Chinese Journal of Information Fusion
http://dx.doi.org/10.62762/CJIF.2024.361895

REVIEW ARTICLE

Bridging Modalities: A Survey of Cross-Modal Image-Text Retrieval

Tieying Li[1,2,3], **Lingdu Kong**[1], **Xiaochun Yang**[1,4,*], **Bin Wang**[1,2,3] **and Jiaxing Xu**[5]

[1] School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China
[2] National Frontiers Science Center for Industrial Intelligence and Systems Optimization, Northeastern University, Shenyang 110819, China
[3] Key Laboratory of Data Analytics and Optimization for Smart Industry (Northeastern University), Ministry of Education, China
[4] Software College, Northeastern University, Shenyang 110169, China
[5] School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore

Abstract

The rapid advancement of Internet technology, driven by social media and e-commerce platforms, has facilitated the generation and sharing of multimodal data, leading to increased interest in efficient cross-modal retrieval systems. Cross-modal image-text retrieval, encompassing tasks such as image query text (IQT) retrieval and text query image (TQI) retrieval, plays a crucial role in semantic searches across modalities. This paper presents a comprehensive survey of cross-modal image-text retrieval, addressing the limitations of previous studies that focused on single perspectives such as subspace learning or deep learning models. We categorize existing models into single-tower, dual-tower, real-value representation, and binary representation models based on their structure and feature representation. A key focus is placed on the fusion of modalities to enhance retrieval performance across diverse data types. Additionally, we explore the impact of multimodal Large Language Models (MLLMs) on cross-modal fusion and retrieval. Our study also provides a detailed overview of common datasets, evaluation metrics, and performance comparisons of representative methods. Finally, we identify current challenges and propose future research directions to advance the field of cross-modal image-text retrieval.

Keywords: multi-modal data, cross-modal retrieval, cross-modal alignment, cross-modal fusion, large language models.

1 Introduction

The advent of Internet technology, driven by social media and e-commerce platforms, offers a convenient way to generate and share multimodal data. Efficient and accurate retrieval of relevant information from vast multimodal data has garnered increased interest from

Academic Editor:
You He

Submitted: 03 April 2024
Accepted: 08 June 2024
Published: 12 June 2024

Vol. 1, **No.** 1, 2024.
10.62762/CJIF.2024.361895

***Corresponding author:**
✉ Xiaochun Yang
yangxc@mail.neu.edu.cn

Citation
Li, T., Kong, L., Yang, X., Wang, B., & Xu, J. (2024). Bridging Modalities: A Survey of Cross-Modal Image-Text Retrieval. *Chinese Journal of Information Fusion*, 1(1), 79–92.

researchers due to its extensive real-world applications. Cross-modal image-text retrieval enables semantic search of instances in one modality (e.g., image) based on queries from another modality (e.g., text). Cross-modal image-text retrieval typically includes two main tasks: image query text (IqT) retrieval and text query image (TqI) retrieval. The formal definition is as follows:

The multimodal training set, denoted as $O = o_i i = 1^n$, consists of n instances. Each instance $o_i = (v_i, t_i, y_i)$ comprises an original image sample v_i, a text sample t_i, and a label annotation vector $y_i = [yi1, \ldots, y_{iC}]$, where C is the number of classes. Each annotation y_{iz} equals 1 if the instance o_i belongs to the z-th class, and y_{iz} equals 0 otherwise $(1 \leq z \leq C)$. The testing set $Q = q_{i=1}^{m}$ consists of m query instances, where $q_i = (v_i, t_i)$. For each query sample v_i or t_i, samples of the other modality that are semantically relevant should be returned.

Deep learning-based cross-modal image-text retrieval has achieved great success due to deep models that can effectively extract semantic information from visual and language data of different modalities.

Furthermore, with the success of large language models (LLMs) like ChatGPT, multimodal Large Language Models (MLLMs) have emerged, drawing more attention from researchers. Several previous efforts have surveyed cross-modal image-text retrieval. However, current surveys often classify cross-modal retrieval models from only a single perspective (e.g., subspace learning model or deep learning model), leading to insufficiently thorough results. Moreover, there is a lack of analysis on the cross-modal retrieval capabilities of the latest multimodal large language models. Inspired by this, we present a more comprehensive and up-to-date survey of cross-modal image-text retrieval in this paper.

Figure 1. Illustration of the classification of cross-modal retrieval model from two perspectives..

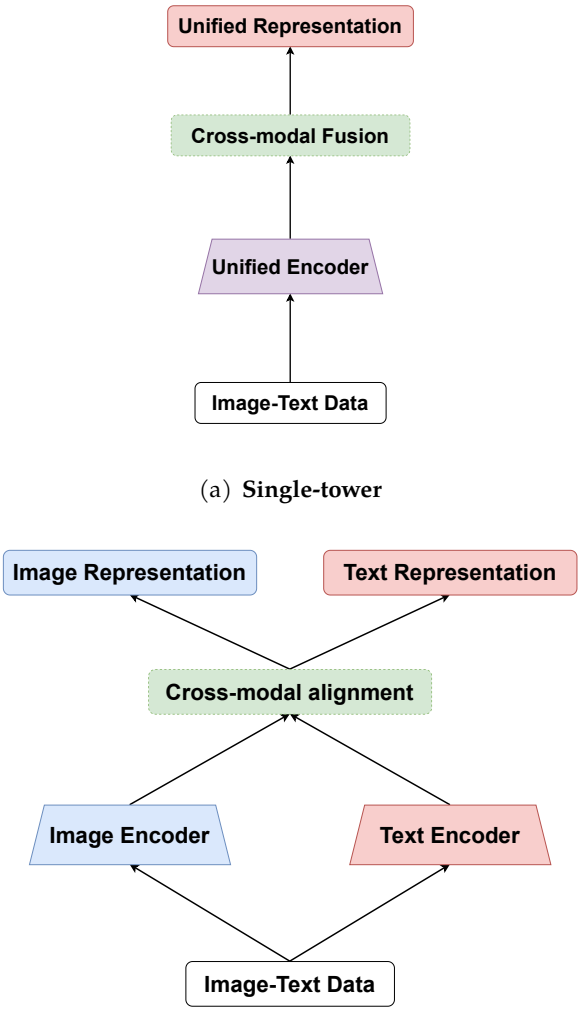
The two most critical factors influencing cross-modal image-text retrieval systems are model structure and feature representation. We classify existing models based on these two key aspects to provide a more thorough analysis of cross-modal image-text retrieval. Figure 1 illustrates our classification of cross-modal retrieval models.

