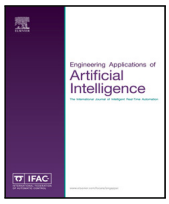
Engineering Applications of Artificial Intelligence 152 (2025) 110777

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Covariance Attention Guidance Mamba Hashing for cross-modal retrieval

Gang Wang, Shuli Cheng *, Anyu Du, Qiang Zou

School of Computer Science and Technology, Xinjiang University, Urumqi, 830046, Xinjiang, China

ARTICLE INFO

Keywords:
Artificial intelligence
Multimedia technology
Cross-modal hashing
Multi-feature fusion
Covariance attention

ABSTRACT

With the rapid development of artificial intelligence and multimedia technology, cross-modal hashing (CMH) has been widely applied in multimedia retrieval, recommendation systems, and large-scale data search due to its efficient query processing and low storage requirements, and has become a pivotal research area in both academia and industry. However, existing CMH algorithms fall short in exploiting the potential inter-modal correlations, leading to considerable semantic gaps. To overcome this issue, this paper proposes an innovative CMH framework called Covariance Attention Guidance Mamba Hashing (CAGMH) for Cross-Modal Retrieval. The framework enables deeper semantic alignment between modalities through a novel multi-feature fusion mechanism. This mechanism narrows the semantic gap and enhances the expressive power of each modality. Specifically, CAGMH exploits the distributional properties of covariance to optimize hash code generation and combined with the Mamba strategy to further improve cross-modal retrieval robustness. In addition, we design a novel loss function computation strategy that combines modal correlation with semantic consistency to optimize the model's convergence and generalization ability. Experiments on four public benchmark datasets show that CAGMH surpasses state-of-the-art CMH methods, offering improved accuracy and efficiency in large-scale cross-modal similarity search. The corresponding code is available at https://github.com/Rooike111/CAGMH.

1. Introduction

With the rapid development of multimedia technology, a large amount of high-dimensional data with different structures has been generated on the Internet (Xie et al., 2022). These data vary in their modalities, such as text, images and video, and are semantically related, driving the advancement of cross-modal retrieval (CMH) techniques. CMH maps data from diverse modalities into a common feature space, enabling similarity searches across modalities to meet user demands for quick information access in large-scale data environments (Xie et al., 2021). For example, users can search for related text using an image or find product images through descriptive text. In this context, hashing has become a key technology in cross-modal retrieval due to its efficiency in query processing and compact storage requirements.

Traditional CMH methods project multimodal data into a shared subspace for feature comparison. However, with the advent of 5G, the growing data dimensions and volumes have made this approach inefficient and costly (Zhang et al., 2024; Cheng et al., 2024). To overcome this, deep learning has been incorporated into CMH, significantly improving feature learning efficiency and reducing retrieval costs.

In comparison to traditional CMH techniques, deep cross-modal hashing retrieval presents significant advantages. By leveraging deep neural networks, it automatically extracts rich and hierarchical features directly from raw data, eliminating the need for manual feature engineering (Chao et al., 2023). This approach generates more precise and semantically meaningful hash codes, which enhance retrieval accuracy. Deep cross-modal hashing techniques have demonstrated extensive applications in multimedia retrieval, recommender systems, and large-scale data search. However, their potential extends far beyond these domains. Integrating cross-modal techniques with other fields, such as object detection and re-identification (Khan et al., 2024a, 2025, 2024b), holds significant promise for advancing scientific research and fostering innovation.

CMH method transforms high-dimensional data into compact, binary representations, which are then used to compute similarity through Hamming distance for retrieval (Irie et al., 2015; Yang et al., 2023). These methods can be classified based on their function into supervised and unsupervised categories: supervised methods use labels to enhance model accuracy, while unsupervised methods rely on data topology for searching. Although supervised methods generally offer higher accuracy, unsupervised methods are valuable when labeled data is scarce.

Despite significant progress in supervised CMH, particularly after the introduction of Contrastive Language-Image Pre-training (CLIP), proposed by OpenAI in 2021 (Radford et al., 2021), several limita-

* Corresponding author.
E-mail addresses: wangg@stu.xju.edu.cn (G. Wang), cslxju@xju.edu.cn (S. Cheng), anydxju@xju.edu.cn (A. Du), zouq@stu.xju.edu.cn (Q. Zou).

https://doi.org/10.1016/j.engappai.2025.110777
Received 28 September 2024; Received in revised form 17 March 2025; Accepted 2 April 2025
Available online 17 April 2025

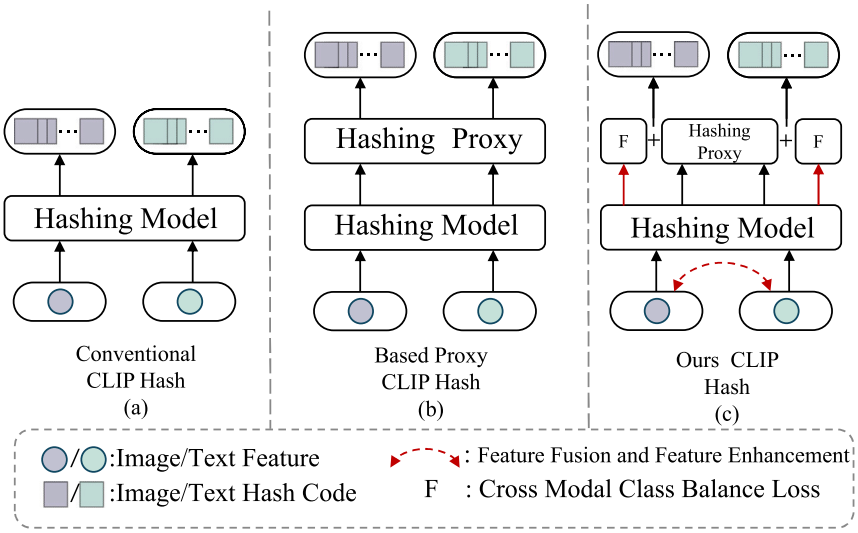


Fig. 1. Our framework VS. existing frameworks, where Fig. (a) shows the traditional CMH method, while Fig. (b) presents the proxy-based CMH method. In Fig. (c), the proposed CAGMH framework builds on Fig. (b) by integrating a feature fusion and enhancement module, as well as an improved loss calculation method, to enhance inter-modal information exchange and robustness.

tions remain: (1) Data from different modalities often carries distinct information, leading to a "semantic gap" that existing models struggle to bridge, resulting in incomplete or inaccurate cross-modal associations. (2) Although some proxy-based CMH methods partially narrow the semantic gap between different modalities, the performance and robustness of these methods are still challenged when dealing with category imbalance or noise-like problems.

To address this issue, we propose a novel framework, Covariance Attention Guidance Mamba Hashing (CAGMH) for Cross-Modal Retrieval. The model framework is illustrated in Fig. 1. This framework not only leverages intra-modal features but also deeply explores inter-modal connections, surpassing the traditional approach of merely sharing attributes. CAGMH enhances retrieval accuracy and accelerates model training through a newly proposed loss function. The primary contributions are as follows:

- Firstly, we propose an end-to-end learning framework based on CLIP, named CAGMH. This framework incorporates a novel Covariance Attention Guidance Mamba Module, which efficiently mines and extracts information across both cross-modal and intra-modal domains, thereby generating semantically aligned and efficient hash codes.
- Additionally, through a balanced approach, we enable the efficient computation of hash loss within the semantic space. This method fully leverages the complementarity between modalities, effectively addressing the issue of imbalance and optimizing performance across multiple dimensions.
- Finally, to evaluate the effectiveness of the proposed method, extensive experiments were conducted on four publicly available datasets designed for image–text retrieval. The experimental results confirm the efficiency of CAGMH in retrieval tasks. Our approach consistently achieves superior performance, as measured by the Mean Average Precision (MAP) metric, and the Precision–Recall curves further highlight its advantage in CMH.

The remainder of this paper is structured as follows: Section 2 reviews representative CMH methods, unsupervised CMH methods, and applications of CLIP in CMH. Section 3 provides a detailed explanation of our proposed CAGMH method. In Section 4, we conducted extensive experiments from multiple perspectives to demonstrate the effectiveness of our proposed CAGMH method. Finally, Section 5 presents the findings of this study and outlines potential directions for future research. Section 6 provides a comprehensive summary of the paper.

2. Related work

As a key technology in artificial neural networks, hashing retrieval not only optimizes memory efficiency but also significantly improves retrieval speed in large-scale search tasks. Historically, hashing algorithms were primarily applied in image retrieval. However, with the increasing demand for cross-modal retrieval, new cross-modal hashing methods have been developed to address this need. These approaches map original information from different modalities into a shared space, subsequently distancing unrelated semantic features and drawing related ones closer together.

CMH methods can be classified into supervised (Wu et al., 2024) and unsupervised (Xie et al., 2024)approaches, Contingent on the use of supervised information throughout the training process. Supervised methods leverage labeled data to guide feature mapping across modalities, thereby optimizing the accuracy of similarity searches. Conversely, unsupervised methods rely on automatic learning for feature mapping without the need for labeled data. Additionally, recent research has explored the enhancement of hashing retrieval tasks using pre-trained visual-language models such as CLIP. These models effectively integrate visual and textual features, providing new insights and a robust experimental foundation for advancing CMH techniques. The Mamba model, introduced by Gu and Dao (2023) in their December 2023 publication, provides a novel framework that significantly influences our approach to CMH.

