmark datasets. The results show that FFCPH effectively bridges the heterogeneity gap across modalities, enabling the hashing network to learn comprehensive semantic representations from cross-modal data and achieve accurate matching and retrieval on multimodal datasets.

Despite the promising results, this study acknowledges certain limitations regarding architectural design, reliance on pre-trained features, and sensitivity to dataset scale, as discussed in Section 5 (Threats to Validity and Reliability). Specifically, the current fusion module faces challenges in capturing long-range dependencies, and the framework's performance is constrained by the quality of external pre-trained encoders and data scale variations. To address these issues and further enhance the model's capability, future research will focus on three main directions: (1) incorporating Graph Convolutional Networks (GCNs) and wider cross-attention layers to better model global semantic relationships; (2) investigating domain adaptation techniques to reduce the dependency on specific pre-trained backbones like CLIP and improve robustness; and (3) integrating semantic data augmentation and few-shot learning paradigms to ensure stable performance even on smaller-scale datasets.

Author Contributions: Conceptualization, H.L. and Y.Z.; methodology, H.L. and Y.Z.; software, H.L.; validation, H.L. and Y.Z.; formal analysis, H.L. and Y.Z.; investigation, H.L.; resources, Y.Z.; data curation, H.L. and Y.Z.; writing—original draft preparation, H.L. and Y.Z.; writing—review and editing, H.L. and Y.Z.; visualization, H.L. and Y.Z.; supervision, Y.Z.; project administration, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets used in this study are publicly available from their respective official websites, and their acquisition methods are detailed as follows: <https://press.liacs.nl/mirflickr/> (accessed on 1 January 2025). <https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html> (accessed on 1 January 2025). <https://cocodataset.org/> (accessed on 1 January 2025).

Acknowledgments: The authors employed ChatGPT-4o for language polishing and readability enhancement. All AI-generated content was meticulously reviewed, verified, and edited by the authors, who bear sole responsibility for the work's accuracy and integrity.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cao, M.; Li, S.; Li, J.; Nie, L.; Zhang, M. Image-text retrieval: A survey on recent research and development. *arXiv* **2022**, arXiv:2203.14713. [CrossRef]
2. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]
3. Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G.R.; Levy, R.; Vasconcelos, N. A new approach to cross-modal multimedia retrieval. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 251–260. [CrossRef]
4. Liu, H.; Wang, R.; Shan, S.; Chen, X. Deep supervised hashing for fast image retrieval. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2064–2072. [CrossRef]
5. Wang, K.; Yin, Q.; Wang, W.; Wu, S.; Wang, L. A comprehensive survey on cross-modal retrieval. *arXiv* **2016**, arXiv:1607.06215. [CrossRef]
6. Peng, Y.; Huang, X.; Zhao, Y. An overview of cross-media retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 2372–2385. [CrossRef]