- Single-tower models, also known as single-stream models, utilize a unified architecture to process both modalities simultaneously. These models integrate the modalities early, aiming to learn joint representations directly. They are beneficial for capturing complex interactions but may face scalability and fusion challenges.
- Dual-tower models, also known as two-stream models, use separate architectures (towers) for each modality. These models process each modality separately, allowing for specialized processing and scalability. However, they must ensure compatibility between the independently learned representations for effective retrieval.
- Real-value representation models involve encoding data into continuous vectors in a high-dimensional space. These vectors typically consist of floating-point numbers. These models are suitable for capturing detailed and complex relationships. However, they incur high computational and storage costs, making them less ideal for large-scale data applications.
- Binary representation models encode data into compact, fixed-length binary codes (e.g., hash vectors of bits). These models offer efficient storage and fast retrieval, making them well-suited for large-scale databases. However, they may sacrifice some accuracy and require sophisticated projection models to learn effective binary codes.

Based on above classification, we summarize the representative cross-modal image-text retrieval methods, as depicted in Table 1. The structure of our study is outlined as follows: First, we summarize cross-modal image-text retrieval models based on the above taxonomy in Section 2. Section 3 introduces MLLMs and focuses on their capabilities in cross-modal retrieval tasks. Section 4 provides a detailed overview of common cross-modal image-text datasets, evaluation metrics, and accuracy comparisons among representative approaches. Section 5 summarizes the challenges identified in the preceding review and outlines meaningful research directions for the future.

2 Deep Learning-Based

This section reviews recent research on cross-modal image-text retrieval using deep-learning neural networks. These models typically involve two main components: feature extraction from each modality



(a) **Single-tower**

(b) **Dual-tower**

Figure 2. Illustration of single-tower and dual-tower structure.

and feature alignment or fusion through an alignment or fusion module. The primary goal is to learn a common semantic subspace that preserves semantic correlations both within and across modalities. We categorize these models based on their structure and feature representation into four categories: single-tower models, dual-tower models, real-valued representation models, and binary representation models.

2.1 Single-tower models

Single-tower (single-stream) architecture models process image and text features through a shared encoder, like a transformer, as shown in Figure 2 (a). These models usually combine the two input modalities early in the network and jointly process them through shared encoders. The main motivation behind single-tower models is their ability to directly learn joint representations of the two modalities, capturing complex interactions between them. By using shared layers to process both modalities together, these models aim to learn rich, fused representations that benefit cross-modal retrieval tasks.

In this review, we focus on the role of modality fusion in enhancing retrieval performance. Particularly, single-tower models exemplify early fusion by embedding image and text inputs into a unified semantic space through shared encoders. This strategy enables deeper interactions between modalities, yielding richer representations. Additionally, MLLM-based approaches, such as BLIP-2, implement hybrid fusion through modules like Q-Former, allowing semantic alignment across modalities at multiple stages.

The ViLT (Vision and Language Transformer) model [8] presents an innovative method for multi-modal training, drawing inspiration from the Vision Transformer (ViT) mechanism. In contrast to earlier methods that needed an object detector for region-level feature extraction, ViLT directly splits images into patches, performs linear embedding, and uses these as transformer inputs. Text data is also embedded and merged with image embeddings for joint training, significantly enhancing learning and inference efficiency. ViLT employs three pre-training objectives: Image-Text Matching (ITM), Masked Language Modeling (MLM), and Word Patch Alignment (WPA). For fine-tuning in cross-modal retrieval, ViLT initializes the similarity score head from the pre-trained ITM head and fine-tunes it with cross-entropy loss to maximize positive pair scores. Experimental results indicate that ViLT drastically reduces per-instance processing time from 900 milliseconds to 15 milliseconds, showcasing its efficiency and innovation in multi-modal learning.

Traditional pre-trained models for computer vision (CV) and natural language processing (NLP) perform well independently but face challenges with cross-modal tasks involving lengthy natural language inputs and intricate visual elements. Unicoder-VL [9] utilizes a multi-layer Transformer to learn joint representations of vision and language via cross-modal pre-training. It uses three tasks: MLM, Masked Object Classification (MOC), and Visual-linguistic Matching (VLM) The model processes linguistic and visual content simultaneously, effectively learning context-aware representations and predicting relationships between images and

Table 1. The compilation of representative cross-modal image-text retrieval methods. In the following, Pro and con represent the advantages and disadvantages of this method respectively.

