IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 June 2024, accepted 13 August 2024, date of publication 16 August 2024, date of current version 29 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3444817

TOPICAL REVIEW

Cross-Modal Retrieval: A Review of Methodologies, Datasets, and Future Perspectives

ZHICHAO HAN, AZREEN BIN AZMAN, (Member, IEEE), MAS RINA BINTI MUSTAFFA, (Senior Member, IEEE), AND FATIMAH BINTI KHALID, (Member, IEEE)
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang 43400, Malaysia
Corresponding authors: Azreen Bin Azman (azreenazman@upm.edu.my) and Zhichao Han (gs61677@student.upm.edu.my)

ABSTRACT With the rapid development of science and technology, all types of mixed media contain large amounts of data. Traditional single multimedia data can no longer satisfy daily requirements. Therefore, the cross-modal retrieval technology has become an urgent requirement. Consequently, there is a pressing need for cross-modal retrieval technology. Its purpose is to mine the connection between different modal samples, that is, to retrieve another modal sample with approximate semantics through one modal sample. For example, users can retrieve multimedia data such as images or videos with text. However, there are differences in the modal representation of different types of multimedia data, and measuring the correlation between different modes is the main problem of cross-modal retrieval. Currently, the most popular deep learning methods have achieved remarkable results in the field of data processing and graphics. Many researchers have applied deep learning methods to cross-modal retrieval to solve the problem of similarity measurement between different multimedia data. By summarizing the relevant paper methods of cross-modal retrieval, this paper provides a definition of cross-modal retrieval problems, reviews the core ideas of the current mainstream cross-modal retrieval methods in the form of three main methods, lists the commonly used data sets and evaluation methods, and finally analyzes the problems and future research trends of cross-modal retrieval.

INDEX TERMS Cross-modal retrieval, deep learning, review.

I. INTRODUCTION

A. BACKGROUND AND MOTIVATION

Cross-modal retrieval, alternatively termed multi-modal retrieval, addresses the task of retrieving pertinent information across diverse modalities, such as text and images. With the substantial surge in online multimedia data and escalating need for efficient information retrieval, cross-modal retrieval has garnered significant interest in recent years. However, the prevalent search technologies in the market predominantly focus on single-mode internal retrieval, such as keyword-based retrieval [1] and content-based retrieval [2]. These methods exclusively conduct similarity searches within the same media type, such as text, images, audio, and video retrieval. In contrast, cross-modal retrieval requires the establishment of a cross-modal relationship model. This enables users to employ a unified pattern for information searches while retrieving pertinent data across different modalities [3], thereby improving the accessibility and comprehensiveness of search results. Figure 1 shows the general framework for cross-modal retrieval. Its basic purpose is to explore the relationship between different modal samples, that is, to retrieve another modal sample with approximate semantics through one modal sample. Thus, the cross-modal retrieval challenge centers on quantifying the content similarity across various modalities, commonly referred to as the heterogeneity gap [4]. The emergence of deep-learning technology, offers a promising avenue for

The associate editor coordinating the review of this manuscript and approving it for publication was Fu Lee Wang.