2.1. Unsupervised cross-modal hashing

Unsupervised CMH algorithms, which do not rely on labeling information, enhance flexibility by learning predefined similarity signals between dissimilar data, making them more compatible with real-world retrieval scenarios. The IMH (Song et al., 2013) method studies both inter- and intra-modal data consistency, projects this consistency into Hamming space, and employs linear regression for hash function learning. Concurrently, deep learning-based unsupervised CMH methods have garnered significant attention in the research community. For example, UGACH (Zhang et al., 2018) utilizes Generative Adversarial Networks (GANs), where a generator produces data from one modality to match another modality, and a discriminator distinguishes between real and generated data. This method not only significantly reduces the semantic disparity across different modalities but also extracts additional semantic information. DJSRH (Su et al., 2019) explores the

shared semantic links between modalities by analyzing the affinity matrix of joint semantics, thereby addressing the challenge of limited annotated data and improving retrieval efficiency, which is crucial for handling large-scale cross-modal dataset. The UKD method (Hu et al., 2020) applies knowledge distillation, employing the similarity matrix derived from the teacher network to guide the hash code learning process in the student network. Additionally, UCCH (Hu et al., 2022) incorporates contrastive learning in unsupervised CMH, emphasizing semantic relevance over irrelevance, which enables the model to better focus on relevant semantic information. Moreover, UCCH designs a hash memory to optimize binary hash codes.

2.2. Supervised cross-modal hashing

Unsupervised CMH, while not requiring label information, lacks strong semantic relevance. In contrast, supervised CMH utilize label information to improve the discriminative power of their hash codes, thereby outperforming unsupervised approaches. For instance, DCMH (Jiang and Li, 2017) was the first model to utilize deep neural networks for processing data from different modalities. This approach overcame the limitations of traditional cross-modal retrieval methods in terms of semantic alignment and retrieval efficiency. By integrating the strengths of deep learning, DCMH achieved end-to-end cross-modal hash learning, advancing the field of cross-modal retrieval and laying the groundwork for subsequent methods based on deep learning. Over time, research shifted from basic deep learning models to more sophisticated approaches, including adversarial learning, attention mechanisms, and self-supervised learning. While these advancements have shown promise, they remain constrained by inefficiencies and limited semantic alignment capabilities in large-scale retrieval tasks. This bottleneck was addressed by the introduction of CLIPMH (Zhu et al., 2023), which incorporates the large-scale pre-trained model CLIP into cross-modal hash retrieval. This innovation enables efficient alignment between images and text, marking a significant milestone in leveraging large-scale pre-trained models for cross-modal retrieval.

2.3. CLIP on cross-modal hashing

Before the introduction of the CLIP framework, various techniques and frameworks, such as Transformer (Vaswani, 2017) and VGG (Simonyan and Zisserman, 2014), were commonly used to extract image and text information. However, these methods struggled with aligning inter-modal features, which significantly impacted the model's learning ability and the quality of the generated hash codes. CLIP4Hash leverages the multimodal capabilities of CLIP to generate high-quality, semantically meaningful hash codes (Zhuo et al., 2022). CLIP, through large-scale unsupervised pre-training, successfully established a strong correspondence between image and text descriptions. CLIP not only learned to recognize image content but also understood and generated natural language descriptions associated with images, resulting in a tight alignment of visual and textual modalities. This alignment was not achieved through manually defined rules or mappings but through the model's self-learning from a large number of instances, thus regarded as a "natural" alignment. Due to this characteristic, CLIP excels in additional supervised training tasks and is naturally suited for cross-modal hashing retrieval tasks. Methods such as DCMHT (Tu et al., 2022), DNPH (Huo et al., 2024a), DSPH (Huo et al., 2023), and DNpH (Qin et al., 2024) have successfully utilized the CLIP framework, achieving significant results.

2.4. Related work on Mamba method

The Mamba method, introduced by Gu and Dao (2023), has garnered significant attention due to its exceptional capabilities in long-sequence selection and memory modeling. By efficiently capturing and dynamically representing spatio-temporal features, it has achieved remarkable success in the field of remote sensing. This has led to the publication of numerous high-quality studies. For instance, RS-Mamba (Zhao et al., 2024) exploits Mamba's unique SSM model to address the limitations of discrete modes effectively. It selectively retains relevant markers while disregarding irrelevant ones. Similarly, HyperMamba (Liu et al., 2024) leverages the Mamba model to implement a Spatial Neighborhood Adaptive Scanning module and a Spectral Adaptive Enhancement Scanning module. These innovations enable the efficient extraction and dynamic representation of spatial and spectral information from hyperspectral images, significantly enhancing hyperspectral image classification performance through its selective mechanism. These advancements are highly inspiring for our research. To the best of our knowledge, the Mamba method has not yet been applied to cross-modal hash retrieval, which we believe represents a meaningful and promising avenue for exploration.

Table 1
Table of notations and definitions.

Notation	Definition
D	Training dataset
N	Number of samples
B	Hash code
K	Hash code length
L	Extracted feature length
x_i	ith image sample
y_i	ith text sample
l_i	Label corresponding to ith sample
C	Number of categories
M	Number of features per sample
M'	Number of irrelevant pairs

3. Methodology

In this section, we present the CAGMH model framework, as illustrated in Fig. 2. Section 3.1 introduces the formal definition of CAGMH, while Section 3.2 describes the architecture of the proposed network. In Section 3.3, we elaborate on the hash learning module, which serves as a key component of our framework. Finally, Section 3.4 details the procedure for generating binary hash codes from the trained CAGMH model.

3.1. Notation

We primarily deal with cross-modal similarity retrieval between image and text modalities. For this task, we define M sample pairs (including images and text descriptions), denoted as $D = \{d_i\}_{i=1}^{M}$, where each $d_i = \{x_i, y_i, l_i\}$, with x_i and y_i representing the ith image sample and the ith text sample, respectively. $l_i = [l_{i1}, l_{i2}, \ldots, l_{ic}]$ denotes the multi-label annotation for the ith sample, where c represents the number of categories. The specific symbols and their meanings used in this section are defined in Table 1.

3.2. Network framework

3.2.1. Feature extraction module

Conventional approaches typically require additional steps to align or map image and text features, such as using manually defined mapping functions or specialized domain adaptation methods. However, these alignment steps may introduce additional complexity and computational overhead and may not adequately capture the underlying relationships between modalities. In contrast, this paper employs the CLIP framework as a feature extraction module. CLIP achieves efficient performance during inference and significantly reduces training time through semantic consistency and optimized large-scale pre-training. It leverages vast amounts of image–text pairs for unsupervised pre-training. This process enables CLIP to automatically learn associations between linguistic descriptions and visual content, generating

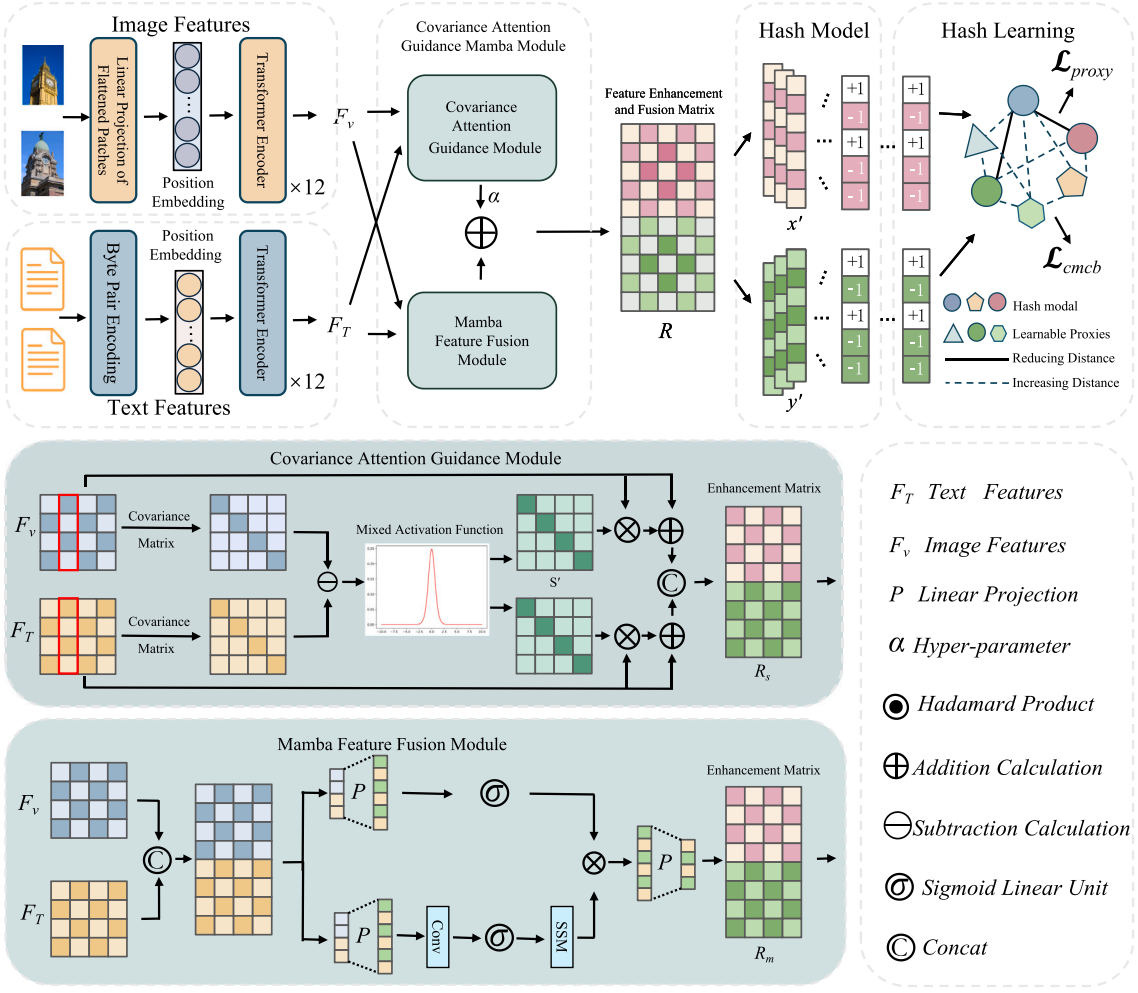


Fig. 2. Overview of CAGMH, which consists of four parts, where (1) Feature Extraction Module. (2) Covariance Attention Guidance Mamba Module, which consists of Covariance Attention Guidance Module and Mamba Feature Fusion Module respectively. (3) Hash Module. (4) Hash Learning Module: The design employs a multimodal proxy loss in conjunction with a cross-modal classification balance loss.

visual and linguistic representations with greater consistency and alignment, thereby effectively addressing the issue of inter-modal feature misalignment.