Categories		Model	Technology	Pros&Cons
Deep learning-based models	Single-tower models	ViLT [8]	Convolution-free encoder Vision-language transformer Contrastive learning strategy	Pro: Reduced computational complexity. Con: Dependence on large datasets.
		Unicoder [9]	Universal encoder structure Masked object classifation Masked language modeling	Pro: Capable of handling multiple tasks across different modalities. Con: Requires significant computational resources for pre-training.
		VisualBERT [10]	OD-based region features Masked language modeling task Sentence-image prediction task	Pro: Achieves high performance with a straightforward and adaptable framework. Con: Requires extensive pre-training on image-caption datasets like COCO for optimal performance.
	Dual-tower models	ViLBERT [11]	Two-stream architecture Co-attentional transformer layers Masked multi-modal learning	Pro: Adaptable to a variety of tasks with minor modifications. Con: Requires substantial resources for pre-training and fine-tuning.
		CLIP [12]	Contrastive language-image pre-training Zero-shot transfer learning	Pro: Capable of zero-shot transfer to multiple tasks without additional training. Pro: Demonstrates competitive performance across a wide range of vision-language benchmarks. Con: Trains on 400 million (image, text) pairs collected from the internet.
		ALIGN [14]	Dual-encoder architecture Normalized softmax loss Noisy data filtering	Pro: Effective zero-shot classification without fine-tuning. Con: Performance might vary significantly across different datasets. Con: Requires extensive preprocessing to handle noisy data.
	Real-value representation models	ACMR [15]	Adversarial learning framework Triplet ranking loss Gradient reversal layer	Pro: Effective Subspace Representation could preserves semantic structure across modalities . Con: Performance depends on careful tuning of hyperparameters .
		DSCMR [16]	Weight sharing strategy Modality invariance loss	Pro: Learns discriminative and modality-invariant features. Con: Performance may vary across different data modalities.
		IEFT [17]	Channelwise feature enhancement	Pro: Mitigates feature misalignment by simultaneous image and text embedding. Con: Focuses on intermodal differences, not affinities.
		COTS [18]	Two-stream pre-training Momentum contrastive learning Masked vision-language modeling	Pro: Highest performance among all two-stream methods. Pro: 10,800X faster in inference compared to the latest single-stream methods. Con: Performance might be affected by the quality of pre-training.
		TEAM [19]	Token embeddings interaction Token embeddings alignment block	Pro: Explicit alignment enhances fine-grained similarity measurement. Con: Potential increase in computational complexity due to explicit alignment.
		CLIP4CMR [34]	Modality-specific MLP Prototype contrastive loss	Pro: Using modality imbalance robustness: maintains performance in imbalanced settings. Pro:The compact representation makes the method effective even in low-dimensional spaces. Con: Involves intricate model and loss function design and tested on a few benchmark datasets only.
	Binary representation models	DCMH [20]	Hash-code learning Deep neural networks for end-to-end learning Cross-modal hashing for multimedia retrieval	Pro: The first method of multimedia retrieval using hash code. Pro: Low storage cost and fast query speed. Con: Performance may be affected by the quality of hand-crafted features.
		UDCMH [21]	Self-taught learning Binary latent factor models	Pro: Enables multimodal data search in a self-taught manner. Pro: Preserves both the nearest and farthest neighbors of data. Con: Performance might be affected by the quality of unsupervised learning.
		SSAH [22]	Adversarial learning framework Self-supervised semantic network Two adversarial networks	Pro: Applicable to diverse forms of multimodal data. Pro: Transformation to a continuous space potentially improves the efficiency of the hashing. Con: Performance may be affected by the quality of self-supervised semantic information.
		Bi-CMR [23]	Bidirectional reinforcement module Adjusted similarity matrix	Pro: Realizing the assumption that "label annotations reliably reflect the instance relevance" conflicts with human perception. Pro: Dynamically adjust the similarity matrix based on the label network. Con: Requires intricate bidirectional learning setup.
		DCHMT [24]	Selecting mechanism Transformer-based encoder Differentiable hashing function	Pro: Incorporates location information using transformers. Pro: Uses gradient descent, bypassing NP-hard discrete optimization issues. Con: Require further validation on a wider range of datasets to assess generalizability.
		DSPH [36]	Semantic-aware proxy loss Multi-Modal Irrelevant Loss	Pro: Maintains fine-grained similarity ranking among samples. Con: Involves intricate proxy-based mechanisms.
Multimodal large language models		BLIP-2 [1]	Q-former Two-stage fine tuning Retrieval model preselection	Pro: Leverage frozen pre-trained image encoders and language models. Pro: Utilize significantly fewer trainable parameters. Con: Lack of in-context learning capability due to pretraining on single image-text pairs.
		InternLM [2]	Perceive sampler Two-stage fine tuning Retrieval model preselection	Pro: Generate coherent articles by intelligently identifying areas for image integration. Pro: Comprehension with rich multilingual knowledge by training on a multi-modal database. Con: Lack of established metrics for quantitatively assessing text-image composition.
		EIIRwQR [3]	Image caption model Zero-shot Queries augmented by LLMs	Pro: Refine queries based on user relevance feedback. Pro: Incorporate VLM and LLM denoiser for refining query expansions. Con: Misleading feedback and query mismatch issues.
		CIREVL [6]	Fixed pattern generated by LLMs Zero-shot	Pro: Without the need for task-specific training. Pro: Offer modular language reasoning. Con: Require further refinement to address nuanced variations in target modifications.
		CbIR [5]	Chat-based text query Single-stage fine tuning Query embeddings generated by LLMs	Pro: A user-friendly chat-based image retrieval system. Pro: Engage in conversation to clarify user search intent. Con: Require extensive training and tuning of dialogue generation modules.
		GRACE [4]	Unique identifier token Two-stage fine tuning	Pro: Memorize and recall images within their parameters. Pro: More comprehensive and contextually relevant responses to user queries. Con: Face challenges in effectively encoding information within the parameters of MLLMs.

texts. Pre-training on large-scale image-caption pairs allows it to excel in downstream tasks such as image-text retrieval and visual commonsense reasoning. Unicoder-VL achieves state-of-the-art results in image-text retrieval on the MSCOCO and Flickr30K datasets, showcasing strong generalization abilities. However, its reliance on pre-training datasets might limit performance on tasks that require domain-specific knowledge.

A flexible model is needed to handle various vision-and-language tasks, capturing detailed semantics from both modalities without complex architectures. VisualBERT [10] integrates BERT with pre-trained object detection systems, processing image features and text together using Transformer layers. It is pre-trained on the COCO dataset using visually-grounded language model objectives such as masked language modeling and sentence-image prediction The model's design enables it to implicitly align language elements and image regions through self-attention, capturing intricate associations without explicit supervision. VisualBERT's design emphasizes simplicity and flexibility in handling diverse tasks.

Single-stream methods represent a powerful approach for cross-modal retrieval, leveraging unified Transformer architectures to effectively bridge the gap between different modalities. While these models perform well on general datasets, fine-tuning them for specific domains may require additional data and computational adjustments.

2.2 Dual-tower models

Dual-stream cross-modal methods, aim to integrate and process information from multiple modalities, such as text, images, and audio. These methods are characterized by their ability to handle the heterogeneity and complexity inherent in multimodal data, thereby facilitating a richer and more comprehensive understanding and generation of content. The primary challenge addressed by dual-stream cross-modal methods is the effective alignment and fusion of disparate data types, which often possess different structures, noise levels, and contextual nuances. The dual stream cross-modal approach typically involves two parallel processing streams, each dedicated to handling a specific modality, as shown in Figure 2 (b).

ViLBERT [11] aims to tackle the challenge of jointly understanding and reasoning about vision and language, which is difficult due to the inherent differences and complexities of each modality. It employs a two-stream model in which one stream processes visual information and the other processes linguistic information. These streams interact via a co-attentional Transformer layer that enables each modality to attend to the other. The key innovation is the co-attentional Transformer layer, which facilitates the interaction between visual and linguistic representations, allowing the model to learn rich, joint representations of both modalities.

CLIP [12] meets the need for models that can understand and connect images and text flexibly, particularly for zero-shot learning tasks where the model must generalize to new concepts without explicit training. CLIP employs separate encoders for images and text, training them with a contrastive loss to align image and text embeddings in a shared space. The model is trained on a vast dataset of images and their corresponding captions from the internet. The key innovation is using contrastive learning to align visual and textual representations, enabling the model to perform zero-shot learning by leveraging the rich, diverse data it was trained on. CLIP demonstrates impressive performance on various tasks without fine-tuning, including image classification, image-text retrieval, object detection, and generating text descriptions for images.