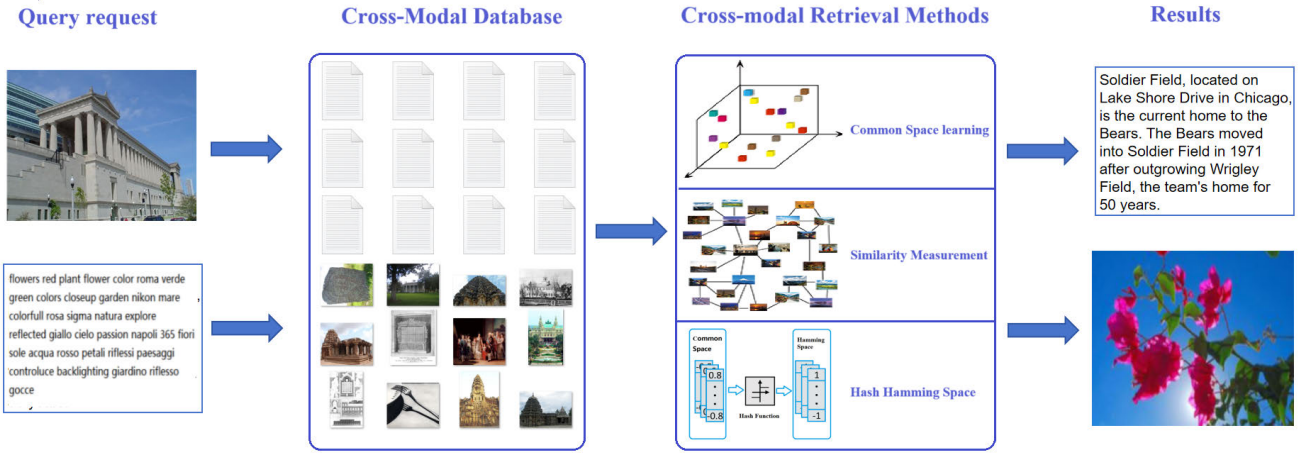

FIGURE 1. The general framework of cross-modal retrieval.

addressing heterogeneity problems through representation learning. yielding remarkable performance and exerting a substantial influence in both academic and industrial domains.

Cross-modal image and text retrieval involves visual and natural language descriptions, emphasizing the interplay between the image and text modalities. Its objective is to retrieve images through text queries and vice versa, without depending on additional auxiliary information. In cross-modal retrieval, different modal data exhibit potential feature heterogeneity and a high-level semantic correlation [5]. Text on web pages is usually represented by dictionary vectors, whereas images are represented by visual features, which are located in completely different feature spaces but represent the same semantic topic. The primary challenge in cross-modal retrieval is quantifying the similarity across distinct modalities, as they have heterogeneous and nonhomologous features. The main research method is to model the relationships between different modalities, learn common latent space representations [6], and then measure the similarity in the common latent space to achieve cross-modal retrieval. Over the years, many researchers and scholars have proposed effective methods for this issue, especially for cross-modal retrieval of text and image modalities. There have also been many review articles on cross-modal retrieval, such as Wang et al., who published a review article titled ''A Comprehensive Survey on Cross-modal Retrieval'' in 2016 [7]. In their review, they divided cross-modal retrieval into real-valued representation learning methods and binary representation hash learning methods and provided a detailed introduction to these methods. Peng et al. from Peking University published an article titled ''An Overview of Cross-media Retrieval: Concepts, Methodologies, Benchmarks and Challenges'' in 2018 [3], which not only elaborates on the concepts and methods of cross-media retrieval, but also further introduces the cross-media retrieval dataset, especially the new dataset created by their team called Xmedia, which includes five categories of data modalities, providing more convenience for future researchers in cross-media retrieval. In 2020, Kaur et al. [8] published a comprehensive review on ''Comparative analysis on cross-modal information retrieval'' which comprehensively elaborated on cross-modal retrieval from various aspects such as research background, research methods, benchmark datasets, etc. The figures and table information in this article are worth learning for researchers and can serve as a benchmark for beginners in the field of cross-modal retrieval. Although these reviews have performed well in the field of cross-modal retrieval, the sources of some proprietary terms are not marked, making it difficult for beginners to further understand and learn. In addition, in the past three years, many new methods and technologies related to cross-modal retrieval have been proposed, and have achieved very good data (compared to the paper three years ago, the retrieval accuracy has improved by approximatedly 10 percent). Therefore, it is time to update the literature review, supplement the latest technologies and methods in recent years, and provide efficient assistance for new scholars in the future.