In this study, we used the pre-trained ViT-B/32 model provided by OpenAI to establish a performance baseline for supervised vector machine hashing CAGMH. The feature extraction method we employed can be represented by the following mathematical formula:

$$f_v = \text{CLIP}_{vision}(x_i) \tag{1}$$

$$f_t = \text{CLIP}_{textual}(y_i) \tag{2}$$

where f_v and f_t represent the extracted image features and text features, respectively, with dimensions both being $N \times M$.

3.2.2. Covariance attention guidance Mamba module

The core module is constructed collaboratively by the Covariance Attention Guidance Module and the Mamba Feature Fusion Module, which work together to achieve enhanced feature alignment and fusion for cross-modal retrieval.

Covariance Attention Guidance Module: This module is purely composed of matrix operations, tensor transformations, and the application of nonlinear functions. These operations do not introduce additional parameters during execution. Notably, this module can integrate data from different modalities, thereby extracting richer and deeper feature representations, which aids in the efficient processing of multimodal datasets. By optimizing the feature weight distribution and fusion strategy, the accuracy and stability of the model in handling complex tasks are improved, thus effectively enhancing relevant features and suppressing irrelevant information.

Compute the column-wise summation of image and text feature matrices to produce the aggregated feature vectors i_1 and t_1:

$$i_1 = \sum_{d=1}^{M} f_v[:, d] \tag{3}$$

$$t_1 = \sum_{d=1}^{M} f_t[:, d] \tag{4}$$

Perform a column-wise summation on the image and text feature matrices to generate the aggregated feature vectors. Then, compute the difference between the covariance matrices to derive the feature similarity matrix S:

$$S = i_1 i_1^T - t_1 t_1^T \tag{5}$$

Notably, smaller values in S indicate a stronger correlation between the corresponding modality features.

To stabilize the similarity matrix and avoid extreme gradients, a composite activation function is applied to S:

$$S' = (1 - \tanh(S^2)) \cdot \text{sigmoid}(S) \cdot (1 - \text{sigmoid}(S)) \tag{6}$$

where S' is the processed feature similarity matrix. This activation constrains the output range between 0 and 0.25, enhancing gradient stability.

This characteristic helps maintain the stability of gradient updates, significantly reducing the likelihood of gradient vanishing or exploding, thereby enhancing the overall stability of the model. Moreover, by effectively constraining extreme values in the input, this composite activation function also improves the robustness of the model. After passing through this activation function:

$$E_i = S' \odot f_v + f_v \tag{7}$$

$$E_t = S' \odot f_t + f_t \tag{8}$$

where, E_i represents the enhanced image feature, E_t represents the enhanced text feature, and $\odot$ denotes the Hadamard operation. By applying the Hadamard operation to multiply two matrices element-wise and then adding the result to the original feature matrix, this method effectively suppresses irrelevant information while amplifying relevant modality features. This process enhances feature representation and uncovers potential semantic information between modalities.

Concatenate the enhanced image and text features to form the final output matrix R_s:

$$R_s = [E_i, E_t] \tag{9}$$

This enhanced feature matrix R_s serves as the input to subsequent modules for further processing.

Mamba Feature Fusion Module: Building on the enhanced features from the previous module, the Mamba model is introduced as the core of the feature fusion module. Mamba, a state-space model (SSM), is specifically designed to efficiently capture long-range dependencies within sequences. Its dynamic selection mechanism enables selective filtering or propagation of information based on token positions, allowing it to effectively process long-sequence features. This capability makes Mamba particularly suitable for feature fusion, as it integrates and refines both intra-modal and inter-modal representations, enhancing the robustness and expressiveness of the learned features. Prior to utilizing this module, the extracted feature values must be concatenated and then input into the module. As shown in Eq. (10), R_m denotes the spliced feature matrix.

$$R_m = \text{MAMBA}\left([f_v, f_t]\right) \tag{10}$$

After the features have been processed by the covariance feature enhancement sub-module and the Mamba feature fusion sub-module, we perform a weighted summation of these features. This operation aims to parse the data through multiple perspectives, thus improving the generalization ability of the model on unseen datasets. In addition, this method effectively reduces the model's dependence on a single feature and lowers the risk of overfitting, thus enhancing the stability and reliability of the model. After a lot of experiments and comparisons, we finally set the weighting parameters to $\alpha = 0.09$. After processing in this module, the features will be restored to their original shapes again.

$$R = \alpha \cdot R_s + R_m \tag{11}$$

$$x' = R(:, 1 : L) \tag{12}$$

$$y' = R(:, L+1 : 2L) \tag{13}$$

where x' and y' are feature-enhanced and fused image features and text features, respectively.

In summary, the Covariance Attention Guidance Module ensures fine-grained alignment between modalities, while the Mamba Feature Fusion Module further integrates and refines these features. Their combination enables CAGMH to effectively address semantic gaps and achieve robust cross-modal retrieval performance.

3.3. Hash learning module

This module is part of the loss function section, where hash loss is a critical factor in reflecting the similarity relationships between images. This study combines multimodal proxy loss with cross-modal classification balance loss, which significantly enhances the stability and robustness of the model while optimizing the generated hash codes. However, most cross-modal methods typically only optimize the model or feature representation without introducing additional intermediate vectors or proxies, which limits the adjustment and optimization of sample representations. Although proxy-based methods perform well in cross-modal hashing retrieval, the initialization and updating of proxy vectors may affect model training and convergence, especially in cases of data class imbalance. Therefore, we propose a cross-modal classification balance loss, which helps the model learn multi-label classification tasks more effectively by balancing the category distribution between different modalities. This loss introduces a balancing factor that enhances the model's ability to differentiate between different classes and effectively combines image and text features. This approach further enhances the model's performance and robustness when handling imbalanced classes or noisy labels.

3.3.1. Multimodal proxy loss

To ensure that related data is embedded in close proximity while unrelated data-proxy pairs are separated, we refine the model parameters using multi-label proxies. For $P = \{p_1, p_2, \ldots, p_c\}$, where P represents the continuous proxy for each class, and p_i is a K-bit vector. In this loss, we not only consider the cosine distance between class binary hash codes and related proxies but also the distance between class binary hash codes and unrelated proxies. Thus, the cosine distance and unrelated proxy distance can be calculated using the following equations:

$$\text{pos}(h, p) = 1 - \frac{h \cdot p^T}{\|h\|_2 \cdot \|p\|_2} \tag{14}$$

$$\text{neg}(h, p) = \max\left(0, \frac{h \cdot p^T}{\|h\|_2 \cdot \|p\|_2} - \text{threshold}\right) \tag{15}$$

where threshold = threshold(C, K) is a parameter used to determine when the corresponding similarity is considered to be negatively impacted, which can be set according to Ref. Xu et al. (2022). Here, h serves as a placeholder for the feature vector representation, allowing for a more flexible explanation of the formula and its components.

$$\begin{aligned} L^{x'}_{pos\_neg} = & \frac{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=1)\cdot \text{pos}(h_i^{x'}, p_j)}{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=1)} \\ & + \frac{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=0)\cdot \text{neg}(h_i^{x'}, p_j)}{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=0)} \\ & + \frac{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(|l_i \cdot l_j|=0)\cdot \text{neg}(h_i^{x'}, h_j^{x'})}{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(|l_i \cdot l_j|=0)} \end{aligned} \tag{16}$$

$$\begin{aligned} L^{y'}_{pos\_neg} = & \frac{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=1)\cdot \text{pos}(h_i^{y'}, p_j)}{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=1)} \\ & + \frac{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=0)\cdot \text{neg}(h_i^{y'}, p_j)}{\sum_{i=1}^{M}\sum_{j=1}^{C} I(l_{ij}=0)} \\ & + \frac{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(|l_i \cdot l_j|=0)\cdot \text{neg}(h_i^{y'}, h_j^{y'})}{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(|l_i \cdot l_j|=0)} \end{aligned} \tag{17}$$

where, I is the indicator function, and $I(l_{ij} = 1)$, $I(l_{ij} = 0)$, and $I(|l_i \cdot l_j| = 0)$ represent 'label-relevant,' 'label-irrelevant,' and 'label-disjoint' pairs, respectively. By constraining different sample pairs, the method pulls positive pairs closer and pushes irrelevant pairs apart, effectively capturing fine-grained semantics to improve retrieval

performance. Specifically, $L^{x'}_{pos\_neg}$ and $L^{y'}_{pos\_neg}$ represent the loss in image mode and text mode, respectively.

If only intra-modal loss is considered without addressing inter-modal loss, it may lead to insufficient semantic consistency, thereby reducing the performance of cross-modal retrieval. Therefore, we calculate inter-modal loss according to Eq. (18). Here, we specifically compute the inter-modal irrelevance loss to mitigate the risk of semantic inconsistency.

$$L_{neg\_pair\_xt} = \frac{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(l_i, l_j = 0) \cdot \mathrm{neg}(h_i^{x'}, h_j^{y'})}{\sum_{i=1}^{M'}\sum_{j=1}^{M'} I(l_i, l_j = 0)} \tag{18}$$

As shown in Eq. (19), the total multimodal proxy loss is calculated as:

$$L_{proxy} = L^{f'_x}_{pos\_neg} + L^{f'_y}_{pos\_neg} + L_{neg\_pair\_xt} \tag{19}$$

3.3.2. Cross modal class balance loss

To enhance hash representation via structured supervision and label constraints, we introduce the Cross Modal Classification Balancing Loss. This loss function is designed to significantly improve retrieval effectiveness and accuracy by refining the hash function and efficiently utilizing the hash target. The cross-modal classification balancing loss is composed of three auxiliary sub-losses: center loss, uniformity loss, and balance loss.