In [14], the authors introduce ALIGN (A Large-scale ImaGe and Noisy-text embedding), which utilizes a massive dataset of over one billion image-alt text pairs collected from the web with minimal filtering. The core of ALIGN is a straightforward dual-encoder architecture that employs contrastive learning to align visual and language representations in a shared embedding space. The ALIGN model uses a dual-encoder architecture with separate encoders for images and text. The encoders are trained with a contrastive loss to align the embeddings of matching image-text pairs. During training, the model applies simple frequency-based filtering on the dataset. The contrastive loss function helps in bringing together the embeddings of matched pairs and separating those of non-matched pairs. ALIGN achieves 76.4% top-1 accuracy on ImageNet without using any of its training samples and sets new state-of-the-art results on Flickr30K and MSCOCO benchmarks. In addition to the basic dual-encoder designs, some recent studies further enhance retrieval quality by promoting representation diversity. For instance, Kim et al. [13] proposed a method that integrates a set of diverse embeddings to enrich the semantic space,

improving the robustness of cross-modal retrieval across varying query intents and data distributions.

Dual-stream methods offer a robust framework for cross-modal retrieval by utilizing specialized pathways for different modalities and aligning their outputs in a shared space. By effectively aligning embeddings and using tailored processing, these models achieve strong performance in retrieving relevant content across heterogeneous data types, showcasing their value in multimodal applications.

2.3 Real-value representation models

Non-hashing methods based on real-valued representations effectively reduce the semantic gap between different modalities by learning dense feature representations, thereby enhancing retrieval precision. By employing deep learning methods to model features of various modalities and extract deep semantic features, these methods effectively address the issue of feature heterogeneity in cross-modal data. They also emphasize semantic correspondence between modalities, narrowing the semantic gap to improve the accuracy of cross-modal data matching, thereby increasing retrieval precision.

ACMR [15] tackles the challenge of aligning visual and textual data for cross-modal retrieval tasks, where traditional methods often fail to bridge the semantic gap between different modalities. The proposed solution involves employing adversarial training to learn robust cross-modal representations. Specifically, ACMR utilizes a dual-stream architecture where each modality is processed separately, with an adversarial loss to align the embeddings in a shared space. The key innovation of ACMR is the integration of adversarial learning, which encourages the model to produce modality-invariant features. This approach ensures that visual and textual representations are more closely aligned, thereby improving retrieval accuracy. ACMR significantly enhances the performance of cross-modal retrieval tasks, demonstrating improved alignment between visual and textual data and higher retrieval accuracy compared to non-adversarial methods. However, adversarial training can be complex and computationally intensive, and it may lead to potential instability during training.

DSCMR [16] addresses the challenge of learning effective representations for cross-modal retrieval tasks, where existing methods often struggle to capture the complex relationships between different modalities. The proposed solution employs a deep supervised approach that utilizes labeled data to learn discriminative features for each modality. DSCMR uses a dual-stream network with deep neural networks for both visual and textual data, supervised by a cross-modal ranking loss. The innovation in DSCMR lies in its application of deep supervision and a cross-modal ranking loss, ensuring that the learned representations are both discriminative and aligned across modalities. DSCMR achieves state-of-the-art performance in cross-modal retrieval tasks, showcasing the effectiveness of deep supervision and ranking-based training objectives in improving retrieval accuracy. However, DSCMR requires large amounts of labeled data and is potentially prone to overfitting to specific datasets.

IEFT [17] tackles the challenge of enhancing feature interactions for cross-modal retrieval, where traditional models often fail to fully capture the intricate relationships between visual and textual data. The proposed solution, Interacting-Enhancing Feature Transformer (IEFT), uses a Transformer-based architecture to enhance feature interactions between modalities. IEFT processes visual and textual features in separate streams and employs attention mechanisms to integrate them. The key innovation of IEFT is its use of Transformer-based attention mechanisms to enhance interactions between visual and textual features, allowing the model to learn richer and more nuanced representations. IEFT demonstrates superior performance on cross-modal retrieval benchmarks, benefiting from enhanced feature interactions and the powerful representation capabilities of Transformers.

COTS [18] addresses the difficulty of effectively combining visual and textual information for cross-modal retrieval, where existing methods may not fully leverage the potential of collaborative learning between modalities. The solution involves a Collaborative Two-Stream (COTS) architecture, where two streams process visual and textual data independently but collaborate through shared intermediate representations and alignment losses. The innovation in COTS lies in its collaborative learning mechanism, which ensures that the two streams not only process their respective modalities effectively but also learn from each other through shared representations. While collaborative learning enhances feature alignment and robust performance across various tasks, it increases complexity due to collaboration mechanisms and potential synchronization issues between streams.

TEAM [19] addresses the issue of aligning token embeddings from different modalities for cross-modal retrieval, where conventional methods may not fully capture the semantic relationships between visual and textual data. The proposed solution, Token Embeddings AlignMent (TEAM), employs alignment strategies to ensure that token embeddings from different modalities are closely related in a shared space. TEAM utilizes dual-stream networks with alignment losses to achieve this goal. TEAM's key innovation is its specific focus on token-level alignment, ensuring that individual tokens from text and corresponding visual elements are accurately aligned in the embedding space. TEAM significantly improves cross-modal retrieval performance by ensuring precise alignment of token embeddings, leading to better semantic understanding and retrieval accuracy. However, it incurs potentially high computational costs for fine-grained alignment and complexity in managing token-level interactions.

2.4 Binary representation models

Real-valued cross-modal image-text retrieval methods based on deep learning use feature vectors directly obtained from feature extraction for modeling and retrieval. However, with the explosive growth of multimedia data, such as short videos on TikTok or image-text information on Weibo, multimodal data often reaches hundreds of thousands, millions, or even billions of instances. This requires that the retrieval process for multimodal data ensures both precision and efficiency. Among various retrieval methods, hashing methods have gained widespread attention due to their low storage cost, efficiency, and fast retrieval speed, making them more suitable for large-scale datasets.

Hashing methods map feature vectors from the original feature space to binary codes (Hamming space) to save storage space and increase retrieval speed while maintaining the similarity between data points during the mapping process. Subsequently, the Hamming distance between the hash codes of the query data and those in the database is calculated for similarity ranking, ultimately yielding the retrieval results. Calculating the Hamming distance is faster than other distance metrics such as Euclidean and cosine distances. Additionally, storing data as binary codes rather than real-valued ones reduces the storage requirements for retrieval tasks.