The main objective of this review is to provide a comprehensive analysis of cross-modal retrieval techniques, including classic canonical correlation analysis (CCA) methods, real-valued representation learning methods, and emerging cross-modal retrieval hashing methods. We aim to introduce the basic principles, models, and implementation processes of these technologies in a detailed and clear manner so that readers can gain a comprehensive understanding of cross-modal retrieval. The main contributions of this study are as follows:

- We have systematically elaborated on the benchmark and latest methods for cross-modal retrieval between images and texts, and classified different implementation methods and technologies, which will contribute to the research on cross-modal retrieval.
- Some terms, benchmark research methods, benchmark datasets, and data measurement methods related to cross-modal retrieval have been annotated and linked to facilitate future researchers in literature search and

reference, leaving them with considerable query time for their future research.
- The current challenges and future development and research directions of cross-modal retrieval between images and text are introduced, providing assistance and significance for future researchers.

B. ORGANIZATIONAL STRUCTURE OF THE ARTICLE

The first chapter of this article provides an introduction, mainly explaining what cross-modal retrieval is, its research methods, the main problems it currently faces, and the current research status. In Chapter 2, based on the structure of cross-modal retrieval models, we first introduced the tools and methods for extracting image and text features. Subsequently, combined with clear diagrams and data tables, we comprehensively reviewed the comparison and performance of traditional CCA methods and other methods extended from CCA research methods, cross-modal retrieval methods based on real-valued representation learning method, and cross-modal retrieval hashing methods. In Section III, we review the benchmark dataset and its sources and measurement methods, which will be of great help to future scholars. In Section IV, we provide an outlook and elaboration of future research methods for cross-modal retrieval. Finally, a summary of the study is presented.

II. CROSS-MODAL RETRIEVAL TECHNIQUES REVIEW

In this section, we provide an overview of cross-modal retrieval based on the organizational structure shown in Figure 2 and the key technologies popular in different periods. Real-value representation learning directly learns features extracted from different modalities; Binary representation learning, on the other hand, involves mapping the features extracted from different modalities to a Hamming binary space, and then learning within this space. Real-value representation learning methods focus more on semantic matching between images and texts. The currently popular deep learning based real-value representation cross-modal image and text retrieval methods can be divided into two categories: common space learning and cross-modal similarity measurement. The former type of method focuses on modeling features between different modalities, effectively solving the problem of heterogeneous features in different modal data through good feature extraction, thereby ensuring retrieval accuracy. The latter type of method focuses on the semantic correspondence between image and text modalities, aiming to improve the accuracy of image text matching and retrieval by reducing the semantic gap.

A. DATA FEATURE EXTRACTION

The first step in cross-modal retrieval is the extraction of features from the images and texts. Convolutional Neural Networks (CNNs) [22] have demonstrated remarkable efficacy in tasks related to image classification and have been are extensively utilized for image representation in cross-modal retrieval. CNNs can automatically learn discriminative

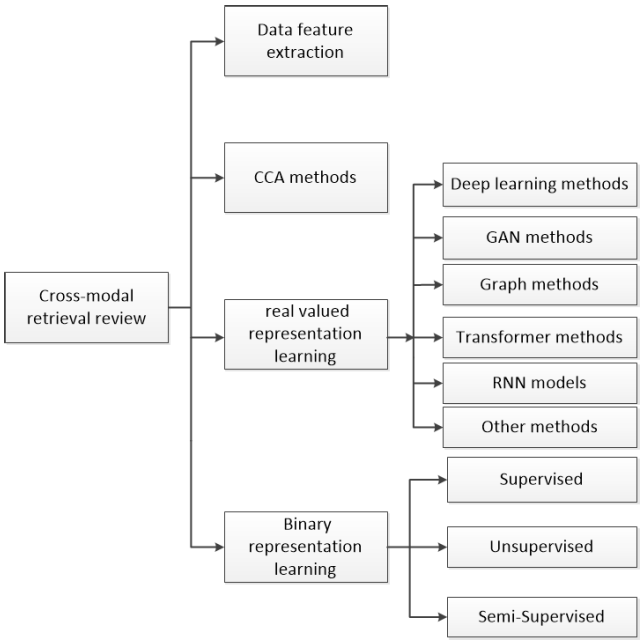

FIGURE 2. An overview of cross-modal retrieval based on the organizational structure.