However, the model may lack an effective centrality constraint, which hinders its ability to cluster similar samples around a centroid and leads to insufficient feature concentration within the same category. We propose a centrality loss to address this issue, thereby enhancing the centroid representation ability for each category, as shown in Eq. (20):

$$L_{center} = \frac{1}{M}\sum_{i=1}^{M} \max(0, 1 - h_i \cdot t_i). \tag{20}$$

where h_i denotes the predicted hash code, constructed by concatenating the Hadamard matrix H with its negation $-H$, and t_i represents the central vector corresponding to the sample.

To ensure the generated hash codes are close to binary values $\{-1, 1\}$, the uniformity loss penalizes deviations from these discrete values. This is defined as:

$$L_{uniform} = \sum_{i=1}^{M} \left(|h_i| - 1\right)^3 \tag{21}$$

The cubic penalty pushes the predicted hash codes h toward the binary constraints while maintaining smooth gradients, effectively reducing the impact of errors and irrelevant information in feature representations.

In the current study, hash code generation tends to concentrate on a few categories, leading to a sparse distribution of hash codes in other categories and a reduction in category differentiation, ultimately affecting retrieval performance. To overcome this issue, we introduced a balanced loss function, as shown in Eq. (22), to ensure that hash codes are dispersed as widely as possible throughout the space, thereby improving retrieval performance.

$$L_{balance} = \beta \cdot \frac{1}{C}\sum_{c=1}^{C} \cdot \left| \frac{1}{N_c}\sum_{i=1}^{N_c} h_{ic} - 0.5 \right| \tag{22}$$

where, N_c denotes the number of samples in the cth category, while h_{ic} represents the feature vector of the ith sample within the cth category, where each element of h_{ic} lies within the range $[-1, 1]$. Through extensive experiments, we set β to 0.004. The purpose of this loss term is to ensure that the hash codes of each category are uniformly distributed around 0.5, thereby promoting an even distribution of hash codes across the entire space. Therefore, as shown in Eqs. (23) and (24), the hash codes of text and image are calculated separately.

$$\begin{aligned} L^{x'}_{cmcb} = {} & \frac{1}{M}\sum_{i=1}^{M}\left(\max(0, 1 - x'_i t_i) + \sum_{i=1}^{M}\left(|x'_i| - 1|\right)^3\right) \\ & + \beta \cdot \frac{1}{C}\sum_{c=1}^{C} \cdot \left| \frac{1}{N}\sum_{i=1}^{N} x'_{ic} - 0.5 \right| \end{aligned} \tag{23}$$

$$\begin{aligned} L^{y'}_{cmcb} = {} & \frac{1}{M}\sum_{i=1}^{M}\left(\max(0, 1 - y'_i t_i) + \sum_{i=1}^{M}\left(|y'_i| - 1|\right)^3\right) \\ & + \beta \cdot \frac{1}{C}\sum_{c=1}^{C} \cdot \left| \frac{1}{N}\sum_{i=1}^{N} y'_{ic} - 0.5 \right| \end{aligned} \tag{24}$$

The Cross Modal Class Balance Loss effectively combines center, uniform, and balance loss to generate compact, discrete, and balanced hash codes. This approach enhances both the robustness and accuracy of the cross-modal retrieval model.

The overall loss calculation for the cross-modal classification balance is given by Eq. (25):

$$L_{cmcb} = L^{x'}_{cmcb} + L^{y'}_{cmcb} \tag{25}$$

The total loss calculation for the hashing learning module is shown in Eq. (26):

$$L_{total} = \sigma(\omega) \cdot L_{cmcb} + L_{proxy} \tag{26}$$

where, $\sigma(\omega)$ represents the sigmoid function, where ω is a learnable parameter initialized to 1.6. By automatically adjusting the weight of the loss term, the model can more quickly converge to the optimal solution, thereby speeding up the training process.

3.4. Out-of-sample extension

Algorithm 1 outlines the steps for generating hash codes. First, the trained CAGMH model generates class-specific binary hash codes for the query samples, using the sign function to convert the results into hash codes.

Algorithm 1 Hash Code Generation Steps for CAGMH

Input: : Query samples q_i; Parameters for CAGMH.
Output: : Binary hash code for q_i.
1: Generate binary-like hash codes by inputting the query data q_i into the trained CAGMH model.
2: Convert these to final hash codes using a sign function:

$$\mathrm{sign}(u) = \begin{cases} +1, & u > 0 \\ -1, & u < 0 \end{cases} \tag{27}$$

4. Experiments

We validated the effectiveness of our CAGMH framework through experiments on four public CMH datasets: MIRFLICKR-25K (Huiskes and Lew, 2008), NUS-WIDE (Chua et al., 2009), MS COCO (Lin et al., 2014), and IAPR TC-12 (Grubinger et al., 2006) . We provide a detailed description of the experimental datasets, the implementation of CAGMH, and the evaluation metrics used.

4.1. Datasets

MIRFLICKR-25K: This dataset from Flickr contains 24,581 image–text pairs across 24 categories, each pair annotated with multi-labels. The dataset covers a broad range of subjects, including natural scenes, objects, people, and events, making it diverse and challenging for tasks involving image classification, annotation, and retrieval. The relatively small scale of MIRFLICKR-25K also allows for rapid experimentation,

Table 2
Detailed settings of each data set.

Dataset	Query	Train	Database	Total
MIRFLICKR-25K	2000	10,000	22,581	24,581
NUS WIDE	2100	10,500	190,679	192,779
MS COCO	5000	10,000	117,218	122,218
IAPR TC-12	5000	10,000	14,626	19,626

which is particularly useful in testing new methods during early-stage development.

NUS-WIDE: This large dataset includes 269,648 image–text pairs labeled across 81 categories, covering a diverse range of scenes and objects. After removing less populated categories, we selected 21 major categories, resulting in a subset with 195,834 pairs, each belonging to at least one category. NUS-WIDE stands out due to its scale and real-world-inspired diversity, reflecting the complexity of user-generated content, where structured and unstructured information often coexist. This makes it particularly suitable for evaluating cross-modal retrieval methods in scenarios such as large-scale multimedia search or recommendation systems.

MS COCO: This dataset, a prominent benchmark for object detection, includes 80 multi-label categories. It contains a total of 123,389 images, consisting of 82,785 training images and 40,504 validation images, each accompanied by five captions. In our experiments, to ensure that each sample contains both image and text modalities, we merged the training and test sets, making sure that each data pair belongs to at least one category. MS COCO reflects real-world applications in tasks like image captioning and multi-modal understanding due to its high-quality annotations and diverse content, including natural environments and human-centric scenes. Its focus on both multi-label categorization and text descriptions aligns well with the requirements of modern cross-modal retrieval tasks.

IAPR TC-12: The IAPR TC-12 dataset consists of 20,000 images across 255 categories, including people, animals, and landscapes. Each image is annotated with captions in multiple languages (English, German, Spanish). For our experiments, we utilized the English captions. This dataset serves as a challenging and valuable resource for cross-modal hashing retrieval.

We adopt the data division strategy proposed by DCMHT (Tu et al., 2022) and DHaPH (Huo et al., 2024b) apply it to our datasets, with necessary modifications. Specifically, we exclude text pairs that do not belong to any class, ensuring the remaining data is properly categorized and aligned with our research objectives. The query set was randomly sampled from the original dataset, with the remaining data used as the retrieval set. Training instances were randomly selected from the retrieval database.

4.2. Experimental settings

4.2.1. Baseline methods

To effectively demonstrate the performance of our designed algorithm, we compared the CAGMH framework with 11 classical and representative cross-modal hashing retrieval methods, including DCMH (Jiang and Li, 2017), GCDH (Bai et al., 2022), SCAHN (Wang et al., 2020), DCMHT (Tu et al., 2022), DHaPH (Huo et al., 2024b), DNPH (Huo et al., 2024a), DNpH (Qin et al., 2024), DSPH (Huo et al., 2023), and TwDH (Tu et al., 2024). All comparative methods were implemented using their official source code. The network parameters were configured according to the official settings provided. If the parameters were not specified in the corresponding papers, we adhered to those provided in the official source code.

4.2.2. Implementation details

We utilized a Linux server equipped with an NVIDIA A40 GPU for model training and comparative experiments, and implemented our model using the open-source framework PyTorch 2.2.1. The ViT-B/32 pre-trained model provided by OpenAI was employed as our feature backbone network. We optimized the model parameters using the Adam optimizer with a learning rate of 0.001, training the model for 50 epochs with a batch size of 128. The hyperparameters α, β, and ω were set to 0.09, 0.004, and 1.6, respectively. During the data processing phase, images were uniformly resized to 224×224, and text was encoded using the BPE method. The query and train settings for different datasets are detailed in Table 2.

4.2.3. Evaluation indicators

To comprehensively validate the performance of CAGMH, we used five key metrics: Mean Average Precision (mAP), mean Average Precision within Hamming radius 2 (P@H $\leqslant$ 2), Precision–Recall (PR) curve, TopN-precision curve and Normalized Discounted Cumulative Gain (NDCG@1000).