Learning hash functions mainly involves dimensionality reduction and quantization. Dimensionality reduction maps the information from the original space to a lower-dimensional space, such as mapping an image's original pixel space information to a lower-dimensional (e.g., tens of dimensions) representation. Quantization involves linear or nonlinear transformations of the original features and binary segmentation of the feature space to produce hash codes. As mentioned in the problem definition section of cross-modal retrieval, there is a semantic gap between different forms (modalities) of data representation. Minimizing this semantic gap remains a primary challenge for cross-modal retrieval hashing methods. Generally, there are two approaches to address this: one is learning a unified hash code, and the other is using supervised information, such as labels, to collaboratively represent and minimize the distance between hash codes of semantically relevant instances.

DCMH [20] addresses the challenge of efficiently retrieving relevant data across different modalities (e.g., text and images) by using hashing techniques to map high-dimensional data into compact binary codes. The proposed solution utilizes a deep learning framework to generate hash codes for each modality through learning shared representations. These representations are optimized to maintain semantic similarity across different modalities, ensuring related items have similar hash codes. This is the first use of deep hashing neural networks to learn these representations, allowing the model to capture complex relationships between modalities and generate more accurate hash codes.

UDCMH [21] addresses the challenge of cross-modal retrieval without labeled data, which is significant since traditional supervised methods rely heavily on labeled training examples. The key innovation is the unsupervised learning approach, which eliminates the need for labeled data and still achieves effective cross-modal retrieval by learning from the data's inherent structure. This approach demonstrates strong performance in cross-modal retrieval tasks, especially in scenarios where labeled data is scarce or unavailable. However, its performance may not match supervised methods on well-labeled datasets and may be sensitive to the quality of the data structure.

SSAH [22] tackles the challenge of generating robust hash codes for cross-modal retrieval by leveraging the advantages of both self-supervised learning and adversarial training. Self-supervised learning generates initial hash codes, while adversarial

training refines these codes to ensure they are modality-invariant and semantically meaningful. This combination enables the model to learn effective representations without the need for extensive labeled data. SSAH achieves enhanced retrieval performance and robustness, demonstrating the effectiveness of its novel training strategy.

Bi-CMR [23] is the first to recognize that the assumption "label annotations reliably reflect instance relevance" conflicts with human perception. It proposes a new evaluation method to guide the learning of instance hash codes consistent with human perception. Bi-CMR introduces a novel bidirectional reinforcement-guided hashing method that reinforces hash code learning through mutual promotion. The key innovation is using reinforcement learning to dynamically adjust and improve the hashing process, ensuring the generated hash codes are effective for cross-modal retrieval. Bi-CMR demonstrates superior performance in cross-modal retrieval tasks, with hash codes that are well-aligned and optimized for retrieval accuracy.

DCHMT [24] tackles the challenge of effectively integrating and hashing data from multiple modalities using a unified framework. It constructs a multi-modal transformer to capture detailed cross-modal semantic information and introduces a micro-hashing module to map modal representations into hash codes. UCMFH tackles the need for effective cross-modal retrieval without labeled data, focusing on learning robust hash codes through unsupervised methods. The proposed solution uses unsupervised contrastive learning to generate hash codes. By leveraging contrastive learning, the model maximizes the similarity between related items across modalities while minimizing the similarity between unrelated items. UCMFH demonstrates strong performance in unsupervised cross-modal retrieval tasks, achieving high accuracy and robustness by effectively learning from the inherent structure of data.

Overall, real-valued representations are suitable for tasks that require high precision, while hashing representations are ideal for applications that need rapid, large-scale retrieval.

3 Multimodal Large Language Models

In the past two years, large language models (LLMs) have made significant strides, demonstrating the ability to perform many NLP downstream tasks in a zero-shot setting. However, their inference capabilities with data from other modalities have been limited. To address this gap, MLLMs have been proposed. These models are capable of not only generating and understanding complex text but also processing image information, allowing a single MLLM to handle multiple multimodal downstream tasks simultaneously. Utilizing MLLMs for image-text retrieval has emerged as a powerful and widely applied technique. By integrating natural language processing and computer vision technologies, MLLMs can efficiently extract information from vast datasets, achieving precise image-text matching and search.

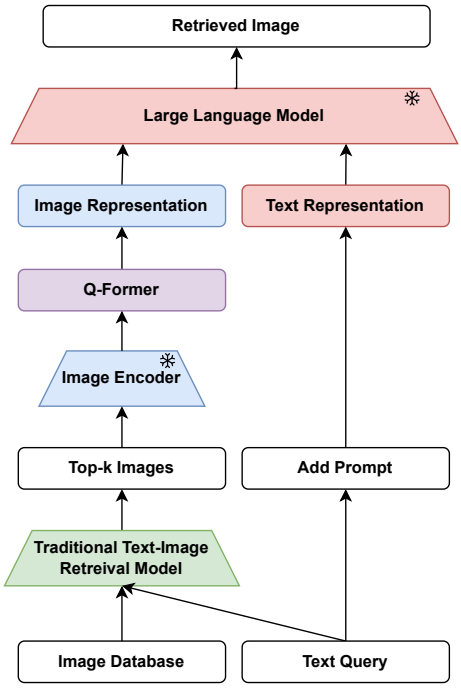


Figure 3. Example of BLIP-2's pipeline for text-to-image retrieval.

Before introducing this section, we first differentiate between VLP models and MLLMs. We define VLP as a multimodal pre-training model tailored for specific tasks involving vision and language. In contrast, MLLMs are pre-trained models capable of addressing multiple complex reasoning tasks across different modalities. The key distinction lies in their ability to handle multiple downstream tasks. Therefore, VLP models are not classified within this section. Our categorization is based on the core components and capabilities of the models.

The process of using MLLMs for image-text retrieval generally includes the following steps:

- Using an MLLM trained on large-scale data and fine-tuning it with an image-text retrieval dataset.
- Employing specific prompts to complete the image-text retrieval task.
- Involving smaller image-text retrieval models to assist the MLLM in the task.

BLIP-2 [1] employs a bidirectional retrieval approach by leveraging pre-trained image models and large language models. The text-to-image retrieval pipeline used by BLIP-2 is illustrated in Figure 3. This pipeline is enhanced with Q-Former to bridge the gap between modalities, using a two-stage training process: initially training the image model, followed by the text model. The retrieval process begins with a common retrieval model selecting 128 candidate images based on image-text similarity. These candidate images, along with the query text, are then input into the model, which selects the most relevant image as the retrieval result. Essentially, this approach utilizes generative models to perform the retrieval task, ensuring accurate and efficient matching of images based on textual input.

InternLM [2] focuses solely on image retrieval. It involves fine-tuning both the Perceive Sampler and the MLLM, followed by fine-tuning Perceive Sampler with LoRA. Initially, CLIP is used to select the top-k candidate images, from which the MLLM selects one image as the final retrieved result. This approach, like the previous one, is fundamentally generative.