features by capturing local patterns and global structures in the images. In CNN-based cross-modal retrieval, images are fed as input to the CNN architecture. The architecture comprises numerous convolutional, pooling, and fully connected layers. Convolutional layers extract local features via filters, whereas pooling layers reduce the spatial dimensions of the feature maps. The fully connected layers capture high-level semantic information. Pre-trained CNN models, such as AlexNet [23], VGGNet [24], and ResNet [25], trained on large-scale image classification datasets (e.g., ImageNet), are commonly used for image representation. These models can extract high-level features that are transferable across different tasks, including cross-modal retrieval. From Table 1, we can see that the choice of model depends on the specific task requirements, available computing resources, and size of the dataset. Usually, for general image classification tasks, VGGNet is a good choice. If the task requires handling very deep network structures or specific complex tasks, ResNet may be more suitable.

Various techniques can be employed to incorporate textual information into CNN-based cross-modal retrieval. One approach is to convert text into fixed-length vectors using word embeddings, such as Word2Vec or GloVe [26]. These embeddings capture semantic word relationships and facilitate text representations within a continuous vector space. Another method involves employing the bag-of-words (BoW) representation, where each word in the text is depicted by either a count or presence indicator. Recurrent Neural Networks (RNNs) capture sequential dependencies and are extensively used for text representation in cross-modal retrieval. RNNs process text input in a sequential manner, maintaining an internal hidden state that retains information about previous inputs. Long Short-Term Memory (LSTM)

TABLE 1. Comparison of AlexNet, VGGNet, and ResNet.

Model	Network Depth	Salient Feature	Number of Layers	Parameter	Application
AlexNet	Shallow	Local Response Normalization	8 layers	60M	Image Classification
VGGNet	Deep	Small Convolution Kernel 3x3	16th or 19th layers	138M-143M	Fine-Grained Recognition
ResNet	Very Deep	Cross-Layer Residual Connection	50, 101, 152 layers, etc.	26M-60M	Processing Complex Tasks

TABLE 2. Comparison of BoW, Word2Vec, GloVe, LSTM, and GRU.

Model	Advantages	Disadvantages	Applications
BoW	Simple and intuitive, easy to implement. Suitable for simple tasks such as short text classification	Ignored the order and contextual information between words. Unable to capture the semantic relationship of words	Short text representation
Word2Vec	Ability to learn distributed representations of words and capture semantic information. Performed well on large-scale corpora	Unable to process unknown vocabulary. Requires a large amount of training data	Word Vector Learning and Semantic Similarity Calculation for Large Scale Corpus
GloVe	Make full use of global statistical information	Longer training time	Semantic similarity calculation, global representation of word vectors
LSTM	Capture long-range dependencies. Performed well in sequence modeling tasks.	Complex calculation, long training time	Sequence tasks like language modeling, machine translation, and text generation
GRU	Simple calculation, fast training speed	Loss of some long-term dependencies	Sequence tasks like language modeling, machine translation, and text generation

[27] and gated recurrent units (GRU) [28] are prevalent RNN variations designed to mitigate the vanishing gradient issue and capture long-term dependencies. Character-level RNNs operate on individual characters instead of words, and can capture fine-grained information. These models are useful when word-level representations are insufficient, such as in tasks involving misspelled or unknown words. Table 2 summarizes the advantages and disadvantages of these text feature extraction models and their applicable environments.

B. CANONICAL CORRELATION ANALYSIS (CCA) METHODS

Cross-modal retrieval data typically appear in pairs, such as image-text pairs, and the premise of cross-modal retrieval is the correlation between text content and image content. Canonical Correlation Analysis (CCA) [9] is a classical statistical method commonly employed in cross-modal retrieval tasks. CCA aims to find linear transformations for two different modalities, images and text, that maximize the correlation between their transformed representations. By maximizing the correlation, CCA seeks to capture the shared information between modalities and align them in a common subspace, thereby facilitating cross-modal retrieval. Figure 3 illustrates the fundamental structure of the CCA cross-modal retrieval technique.