Table 3 presents the mAP results for text-to-image (T2I) and image-to-text (I2T) retrieval tasks, with the best and second-best results highlighted in **bold** and underlined, respectively. Compared to baseline methods, our CAGMH model consistently achieves superior performance. On the MIRFLICKR-25K dataset, CAGMH outperforms the best CMH method, DNpH, by 5.31% in I2T and 7.49% in T2I retrieval. Similarly, on the NUSWIDE dataset, it surpasses GCDH by 3.45% and 2.96% in I2T and T2I retrieval, respectively. On the COCO dataset, our approach also shows a leading advantage. Finally, we achieved excellent results on the IAPR TC-12 dataset. Additionally, our model achieves optimal performance after just 50 training epochs, demonstrating fast convergence. This is primarily attributed to the proposed loss function, which integrates Covariance Attention Guidance and Cross-modal Class Balance Loss. The Covariance Attention Guidance module aligns features across modalities by considering their covariance, accelerating convergence through improved hash code optimization. Its convergence analysis and mAP changes are shown in Fig. 3. Additionally, the Cross-modal Class Balance Loss mitigates the prevalent issue of class imbalance in real-world datasets, promoting more stable and balanced learning. By addressing this imbalance, the model generalizes better to unseen data and is less susceptible to overfitting, resulting in improved performance across various tasks.

Moreover, we evaluated our model from multiple perspectives. Figs. 4, 5, 6, and 7 compare the performance of different datasets under 16-bit and 32-bit hash codes using PR curves and Top-N precision curves. The results show that our method significantly outperforms others, mainly due to the Covariance Attention Guidance Mamba Module, which effectively fuses extracted features to enhance semantic alignment. In contrast, baseline methods rely solely on CLIP's direct feature extraction, lacking further refinement, which limits their performance. Additionally, the superior results in Top-N precision curves are attributed to the Cross Modal Class Balance Loss, which mitigates class imbalance, thereby improving retrieval accuracy and robustness.

In Fig. 8, subfigures (a), (b), (c), and (d) present the mAP@H $\leq$ 2 results for I2T retrieval, while subfigures (e), (f), (g), and (h) display the corresponding results for T2I retrieval. When the hash code length is 64 bits, the mAP@H $\leq$ 2 values for most hashing methods tend to decrease, likely due to the increased sparsity of the discrete space, which causes fewer data points to fall within a Hamming radius of 2. However, our method maintains a high mAP@H $\leq$ 2, demonstrating the effectiveness of our proposed loss function. This further demonstrates that the hash codes generated by CAGMH maintain high retrieval accuracy and performance, particularly in similarity matching and result ranking.

Fig. 9 shows a comparison of our method and baselines using the NDCG@1000 metric, which evaluates both relevance and ranking quality of retrieval results. Our method outperforms all baselines, especially on the MIRFLICKR-25K dataset, due to the Covariance Attention Guidance Mamba Module and the designed loss function, which enhance feature representation and ranking precision, ensuring more efficient and accurate cross-modal retrieval.

Table 3
Comparison with baselines in terms of map results w.r.t. 16 bit, 32 bit and 64 bit on four datasets.

Task	Method	MIRFLICKR-25K			NUS WIDE			MS COCO			IAPR TC-12		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
I2T	DCMH	0.7288	0.7411	0.7490	0.5238	0.5995	0.6195	0.5177	0.5311	0.5471	0.4584	0.4882	0.5045
	SCAHN	0.7647	0.7713	0.7743	0.6376	0.6577	0.6618	0.6417	0.6668	0.6516	0.5213	0.5351	0.5528
	GCDH	0.7991	0.8123	0.8204	0.7142	0.7367	0.7498	0.7297	0.7651	0.7905	–	–	–
	DCMHT	0.8254	0.8284	0.8257	0.6832	0.6892	0.7025	0.6440	0.6465	0.6552	0.5749	0.5960	0.6132
	DHaPH	0.8271	0.8351	0.8324	0.7215	0.7333	0.7410	**0.7310**	0.7402	0.7496	0.5999	0.6207	0.6307
	DNPH	0.7899	0.8225	0.8201	0.6668	0.6872	0.7004	0.6356	0.6865	0.7395	0.4573	0.5240	0.5671
	DNpH	0.8399	0.8487	0.8500	0.7135	0.7169	0.7247	0.6754	0.6897	0.6862	**0.7012**	**0.7272**	0.7473
	DSPH	0.8036	0.8286	0.8439	0.6756	0.6898	0.7161	0.6916	0.7416	0.7706	0.5281	0.6082	0.6629
	TwDH	0.7613	0.7544	0.7834	0.6226	0.6636	0.6668	0.6091	0.6808	0.7114	–	–	–
	Ours	**0.8850**	**0.9003**	**0.9128**	**0.7507**	**0.7678**	**0.7857**	0.7131	**0.7732**	**0.7981**	0.5791	0.6780	**0.7583**
T2I	DCMH	0.7520	0.7696	0.7776	0.5440	0.5901	0.5956	0.5510	0.5883	0.6050	0.5180	0.5368	0.5464
	SCAHN	0.7672	0.7823	0.7845	0.6676	0.6739	0.6754	0.6417	0.6645	0.6533	0.5188	0.5252	0.5398
	GCDH	0.7849	0.8022	0.8067	0.7215	0.7423	0.7534	**0.7261**	0.7650	0.7885	–	–	–
	DCMHT	0.8115	0.8179	0.8200	0.6920	0.7081	0.7208	0.6282	0.6361	0.6484	0.5770	0.6185	0.6271
	DHaPH	0.8089	0.8170	0.8194	0.7203	0.7284	0.7388	0.6999	0.7037	0.7168	0.6011	0.6134	0.6297
	DNPH	0.7880	0.8123	0.8062	0.6860	0.7066	0.7204	0.6346	0.6932	0.7438	0.4479	0.5059	0.5580
	DNpH	0.8151	0.8249	0.8329	0.7222	0.7265	0.7313	0.6595	0.6801	0.6990	**0.7010**	**0.7260**	0.7472
	DSPH	0.7909	0.8044	0.8298	0.6910	0.7054	0.7304	0.6941	0.7445	0.7674	0.5119	0.6019	0.6611
	TwDH	0.7508	0.7476	0.7788	0.6298	0.6731	0.6772	0.6004	0.6701	0.7056	–	–	–
	Ours	**0.8853**	**0.9000**	**0.9123**	**0.7497**	**0.7690**	**0.7874**	0.7140	**0.7736**	**0.7979**	0.5788	0.6770	**0.7579**

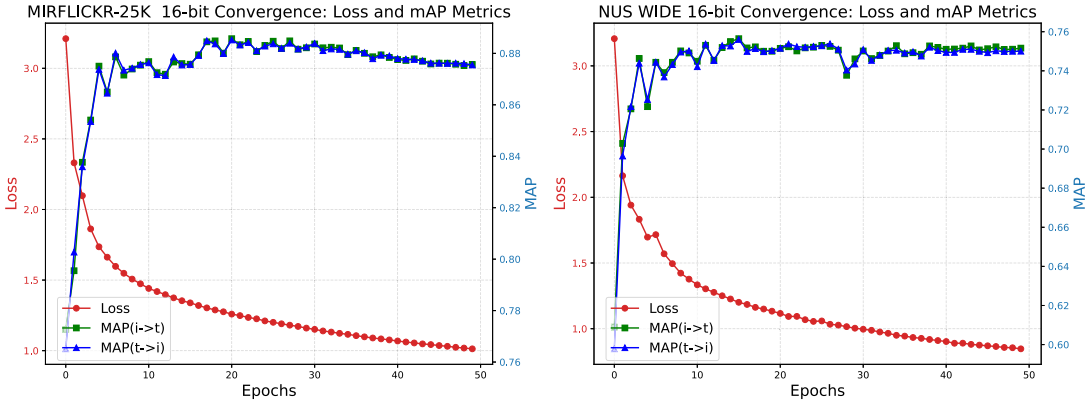


Fig. 3. MIRFLICKR-25K and NUS-WIDE 16-bit convergence curves: Loss and mAP metrics.

4.3. Ablation study

We conducted ablation experiments on MIRFLICKR-25K and NUS-WIDE to demonstrate the effectiveness of each proposed method, introducing three model variants: (1) CAGMH-0, which omits both the Covariance Attention Guidance Mamba Module and the cross-modal classification balance loss; (2) CAGMH-1, which omits only the Covariance Attention Guidance Mamba Module; (3) CAGMH-2, which omits only the cross-modal classification balance loss. The experimental results, presented in Table 4, indicate that the cross-modal classification balance loss effectively addresses the proxy loss issue, while the Covariance Attention Guidance Mamba Module significantly enhances model performance by facilitating information exchange and reinforcement across modalities. The combined application of these components results in average performance improvements of 8.91% and 7.51% on the two datasets, respectively.

Table 4
Ablation experiments.

Task	Method	MIRFLICKR-25K			NUS-WIDE		
		16 bits	32 bits	64 bits	16 bits	32 bits	64 bits
I2T	CAGMH-0	0.8036	0.8286	0.8439	0.6756	0.6898	0.7162
	CAGMH-1	0.8324	0.8429	0.8530	0.6818	0.7009	0.7175
	CAGMH-2	0.8648	0.8881	0.9083	0.7440	0.7601	0.7803
	CAGMH	**0.8850**	**0.9003**	**0.9128**	**0.7507**	**0.7678**	**0.7857**
T2I	CAGMH-0	0.7919	0.8044	0.8298	0.6910	0.7054	0.7304
	CAGMH-1	0.8121	0.8169	0.8332	0.6995	0.7139	0.7352
	CAGMH-2	0.8655	0.8876	0.9081	0.7424	0.7688	0.7860
	CAGMH	**0.8853**	**0.9000**	**0.9123**	**0.7497**	**0.7690**	**0.7874**

4.4. Hyperparameter experiments

Fig. 10 presents the experimental results that compare the impact of the hyperparameters α, β, and ω on the performance of the MIRFLICKR-25K dataset. These results clearly demonstrate the critical role that each hyperparameter plays in the overall model performance. From the experimental results, we can conclude that when α, β, and ω are set to 0.09, 0.004, and 1.6, respectively, the model achieves better performance. Furthermore, this specific configuration consistently produces optimal outcomes when the parameters are used in conjunction, thereby enhancing the model's overall robustness and reliability.