EIIRwQR [3] also focuses on image retrieval, utilizing a VLM to generate a set of candidate images. Each candidate image is described with a caption generated by an image description model. The MLLM takes the original query and these generated captions as input, modifying each query. The VLM then uses the modified queries for image retrieval. This process is iterated multiple times to refine the final retrieval result. The MLLM is employed only during the inference stage without any fine-tuning, categorizing this approach as using MLLMs for data augmentation.

CIREVL [6] focuses on image retrieval without any training process, addressing the high labor costs associated with annotated data. It employs an MLLM to transform the text into a fixed descriptive sentence format, which is then used by a traditional model for image retrieval. he MLLM is utilized only during the inference stage and is not fine-tuned, effectively categorizing this approach as using MLLMs for data augmentation.

In CbIR [5], dialogues are used as input. The accumulated dialogue information, processed with a contrastive loss function, fine-tunes the large model to obtain 256-dimensional retrieval vectors. These vectors are then compared with 256-dimensional image vectors using cosine similarity to retrieve the images.

GRACE [4] involves assigning each image a unique image token and training the instruction to predict the identifier for the <image token>. During inference, the model predicts the image identifier corresponding to the given query.

In fact, aside from the methods mentioned above, most MLLMs can potentially be employed for image-text retrieval tasks, although many of these models have not been specifically tested for this purpose. Additionally, existing MLLM methods tested for image-text retrieval typically involve LLMs trained solely on text data. However, there are models like Google's Gemini [7], which are inherently multimodal. Instead of a two-stage process where the model is first trained on text and then on images, these models are pre-trained on multimodal data from the beginning. Such inherently multimodal models exhibit greater adaptability and robustness with multimodal data. Future exploration of these native multimodal LLMs may further enhance the performance of image-text retrieval.

In summary, the existing works highlight various approaches to utilizing MLLMs for image-text retrieval. The methods range from leveraging pre-trained models and fine-tuning specific components to employing generative techniques and using MLLMs for data augmentation without additional training. These diverse strategies underscore the flexibility and potential of MLLMs in enhancing image-text retrieval tasks, paving the way for more accurate and efficient retrieval systems in the future.

4 Datasets and Evaluation

4.1 Datasets

The researchers have proposed various datasets for cross-modal image-text retrieval, including Wikipedia [25], NUS-WIDE [26], TC-12 [27], Flickr [28], Pascal Sentence [29], etc. The most frequently used datasets are summarized as MSCOCO [30] and Flickr30K [28]. MS COCO dataset contains 123,287 images from the Microsoft Common Objects in Context (COCO) dataset, each paired with five human-generated textual captions.

Table 2. The MAP@ALL results of real-value cross-modal image-text retrieval methods. The experiment results are from [38]and [37].

Task	Methods	Source	Wikipedia	Pascal-Sentence	NUS-WIDE	Xmedia
IqT	ACMR [15]	ACM MM17	0.468	0.538	0.519	0.536
	CM-GANS [32]	TMM18	0.521	0.603	0.536	0.567
	DSCMR [16]	CVPR19	0.521	0.674	0.611	0.697
	AGCN [33]	IEEE CSVT22	0.620	0.683	-	-
	CLIP4CMR [34]	ARXIV22	0.592	0.698	0.609	0.746
IqT	ACMR [15]	ACM MM17	0.412	0.544	0.542	0.519
	CM-GANS [32]	TMM18	0.466	0.604	0.551	0.551
	DSCMR [16]	CVPR19	0.478	0.682	0.615	0.693
	AGCN [33]	IEEE CSVT22	0.532	0.683	-	-
	CLIP4CMR [34]	ARXIV22	0.574	0.692	0.621	0.758

Table 3. The MAP@ALL results of binary cross-modal image-text retrieval methods. The experiment results are from [38]and [37].

Task	Methods	Source	MirFlickr			NUS-WIDE			MS COCO		
			16bits	32bits	64bits	16bits	32bits	64bits	16bits	32bits	64bits
IqT	DCMH [20]	CVPR17	0.724	0.731	0.731	0.568	0.561	0.596	0.505	0.536	0.557
	SSAH [22]	CVPR18	0.903	0.922	0.925	0.691	0.727	0.728	0.632	0.669	0.668
	DCHUC [35]	TKDE20	0.895	0.916	0.926	0.707	0.672	0.738	0.513	0.550	0.558
	HMAH [31]	TMM23	0.960	0.965	0.969	0.813	0.825	0.840	0.691	0.732	0.763
	DSPH [36]	TCSVT23	0.925	0.940	0.945	0.852	0.905	0.929	0.793	0.815	0.833
TqI	DCMH [20]	CVPR17	0.764	0.749	0.780	0.558	0.591	0.616	0.549	0.572	0.605
	SSAH [22]	CVPR18	0.896	0.906	0.915	0.658	0.673	0.666	0.583	0.556	0.664
	DCHUC [35]	TKDE20	0.764	0.749	0.780	0.558	0.591	0.616	0.549	0.572	0.605
	HMAH [31]	TMM23	0.915	0.925	0.938	0.783	0.796	0.814	0.800	0.869	0.904
	DSPH [36]	TCSVT23	0.897	0.904	0.911	0.859	0.920	0.935	0.792	0.800	0.819

After removing rare words, the average caption length is 8.7 words. The dataset is divided into $82,783$ training image-text pairs, $5,000$ validation pairs, and $5,000$ test pairs. Model evaluations are conducted on five folds of $1,000$ test pairs and the entire set of $5,000$ test pairs. Flickr30K0 comprising $31,000$ images sourced from the Flickr website, each image in this dataset is annotated with five textual descriptions. The dataset is split into three sections: $1,000$ image-text pairs for validation, $1,000$ pairs for testing, and the remaining for training.

4.2 Evaluation

We summarize the following evaluation metrics widely used to assess cross-modal retrieval tasks: Mean Average Precision@K (MAP@K): MAP calculates the average precision for each query and then averages these values over all queries. In the experimental validation of MLLMs, the R@n metric is commonly used, indicating the proportion of queries for which at least one correct result is retrieved within the top-n results.

4.3 Result Analysis

In this section, we present the accuracy of several representative methods in cross-modal retrieval tasks. As shown in Tables 2- 4, we compare the accuracy of cross-modal retrieval methods using common measures for each task. Based on the presented performance, we can summarize the following observations:

- As shown in Table 2, in cross-modal real-valued retrieval, methods based on VLP (Vision-Language Pre-training) or transformer structures often achieve better accuracy. This improvement is due to the enhanced ability of encoders to extract semantic information, as demonstrated by the performance of CLIP4CMR.