The basic CCA model consists of two steps: feature extraction and correlation maximization. In the feature extraction step, the raw data from each modality are transformed into a lower-dimensional feature representation. For instance, in image retrieval, features can be extracted using techniques such as Scale-Invariant Feature Transform (SIFT) [10] or Convolutional Neural Networks (CNNs) [11]. In text retrieval, features can be derived from Word2Vec [12] or bag-of-words [13] representations. Once the features are obtained, CCA finds two projection matrices that maximize the correlation between the transformed representations. These projection matrices map the features of each modality into a common subspace, where the correlation between modalities is maximized. The CCA objective function is formulated to maximize the correlation between the projected features.

Although CCA is widely popular for its simplicity and efficiency and has quickly become a benchmark algorithm for similar algorithms, the CCA model is an unsupervised method that does not use semantic category labels, CCA cross-modal retrieval it faces several challenges and limitations. One significant challenge is the semantic gap [14] between different modalities, as CCA focuses on capturing statistical correlations rather than semantic relationships. In addition, CCA suffers from the curse of dimensionality when the number of features is large, leading to decreased retrieval performance. Furthermore, CCA assumes a linear relationship between modalities, which may limit its ability to capture complex nonlinear correlations. Some extended models based on CCA have been proposed to address the challenges and limitations of CCA. In 2010, Rasiwasia et al. [15] introduced a joint model for images and text, and analyzed the retrieval performance of different combinations of image and text representations. These combinations encompass all

permutations of the two guiding hypotheses they suggest, thereby reinforcing the notion that incorporating cross-modal correlation and semantic abstraction can enhance retrieval accuracy. In fact, a certain modality of data may have more than one semantics; therefore, focusing solely on pairwise coupling is far from sufficient. The universal representation acquired using this method may not entirely retain the inherent cross-modal semantic structure within the data. Standard linear CCA is limited to investigating the linear relationships between two sets of random variables. However, in real-world scenarios, the relationship between variables frequently exhibits non-linear characteristics. Consequently, non-linear CCA methods have been developed. Kernel CCA (KCCA), proposed by Hwang et al. in 2012 [16], is a popular nonlinear CCA algorithm. KCCA integrates the concept of a kernel function into CCA, mapping low-dimensional data to a high-dimensional feature space (referred to as the kernel function space) and facilitating correlation analysis within this space through the kernel function. Although the KCCA method solves the non-linear problem of data, owing to the unknowability of its kernel function selection, the training cost is high, and the model is relatively complex. Therefore, the DCCA(Deep Canonical Correlation Analysis) was proposed to solve these problems. With the advancement of deep neural networks, scholars have observed that methods based on neural networks are adept at managing exceedingly intricate nonlinear relationships within the data. The strategy initially involves employing neural networks for nonlinear data mapping, followed by linear CCA for correlation calculation post data transformation. In 2013, Andrew et al. [17] first proposed DCCA based on deep neural networks, which has received widespread attention. Table 3 summarizes the differences between CCA model-based methods. Nevertheless, this approach suffers from the drawbacks of having numerous model parameters and necessitating a relatively large amount of data.

Overall, none of the above methods utilize multi-label information. To accurately express multiple concepts in an image, it is necessary to fully consider multi-label information and accurately model the correlation between different modalities. Ranjan et al. [18] in 2015 introduced a Multi-Label CCA, proposing the use of cross-type data for retrieval, where each item can be associated with multiple labels while learning the common semantic space of two modalities, solving the limitation of CCA's inability to consider advanced semantic information and improving retrieval efficiency. Unlike CCA, Multi-Label CCA does not depend on explicit one-to-one matching between modalities. Instead, it utilizes multi-label information to establish corresponding relationships, resulting in one-to-many, many-to-one, and other pairing situations, forming a discriminative subspace that is more suitable for cross-modal retrieval tasks.

Similar to CCA, methods such as Partial Least Squares (PLS) [19] and Bilinear Models [20] also attempt to perform cross-modal retrieval by learning subspaces; however, these methods rely on explicit pairing between two modalities

TABLE 3. Comparison of four CCA methods.