5. Discussion

We propose an end-to-end deep CMH hashing method. Our approach aligns with existing deep CMH methods during feature extrac-

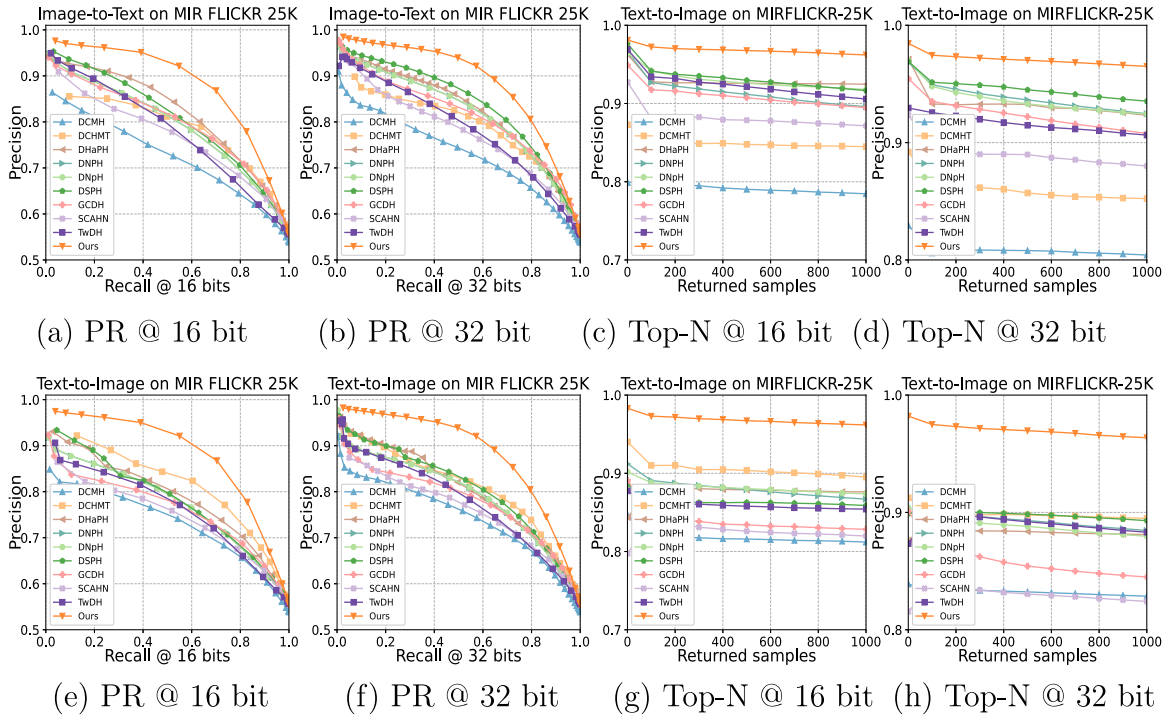


Fig. 4. MIRFLICKR-25K results of Precision–Recall curves, TopN precision curves on 16 bit and 32 bit.

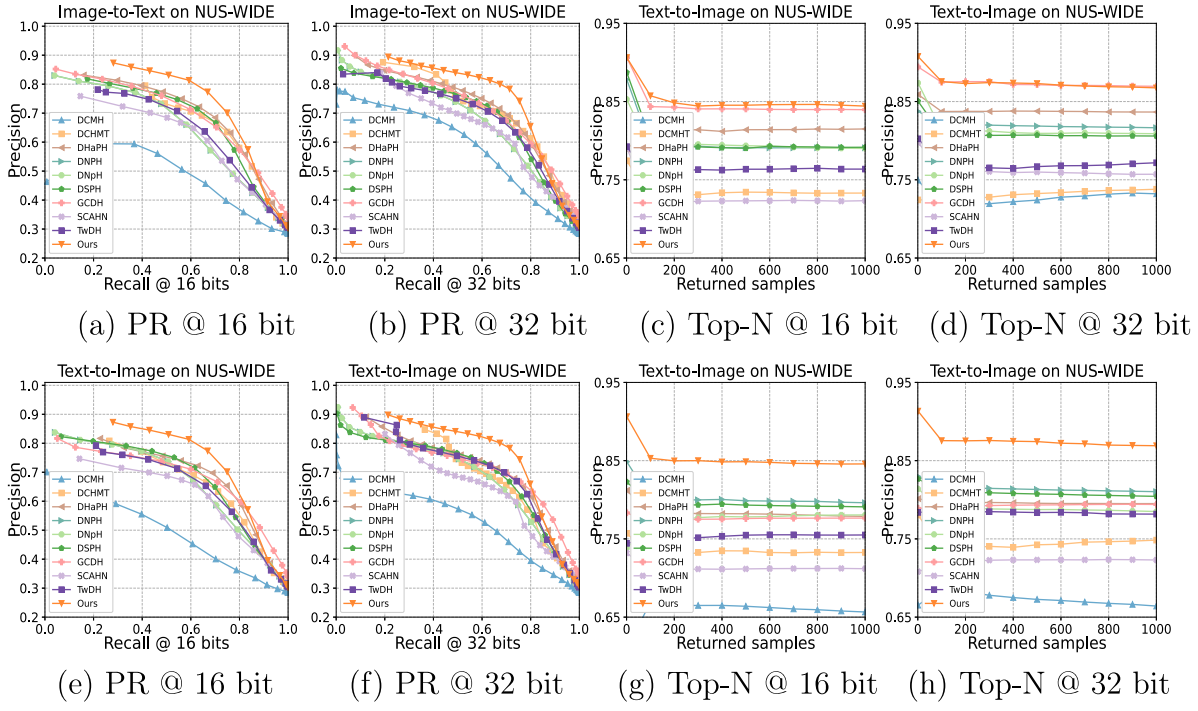


Fig. 5. NUS-WIDE results of Precision–Recall curves, TopN precision curves on 16 bit and 32 bit.

tion, focusing on establishing associations between cross-modal data and semantic labels.

Traditional methods often overlook information exchange between modalities, leading to insufficient fusion of cross-modal information. To address this, our model introduces the Mamba module, which effectively integrates information across different modalities. Additionally, we employ a zero-parameter covariance feature enhancement method to further amplify significant inter-modal information. This enhancement enables the model to retain single-modality information while simultaneously learning latent connections between different modalities, thereby improving the representation of similar information and enhancing overall feature quality.

To tackle the limitations of traditional proxy loss functions, which lack effective center constraints, we propose a novel loss function. Based on extensive experiments, it can be concluded that our method significantly enhances retrieval accuracy under short hash codes, as measured by the mAP@H $\leqslant$ 2 metric. While our SMPH model shows notable improvements in accuracy compared to DSPH, the gains on

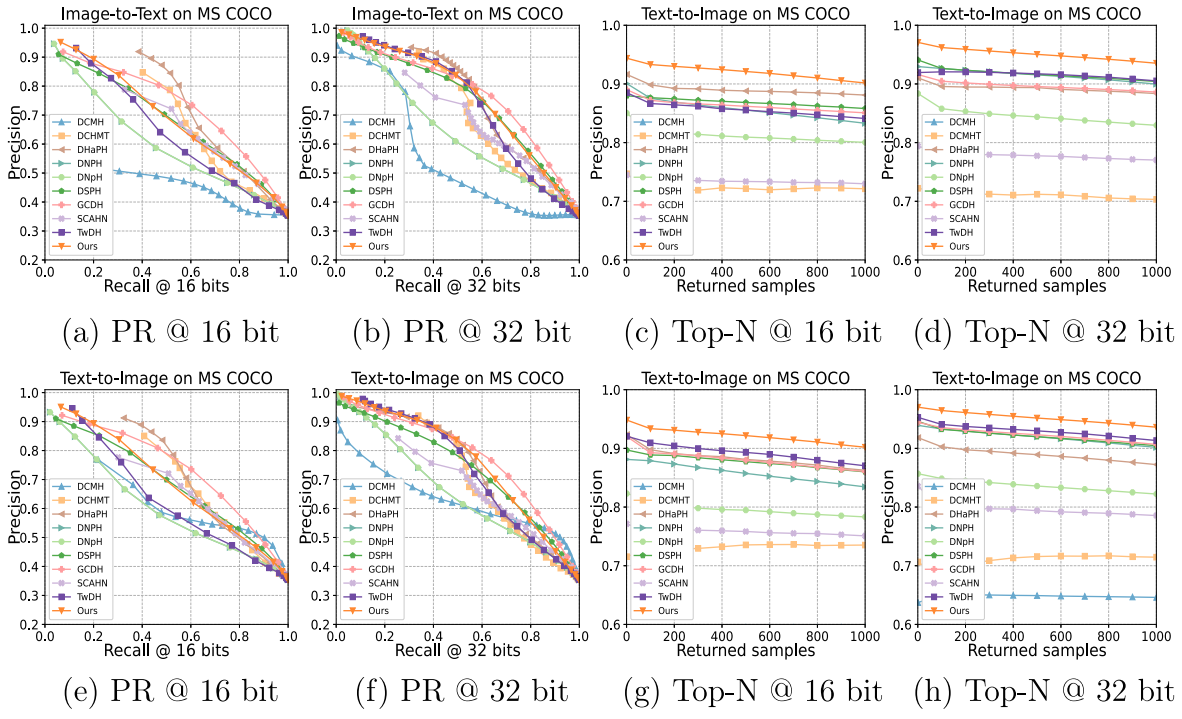


Fig. 6. MS COCO results of Precision–Recall curves, TopN precision curves on 16 bit and 32 bit.