- As shown in Table 3, cross-modal hashing retrieval methods exhibit progressive accuracy with different hash code lengths. Most methods show an increase in accuracy as the code length increases, indicating that longer codes can represent more semantic information, thereby

Table 4. The R@n results of MLLM methods. The experiment results are from their papers.

Task	Methods	Source	Flickr30K			MS-COCO(5K)		
			R@1	R@5	R@10	R@1	R@5	R@10
IqT	AGREE (FT only) [39]	WSDM23	0.916	0.987	0.992	-	-	-
	AGREE [39]		0.921	0.987	0.992	-	-	-
	BLIP-2 ViT-L [1]	ICML23	0.969	1.000	1.000	0.835	0.960	0.980
	BLIP-2 ViT-g [1]		0.976	1.000	1.000	0.854	0.970	0.985
TqI	GRACE [4]	ARXIV24	0.684	0.889	0.937	0.415	0.691	0.791
	AGREE (FT only) [39]	WSDM23	0.781	0.951	0.978	-	-	-
	AGREE [39]		0.828	0.959	0.978	-	-	-
	BLIP-2 ViT-L [1]	ICML23	0.886	0.976	0.989	0.663	0.865	0.918
	BLIP-2 ViT-g [1]		0.897	0.981	0.989	0.683	0.877	0.926

improving retrieval accuracy. However, the accuracy improvement from 32-bit to 64-bit codes is often not as significant as the improvement from 16-bit to 32-bit codes. This may be because once an optimal hash code length is achieved, longer vector lengths do not provide additional valuable semantic information for retrieval.

- As shown in Table 4, the experimental results of MLLMs demonstrate that most methods can retrieve the correct result within the top-5 results. Some models even achieve a 100% recall rate on the validation set. These results highlight that training or fine-tuning MLLMs on large-scale language and image datasets enables the models to capture subtle details and semantic variations in both text and images. This approach not only enhances the models' generalization capabilities but also reduces the dependency on large amounts of annotated data, a significant advantage over traditional models. However, this benefit comes at the cost of requiring substantially more computational resources for training and inference due to the large number of parameters in these models.

5 Conclusion and Future Works

This survey has comprehensively reviewed the field of cross-modal image-text retrieval, categorizing existing methods and highlighting their strengths and limitations. Current cross-modal retrieval methods can be broadly classified into single-tower, dual-tower, real-value representation, and binary representation models.

1) Summary of Existing Methods. Single-tower models integrate modalities early, learning joint representations that capture complex interactions. Their unified architecture, however, may struggle with scalability and efficient fusion of different data types. Dual-tower models process each modality separately through specialized architectures, enhancing scalability and tailored processing. Yet, they face challenges in ensuring compatibility between separately learned representations. Real-value representation models encode data into continuous, high-dimensional vectors, effectively capturing detailed and complex relationships. Despite their accuracy, they are computationally intensive and costly in terms of storage, making them less suitable for large-scale applications. Binary representation models use compact, fixed-length binary codes for data encoding, offering efficient storage and fast retrieval. These models are ideal for large-scale databases but often trade-off some accuracy and require sophisticated techniques to learn effective binary codes.

2) Advantages and Problems. Advantages: Single-tower models. Effective in capturing intricate interactions between modalities. Dual-tower models. Highly scalable and adaptable to specialized processing needs. Real-value representation models. High accuracy in representing complex relationships. Binary representation models. Efficient in storage and fast in retrieval, suitable for large datasets.

Problems: Single-tower models. Scalability issues and challenges in modality fusion. Dual-tower models. Difficulty in ensuring compatibility of learned representations. Real-value representation models. High computational and storage costs. Binary representation models. Potential loss of accuracy and complexity in learning effective binary codes.

3) Future Directions. To advance the field of cross-modal retrieval, future research should focus on several key areas: 1. Improving Model Compatibility and Fusion: Developing hybrid models that leverage

the strengths of both single-tower and dual-tower architectures to enhance compatibility and fusion efficiency. 2. Enhancing Computational Efficiency: Designing novel methods that reduce computational and storage demands of real-value representation models without compromising accuracy. 3. Advanced Binary Coding Techniques: Innovating more sophisticated binary coding methods that balance accuracy and efficiency, making them viable for large-scale applications. 4. Leveraging Multimodal Large Language Models (MLLMs): Further exploring the potential of MLLMs in enhancing cross-modal retrieval tasks, particularly in improving semantic understanding and retrieval accuracy. 5. Comprehensive Benchmarking: Establishing more robust benchmarking frameworks that include diverse datasets and comprehensive evaluation metrics to better assess model performance. 6. Addressing Scalability and Real-world Applications: Developing scalable solutions that can handle real-world data complexities and large-scale multimodal databases, ensuring the practical applicability of cross-modal retrieval systems.

By addressing these challenges and focusing on these future directions, the field of cross-modal image-text retrieval can achieve more robust, efficient, and accurate systems, enhancing the practical utility of these technologies in various real-world applications.

Data Availability Statement

Not applicable.

Funding

This work was supported in part by the National Natural Science Foundation of China under Grant U22A2025, Grant 62072088, Grant 62232007, Grant U23A20309, and Grant 61991404; in part by the Liaoning Provincial Science and Technology Plan Project - Key R&D Department of Science and Technology under Grant 2023JH2/101300182; in part by the 111 Project under Grant B16009.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

[2] Zhang, P., Wang, X. D. B., Cao, Y., Xu, C., Ouyang, L., Zhao, Z., ... & Wang, J. (2023). Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*. [CrossRef]

[3] Zhu, H., Huang, J. H., Rudinac, S., & Kanoulas, E. (2024). Enhancing Interactive Image Retrieval With Query Rewriting Using Large Language Models and Vision Language Models. *arXiv preprint arXiv:2404.18746*. [CrossRef]

[4] Li, Y., Wang, W., Qu, L., Nie, L., Li, W., & Chua, T. S. (2024). Generative Cross-Modal Retrieval: Memorizing Images in Multimodal Language Models for Retrieval and Beyond. *arXiv preprint arXiv:2402.10805*. [CrossRef]

[5] Levy, M., Ben-Ari, R., Darshan, N., & Lischinski, D. (2024). Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems, 36*.

[6] Karthik, S., Roth, K., Mancini, M., & Akata, Z. (2023). Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*. [CrossRef]

[7] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J. B., Yu, J., ... & Ahn, J. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*. [CrossRef]

[8] Kim, W., Son, B., & Kim, I. (2021, July). Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning* (pp. 5583-5594). PMLR.