Methods	Year	pairwise	linear/non-linear	Training cost
CCA	2010	yes	linear	no
KCCA	2012	yes	non-linear	yes
DCCA	2013	yes	non-linear	yes
mlCCA	2015	no	non-linear	yes

to establish corresponding relationships. In 2003, Li et al. introduced a novel cross-pattern association method, called cross-pattern-factor-analysis (CFA). In CFA, queries from one modality are utilized to retrieve content from another modality that employs a low-level functionality. The primary objective of subspace class methods is to learn discriminative shared subspaces primarily by maximizing the correlation. These techniques demonstrated promising outcomes in cross-modal retrieval, a prevalent limitation is their failure to consider the local data structure within each modality and the structural alignment between modalities. In fact, samples in one modality corresponding to neighboring samples in another modality should also have adjacent relationships, and vice versa.

Canonical correlation analysis enables us to summarize relationships into fewer statistical dataset while retaining the main aspects of the relationship. To some extent, the motivation for typical correlations is very similar to that of PCA. This is another dimensionality reduction technique. The CCA formulation can be expressed as follows: We have two sets of variables, I and T, which are the image and text, respectively.

$$\mathbf{I} = \begin{pmatrix} I_1 \\ I_2 \\ \vdots \\ I_p \end{pmatrix}, \quad \mathbf{T} = \begin{pmatrix} T_1 \\ T_2 \\ \vdots \\ T_q \end{pmatrix} \tag{1}$$

As shown in formula (2), we define a linear combination called U and V. U corresponds to the linear combination of the first set of variables I, and V corresponds to the second set of variables T. Each member of U is paired with a member of V, we represent member pairing combinations as (U_i, V_i).

$$\begin{aligned} U_1 &= a_{11}I_1 + a_{12}I_2 + \cdots + a_{1p}I_p \\ U_2 &= a_{21}I_1 + a_{22}I_2 + \cdots + a_{2p}I_p \\ &\vdots \\ U_p &= a_{p1}I_1 + a_{p2}I_2 + \cdots + a_{pp}I_p \\ V_1 &= b_{11}T_1 + b_{12}T_2 + \cdots + b_{1q}T_q \\ V_2 &= b_{21}T_1 + b_{22}T_2 + \cdots + b_{2q}T_q \\ &\vdots \\ V_p &= b_{p1}T_1 + b_{p2}T_2 + \cdots + b_{pq}T_q \end{aligned} \tag{2}$$

We aim to find a linear combination that maximizes the correlation between each typical variable and its members. We calculate the variance U_i and V_i for variables using the