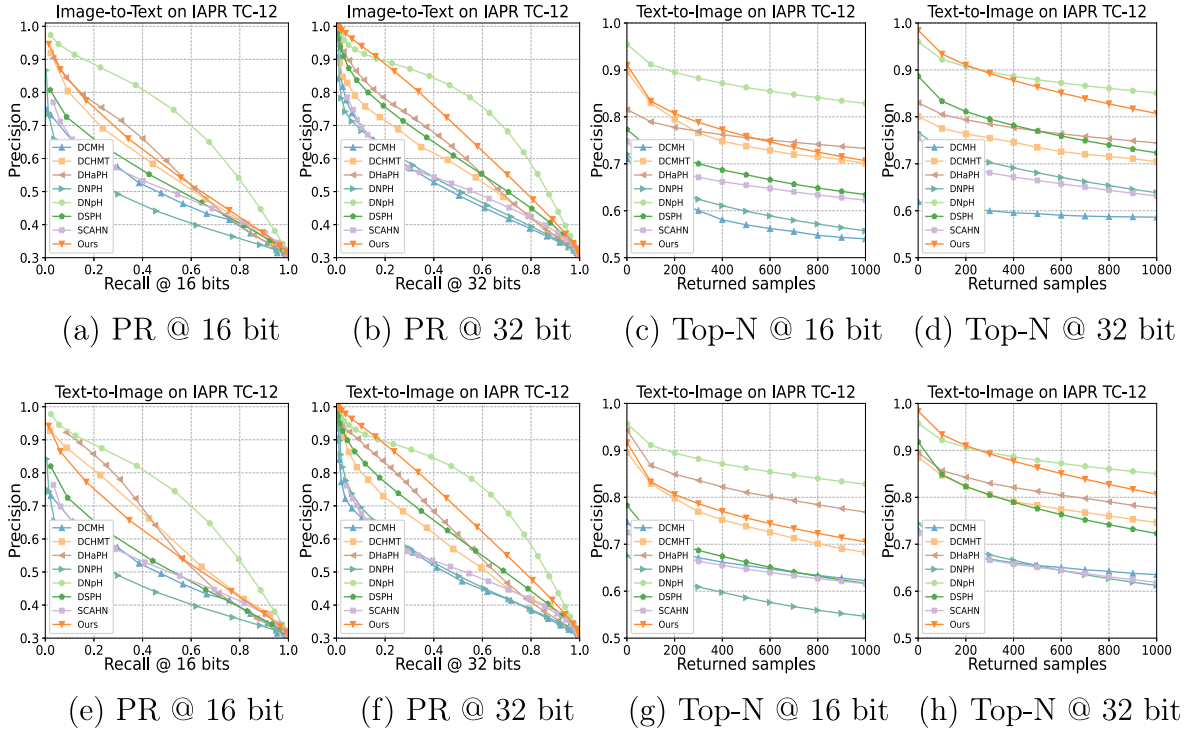


Fig. 7. IAPR TC-12 results of Precision–Recall curves, TopN precision curves on 16 bit and 32 bit.

the MS COCO and IAPR TC-12 datasets remain limited. The MS COCO and IAPR TC-12 datasets differs significantly from the NUS-WIDE and MIRFLICKR-25K datasets in terms of textual description style. While the MS COCO and IAPR TC-12 dataset uses sentence-based descriptions with strong contextual associations, the NUS-WIDE and MIRFLICKR-25K datasets rely on textual descriptions composed of related keywords. This distinction is a critical factor contributing to the challenge of improving model performance on the MS COCO and IAPR TC-12 datasets. To address this, future research could explore strategies such as prompt learning to enhance the model's ability to focus on keywords while mitigating the influence of contextual interference.

6. Conclusion

This paper presents an efficient cross-modal hashing (CMH) retrieval method, named Covariance Attention Guidance Mamba Hashing

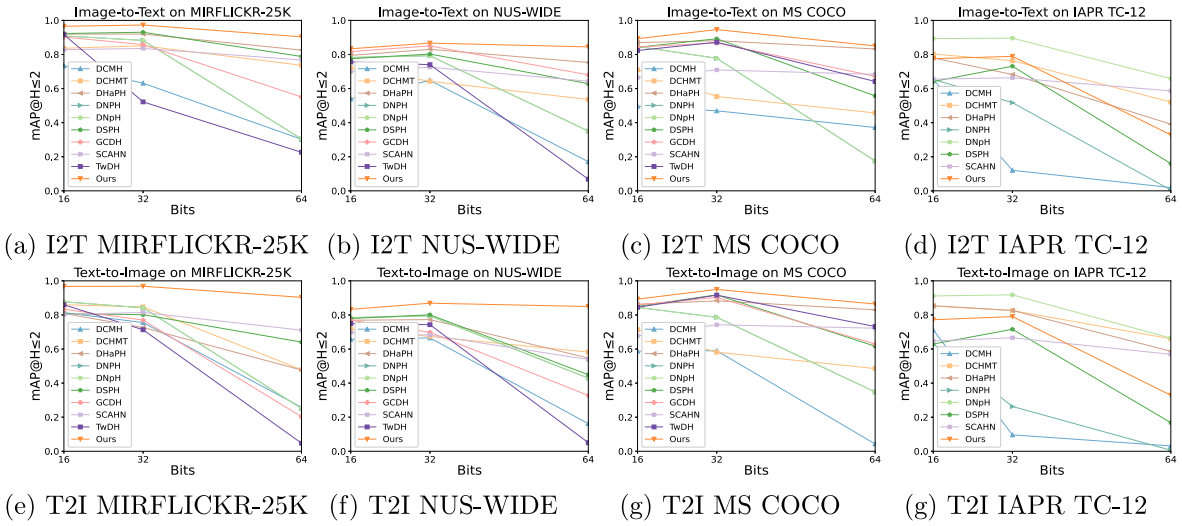


Fig. 8. The mAP@H ⩽ 2 w.r.t. different code lengths on the four datasets.

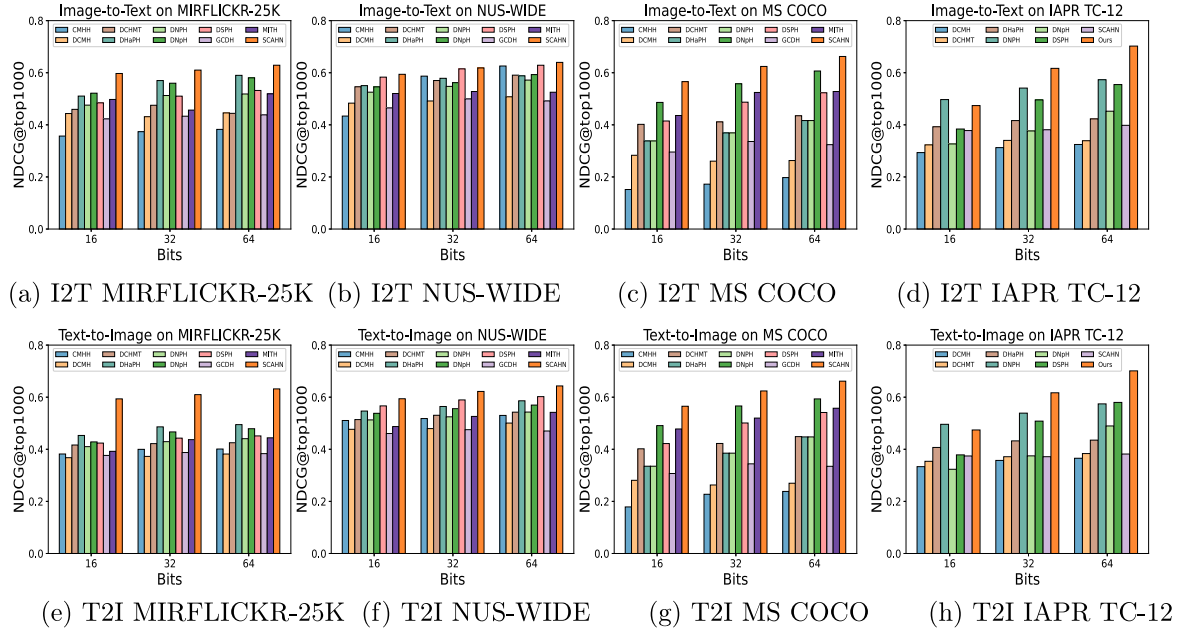


Fig. 9. Comparison with baselines in terms of NDCG@1000 w.r.t. fferent code lengths on the four datasets.

(CAGMH) for Cross-Modal Retrieval, which leverages covariance attention and the Mamba feature fusion module. Compared to existing CMH approaches, CAGMH offers distinct advantages. First, it extracts intra-modal features through separate networks, while a fusion and enhancement module uncovers latent inter-modal relationships. Moreover, by introducing a novel cross-modal classification balance loss and multimodal proxy loss, CAGMH effectively overcomes limitations in traditional proxy-based methods, particularly in addressing cross-modal balance, diversity loss, and robustness. Extensive experiments on benchmark datasets show that CAGMH consistently surpasses state-of-the-art CMH methods, demonstrating superior retrieval accuracy and efficiency.

CRediT authorship contribution statement

Gang Wang: Writing – original draft, Software, Methodology. **Shuli Cheng:** Supervision, Funding acquisition, Formal analysis. **Anyu Du:** Supervision, Project administration. **Qiang Zou:** Validation, Data curation.