7. Lin, Z.; Ding, G.; Hu, M.; Wang, J. Semantics-preserving hashing for cross-view retrieval. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3864–3872. [\[CrossRef\]](#)
8. Zhang, D.; Li, W.-J. Large-scale supervised multimodal hashing. In Proceedings of the AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014; pp. 2177–2183.
9. Xu, X.; Shen, F.; Yang, Y.; Shen, H.T.; Li, X. Learning discriminative binary codes. *IEEE Trans. Image Process.* **2017**, *26*, 2494–2507. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Shi, Y.; You, X.; Zheng, F.; Wang, S.; Peng, Q. Equally-guided discriminative hashing. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 4767–4773.
11. Lin, Q.; Cao, W.; He, Z.; He, Z. Semantic deep cross-modal hashing. *Neurocomputing* **2020**, *396*, 113–122. [\[CrossRef\]](#)
12. Liong, V.E.; Lu, J.; Tan, Y.P.; Zhou, J. Cross-modal deep variational hashing. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4077–4085. [\[CrossRef\]](#)
13. Tu, J.; Liu, X.; Lin, Z.; Hong, R.; Wang, M. Differentiable cross-modal hashing via multimodal transformers. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 453–461. [\[CrossRef\]](#)
14. Zhu, L.; Cai, L.; Song, J.; Zhu, X.; Zhang, C.; Zhang, S. MSSPQ: Multiple Semantic Structure-Preserving Quantization for Cross-Modal Retrieval. In Proceedings of the 2022 International Conference on Multimedia Retrieval, Newark, NJ, USA, 27–30 June 2022; pp. 631–638. [\[CrossRef\]](#)
15. Bai, K.; Gao, P.; Chen, K.; Nie, X.; Li, S.; Li, B. Mutual compromised multi-feature fusion. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME), Niagara Falls, ON, Canada, 15–19 July 2024.
16. Weiss, Y.; Torralba, A.; Fergus, R. Spectral hashing. In Proceedings of the Advances in Neural Information Processing Systems 21, Vancouver, BC, Canada, 8–11 December 2008.
17. Kumar, S.; Udupa, R. Learning hash functions for cross-view similarity search. In Proceedings of the IJCAI Proceedings-International Joint Conference on Artificial Intelligence, Catalonia, Spain, 16–22 July 2011.
18. Song, J.; Yang, Y.; Huang, Z.; Shen, H.T. Inter-media hashing. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 22–27 June 2013; pp. 785–796. [\[CrossRef\]](#)
19. Zhou, J.; Ding, G.; Guo, Y. Latent semantic sparse hashing. In Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, Gold Coast, Australia, 6–11 July 2014; pp. 415–424. [\[CrossRef\]](#)
20. Su, S.; Zhong, Z.; Zhang, C. Deep joint-semantics reconstructing hashing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3027–3035. [\[CrossRef\]](#)
21. Hu, H.; Xie, L.; Hong, R.; Tian, Q. Unsupervised knowledge distillation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3123–3132. [\[CrossRef\]](#)
22. Zhang, P.F.; Li, Y.; Huang, Z.; Xu, X.-S. Aggregation-based graph convolutional hashing. *IEEE Trans. Multimed.* **2022**, *24*, 466–479. [\[CrossRef\]](#)
23. Yu, J.; Zhou, H.; Zhan, Y.; Tao, D. Deep graph-neighbor coherence preserving network. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 4626–4634. [\[CrossRef\]](#)
24. Zhu, L.; Wu, X.; Li, J.; Zhang, Z.; Guan, W.; Shen, H.T. Correlation-identity reconstruction hashing. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 8838–8851. [\[CrossRef\]](#)
25. Zhang, Z.; Luo, H.; Zhu, L.; Lu, G.; Shen, H.T. Modality-invariant asymmetric networks. *IEEE Trans. Knowl. Data Eng.* **2022**, *35*, 5091–5104. [\[CrossRef\]](#)
26. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021.
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
28. Xu, C.; Chai, Z.; Xu, Z.; Yuan, C.; Fan, Y.; Wang, J. Hyp2 loss. In Proceedings of the 30th ACM International Conference on Multimedia, Lisbon, Portugal, 10–14 October 2022; pp. 3173–3184. [\[CrossRef\]](#)
29. Jiang, Q.Y.; Li, W.J. Deep cross-modal hashing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3232–3240. [\[CrossRef\]](#)
30. Li, C.; Deng, C.; Li, N.; Liu, W.; Gao, X.; Tao, D. Self-supervised adversarial hashing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 July 2018; pp. 4242–4251. [\[CrossRef\]](#)
31. Gu, W.; Gu, X.; Gu, J.; Li, B.; Xiong, Z.; Wang, W. Adversary guided asymmetric hashing. In Proceedings of the 2019 International Conference on Multimedia Retrieval, Ottawa, ON, Canada, 10–13 June 2019; pp. 159–167. [\[CrossRef\]](#)
32. Bai, C.; Zeng, C.; Ma, Q.; Zhang, J.; Chen, S. Deep adversarial discrete hashing. In Proceedings of the 2020 International Conference on Multimedia Retrieval, Dublin, Ireland, 8–11 June 2020; pp. 525–531. [\[CrossRef\]](#)
33. Tu, R.C.; Mao, X.L.; Ji, W.; Wei, W.; Huang, H. Data-aware proxy hashing. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, Taipei, Taiwan, 23–27 July 2023; pp. 686–696. [\[CrossRef\]](#)

34. Huo, Y.; Qian, Q.; Dai, J.; Zhang, W.; Huang, L.; Wang, C. Deep neighborhood-aware proxy hashing. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 1–23. [[CrossRef](#)]
35. Wang, J.; Liu, W.; Kumar, S.; Chang, S.F. Learning to hash for indexing big data—A survey. *Proc. IEEE* **2016**, *104*, 34–57. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.