[9] Li, G., Duan, N., Fang, Y., Gong, M., & Jiang, D. (2020, April). Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 07, pp. 11336-11344). [CrossRef]

[10] Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2019). Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*. [CrossRef]

[11] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems, 32*.

[12] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.

[13] Kim, D., Kim, N., & Kwak, S. (2023). Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 23422-23431).

[14] Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., ... & Duerig, T. (2021, July). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning* (pp. 4904-4916). PMLR.

[15] Wang, B., Yang, Y., Xu, X., Hanjalic, A., & Shen, H. T. (2017, October). Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 154-162). [CrossRef]

[16] Zhen, L., Hu, P., Wang, X., & Peng, D. (2019). Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10394-10403).

[17] Tang, X., Wang, Y., Ma, J., Zhang, X., Liu, F., & Jiao, L. (2023). Interacting-Enhancing Feature Transformer for Cross-modal Remote Sensing Image and Text Retrieval. *IEEE Transactions on Geoscience and Remote Sensing*. [CrossRef]

[18] Lu, H., Fei, N., Huo, Y., Gao, Y., Lu, Z., & Wen, J. R. (2022). Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer Vision and pattern recognition* (pp. 15692-15701).

[19] Xie, C. W., Wu, J., Zheng, Y., Pan, P., & Hua, X. S. (2022, October). Token embeddings alignment for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 4555-4563). [CrossRef]

[20] Jiang, Q. Y., & Li, W. J. (2017). Deep cross-modal hashing. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3232-3240).

[21] Wu, G., Lin, Z., Han, J., Liu, L., Ding, G., Zhang, B., & Shen, J. (2018, July). Unsupervised Deep Hashing via Binary Latent Factor Models for Large-scale Cross-modal Retrieval. In *IJCAI* (Vol. 1, No. 3, p. 5).

[22] Li, C., Deng, C., Li, N., Liu, W., Gao, X., & Tao, D. (2018). Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4242-4251).

[23] Li, T., Yang, X., Wang, B., Xi, C., Zheng, H., & Zhou, X. (2022, June). Bi-CMR: bidirectional reinforcement guided hashing for effective cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 9, pp. 10275-10282). [CrossRef]

[24] Tu, J., Liu, X., Lin, Z., Hong, R., & Wang, M. (2022, October). Differentiable cross-modal hashing via multimodal transformers. In *Proceedings of the 30th ACM International Conference on Multimedia* (pp. 453-461). [CrossRef]

[25] Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., & Vasconcelos, N. (2010, October). A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 251-260). [CrossRef]

[26] Chua, T. S., Tang, J., Hong, R., Li, H., Luo, Z., & Zheng, Y. (2009, July). 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). [CrossRef]

[27] Escalante, H. J., Hernández, C. A., Gonzalez, J. A., López-López, A., Montes, M., Morales, E. F., ... & Grubinger, M. (2010). The segmented and annotated IAPR TC-12 benchmark. *Computer vision and image understanding, 114*(4), 419-428. [CrossRef]

[28] Huiskes, M. J., & Lew, M. S. (2008, October). The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval* (pp. 39-43). [CrossRef]

[29] Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010, June). Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk* (pp. 139-147).

[30] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & 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 (pp. 740-755). Springer International Publishing. [CrossRef]

[31] Tan, W., Zhu, L., Li, J., Zhang, H., & Han, J. (2022). Teacher-student learning: Efficient hierarchical message aggregation hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*. [CrossRef]

[32] Peng, Y., & Qi, J. (2019). CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 15*(1), 1-24. [CrossRef]

[33] Dong, X., Liu, L., Zhu, L., Nie, L., & Zhang, H. (2021). Adversarial graph convolutional network for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(3), 1634-1645. [CrossRef]

[34] Zeng, Z., & Mao, W. (2022). A comprehensive empirical study of vision-language pre-trained model for supervised cross-modal retrieval. *arXiv preprint arXiv:2201.02772*. [CrossRef]

[35] Tu, R. C., Mao, X. L., Ma, B., Hu, Y., Yan, T., Wei, W., & Huang, H. (2020). Deep cross-modal hashing with hashing functions and unified hash codes jointly learning. *IEEE Transactions on Knowledge and Data Engineering, 34*(2), 560-572. [CrossRef]

[36] 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 Transactions on Circuits and Systems for Video Technology*. [CrossRef]

[37] Zhu, L., Wang, T., Li, F., Li, J., Zhang, Z., & Shen, H. T. (2023). Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions. *arXiv preprint arXiv:2308.14263*. [CrossRef]

[38] Zhou, K., Hassan, F. H., & Hoon, G. K. (2023). The State of the Art for Cross-Modal Retrieval: A Survey. *IEEE Access*. [CrossRef]

[39] Wang, X., Li, L., Li, Z., Wang, X., Zhu, X., Wang, C., ... & Xiao, Y. (2023, February). AGREE: aligning cross-modal entities for image-text retrieval upon vision-language pre-trained models. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining* (pp. 456-464). [CrossRef]

Tieying Li is currently a PhD student at the School of Computer Science and Engineering, Northeastern University. His current research interests include cross-modal retrieval and multi-task learning. (Email: tieying@stumail.neu.edu.cn)

Lingdu Kong is currently a PhD student at the School of Computer Science and Engineering, Northeastern University. His research interests include visual content-based indexing and retrieval. (Email: 2210707@stu.neu.edu.cn)

Xiaochun Yang received the Ph.D. degree in Computer Software and Theory from Northeastern University, 110004, China, in 2001. She is currently the dean of Software College and a professor in the School of Computer Science and Engineering with Northeastern University, China. Her research interests include Big Data management and knowledge engineering, data quality management, data privacy preserving, and recommender systems. More details about her research can be found at http://faculty.neu.edu.cn/yangxiaochun. (Email: yangxc@mail.neu.edu.cn)

Bin Wang received the Ph.D. degree in computer science and technology from Northeastern University, Shenyang, China, in 2008. He is currently a Professor with the School of Computer Science and Engineering, Northeastern University. His research interests include big data management and knowledge engineering, database theory and technology, cloud computing, and data privacy preserving. More details about his research can be found at http://faculty.neu.edu.cn/wangbin. (Email: binwang@mail.neu.edu.cn)

Jiaxing Xu is a PhD candidate at School of Computer Science and Engineering, Nanyang Technological University, Singapore. His research interests include graph representation learning and brain network analysis. (Email: jiaxing003@e.ntu.edu.sg)