Funding statement

This research was funded by the National Natural Science Foundation of China under Grant 62441213, the Key Laboratory Open Projects in Xinjiang Uygur Autonomous Region, China under Grant 2023D04028, and the Graduate Research and Innovation Project of Xinjiang Uygur Autonomous Region, China under Grant XJ2024G087.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

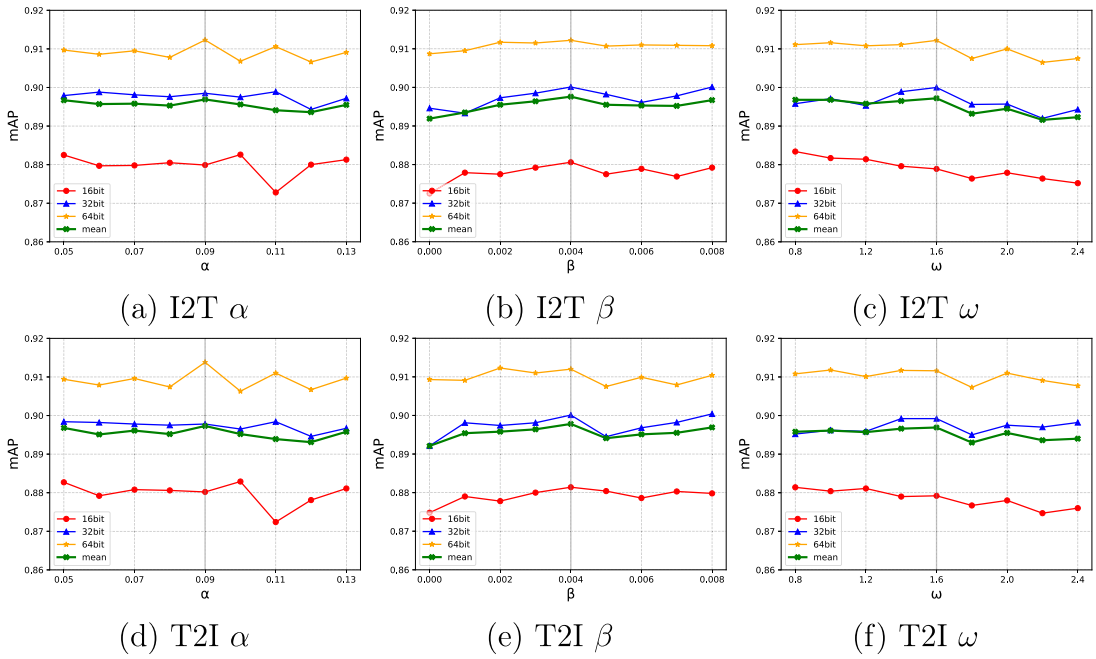


Fig. 10. Hyperparameter experiment on MIRFLICKR-25K.

References

Bai, C., Zeng, C., Ma, Q., Zhang, J., 2022. Graph convolutional network discrete hashing for cross-modal retrieval. IEEE Trans. Neural Netw. Learn. Syst. 35 (4), 4756–4767.

Chao, Z., Cheng, S., Li, Y., 2023. Deep internally connected transformer hashing for image retrieval. Knowl.-Based Syst. 279, 110953, URL: https://api.semanticscholar.org/CorpusID:261544636.

Cheng, S., Chan, R., Du, A., 2024. CACFTNet: A hybrid cov-attention and cross-layer fusion transformer network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 62, 1–17, URL: https://api.semanticscholar.org/CorpusID:268313599.

Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., 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. pp. 1–9.

Grubinger, M., Clough, P.D., Müller, H., Deselaers, T., 2006. The IAPR TC-12 benchmark: A new evaluation resource for visual information systems. URL: https://api.semanticscholar.org/CorpusID:18883184.

Gu, A., Dao, T., 2023. Mamba: linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752.

Hu, H., Xie, L., Hong, R., Tian, Q., 2020. Creating something from nothing: Unsupervised knowledge distillation for cross-modal hashing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3123–3132.

Hu, P., Zhu, H., Lin, J., Peng, D., Zhao, Y.-P., Peng, X., 2022. Unsupervised contrastive cross-modal hashing. IEEE Trans. Pattern Anal. Mach. Intell. 45 (3), 3877–3889.

Huiskes, M.J., Lew, M.S., 2008. The mir flickr retrieval evaluation. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. pp. 39–43.

Huo, Y., Qin, Q., Dai, J., Wang, L., Zhang, W., Huang, L., Wang, C., 2023. Deep semantic-aware proxy hashing for multi-label cross-modal retrieval. IEEE Trans. Circuits Syst. Video Technol. 34 (1), 576–589.

Huo, Y., Qin, Q., Dai, J., Zhang, W., Huang, L., Wang, C., 2024a. Deep neighborhood-aware proxy hashing with uniform distribution constraint for cross-modal retrieval. ACM Trans. Multimed. Comput. Commun. Appl. 20, 1–23, URL: https://api.semanticscholar.org/CorpusID:267352535.

Huo, Y., Qin, Q., Zhang, W., Huang, L., Nie, J., 2024b. Deep hierarchy-aware proxy hashing with self-paced learning for cross-modal retrieval. IEEE Trans. Knowl. Data Eng. 36 (11), 5926–5939. http://dx.doi.org/10.1109/TKDE.2024.3401050.

Irie, G., Arai, H., Taniguchi, Y., 2015. Alternating co-quantization for cross-modal hashing. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1886–1894.

Jiang, Q.-Y., Li, W.-J., 2017. Deep cross-modal hashing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 3270–3278.

Khan, M., Ahmad, J., El-Saddik, A., Gueaieb, W., Masi, G.D., Karray, F., 2024a. Drone-HAT: Hybrid attention transformer for complex action recognition in drone surveillance videos. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. CVPRW, pp. 4713–4722, URL: https://api.semanticscholar.org/CorpusID:272915606.

Khan, M., Gueaieb, W., Elsaddik, A., De Masi, G., Karray, F., 2025. Graph-based knowledge driven approach for violence detection. IEEE Consum. Electron. Mag. 14 (1), 77–85. http://dx.doi.org/10.1109/MCE.2024.3446192.

Khan, M., Saad, M., Khan, A., Gueaieb, W., Saddik, A.E., Masi, G.D., Karray, F., 2024b. Action knowledge graph for violence detection using audiovisual features. In: 2024 IEEE International Conference on Consumer Electronics. ICCE, pp. 1–5, URL: https://api.semanticscholar.org/CorpusID:268043908.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.

Liu, Q., Yue, J., Fang, Y., Xia, S., Fang, L., 2024. HyperMamba: A spectral-spatial adaptive mamba for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 62, 1–14. http://dx.doi.org/10.1109/TGRS.2024.3482473.

Qin, Q., Huo, Y., Huang, L., Dai, J., Zhang, H., Zhang, W., 2024. Deep neighborhood-preserving hashing with quadratic spherical mutual information for cross-modal retrieval. IEEE Trans. Multimed..

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. PMLR, pp. 8748–8763.

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

Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T., 2013. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. pp. 785–796.

Su, S., Zhong, Z., Zhang, C., 2019. Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3027–3035.

Tu, J., Liu, X., Hao, Y., Hong, R., Wang, M., 2024. Two-step discrete hashing for cross-modal retrieval. IEEE Trans. Multimed. 1–12.

Tu, J., Liu, X., Lin, Z., Hong, R., Wang, M., 2022. Differentiable cross-modal hashing via multimodal transformers. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 453–461.

Vaswani, A., 2017. Attention is all you need. arXiv preprint arXiv:1706.03762.

Wang, X., Zou, X., Bakker, E.M., Wu, S., 2020. Self-constraining and attention-based hashing network for bit-scalable cross-modal retrieval. Neurocomputing 400, 255–271.

Wu, S., Yuan, X., Xiao, G., Lew, M.S., Gao, X., 2024. Deep cross-modal hashing with multi-task latent space learning. Eng. Appl. Artif. Intell. 136, 108944.

Xie, X., Li, Z., Li, B., Zhang, C., Ma, H., 2024. Unsupervised cross-modal hashing retrieval via dynamic contrast and optimization. Eng. Appl. Artif. Intell. 136, 108969.

Xie, Y., Zeng, X., Wang, T., Xu, L., Wang, D., 2021. Matching images and texts with multi-head attention network for cross-media hashing retrieval. Eng. Appl. Artif. Intell. 106, 104475.

Xie, Y., Zeng, X., Wang, T., Xu, L., Wang, D., 2022. Multiple deep neural networks with multiple labels for cross-modal hashing retrieval. Eng. Appl. Artif. Intell. 114, 105090.

Xu, C., Chai, Z., Xu, Z., Li, H., Zuo, Q., Yang, L., Yuan, C., 2022. HHF: Hashing-guided hinge function for deep hashing retrieval. IEEE Trans. Multimed. 25, 7428–7440.

Yang, F., Han, M., Ma, F., Ding, X., Zhang, Q., 2023. Label embedding asymmetric discrete hashing for efficient cross-modal retrieval. Eng. Appl. Artif. Intell. 123, 106473.

Zhang, H., Cheng, S., Du, A., 2024. Multi-stage auxiliary learning for visible-infrared person re-identification. IEEE Trans. Circuits Syst. Video Technol. 34 (11), 12032–12047. http://dx.doi.org/10.1109/TCSVT.2024.3425536.

Zhang, J., Peng, Y., Yuan, M., 2018. Unsupervised generative adversarial cross-modal hashing. In: Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32.

Zhao, S., Chen, H., Zhang, X., Xiao, P., Bai, L., Ouyang, W., 2024. RS-mamba for large remote sensing image dense prediction. IEEE Trans. Geosci. Remote Sens. 62, 1–14. http://dx.doi.org/10.1109/TGRS.2024.3425540.

Zhu, J., Sheng, M., Ke, M.-C., Huang, Z., Chang, J.-Y., 2023. CLIP multi-modal hashing: A new baseline CLIPMH. arXiv abs/2308.11797. URL: https://api.semanticscholar.org/CorpusID:261075920.

Zhuo, Y., Li, Y., Hsiao, J., Ho, C., Li, B., 2022. Clip4hashing: Unsupervised deep hashing for cross-modal video-text retrieval. In: Proceedings of the 2022 International Conference on Multimedia Retrieval. pp. 158–166.