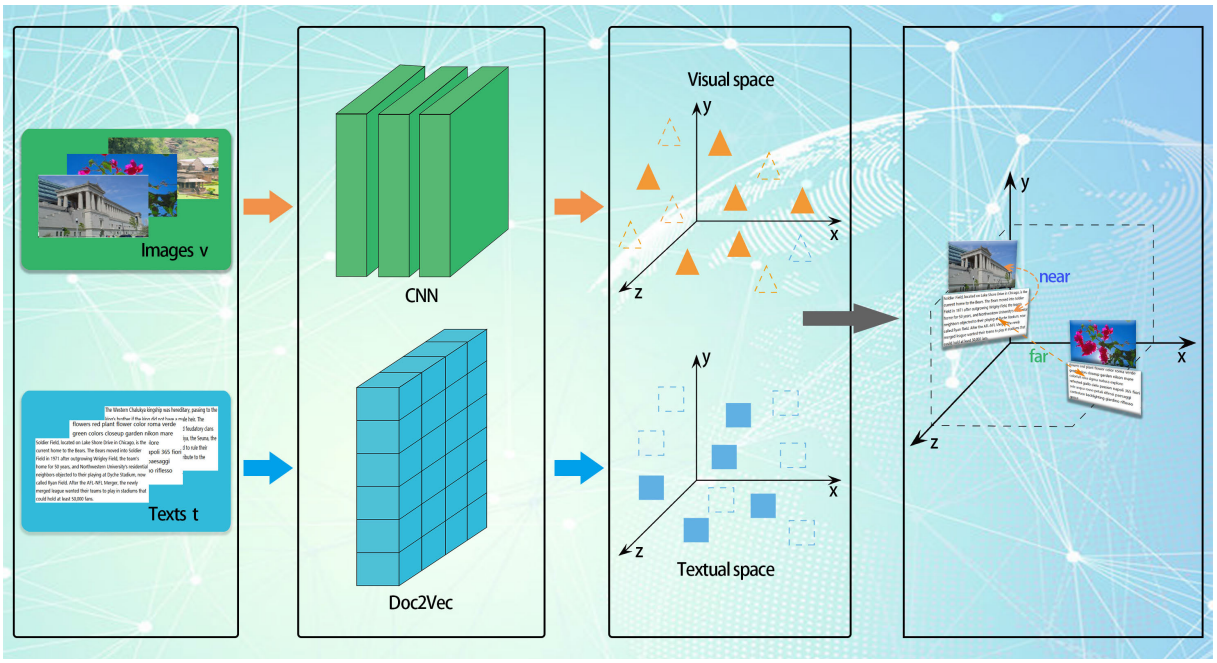

FIGURE 3. The fundamental architecture of the CCA cross-modal retrieval method.

following expression, where a is the coefficient:

$$\text{var}\,(U_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} a_{ik}a_{il}\,\text{cov}\,(X_k, X_l) \tag{3}$$

$$\text{var}\,(V_j) = \sum_{k=1}^{p}\sum_{l=1}^{q} b_{jk}b_{jl}\,\text{cov}\,(Y_k, Y_l) \tag{4}$$

The covariance between U_i and V_i is:

$$\text{cov}\,(U_i, V_j) = \sum_{k=1}^{p}\sum_{l=1}^{q} a_{ik}b_{jl}\,\text{cov}\,(X_k, Y_l) \tag{5}$$

For the correlation between U_i and V_i, we take the covariance between the two variables and divide it by the square root of the variance product:

$$\frac{\text{cov}\,(U_i, V_j)}{\sqrt{\text{var}\,(U_i)\,\text{var}\,(V_j)}} \tag{6}$$

Ultimately, the correlation between image text pairs i can be expressed as ρ_i^*.

$$\rho_i^* = \frac{\text{cov}\,(U_i, V_i)}{\sqrt{\text{var}\,(U_i)\,\text{var}\,(V_i)}} \tag{7}$$

III. REAL-VALUED REPRESENTATION LEARNING TECHNIQUES

As shown in Figure 4, in this section, we further subdivide cross-modal retrieval real-value representation methods into deep learning, adversarial neural network, graph network, transformer, and other methods.

A. DEEP LEARNING TECHNIQUES

Scholars have investigated deep learning methodologies for facilitating cross-modal retrieval tasks. Deep-learning methods can automatically learn high-level representations from raw data and capture complex relationships and semantic information. The capacity of deep learning to acquire hierarchical representations directly from unprocessed data underscores its efficacy as a potent instrument for cross-modal retrieval tasks. Ngiam et al. [29] presented a series of tasks for multimodal learning and demonstrated how to train deep networks to learn cross-modality features and how to learn a shared representation between modalities. Srivastava and Salakhutdinov [30] developed a joint density model in a multimodal input space. Subsequently, given the observed modes, the missing modes can be filled by sampling from the conditional distribution above them. Feng et al. introduced a deep- learning framework called the cross-modal correspondence autoencoder (Corr-AE) [31]. This model integrates single-modal representation learning and inter-modal correlation learning through the joint minimization of reconstruction errors from single-modal autoencoders and correlation errors across distinct modal representation layers. Wu et al. [32] treated click data as a large click graph, where vertices denote images or text queries, and edges represent the connections formed by clicks between an image and a query. Their objective was to create a multimodal representation that captures both explicit and implicit relevance relationships between these vertices within the click graph. He et al. [33] proposed a network architecture to effectively capture cross-modal retrieval properties, specifically enabling a bidirectional search. This architecture

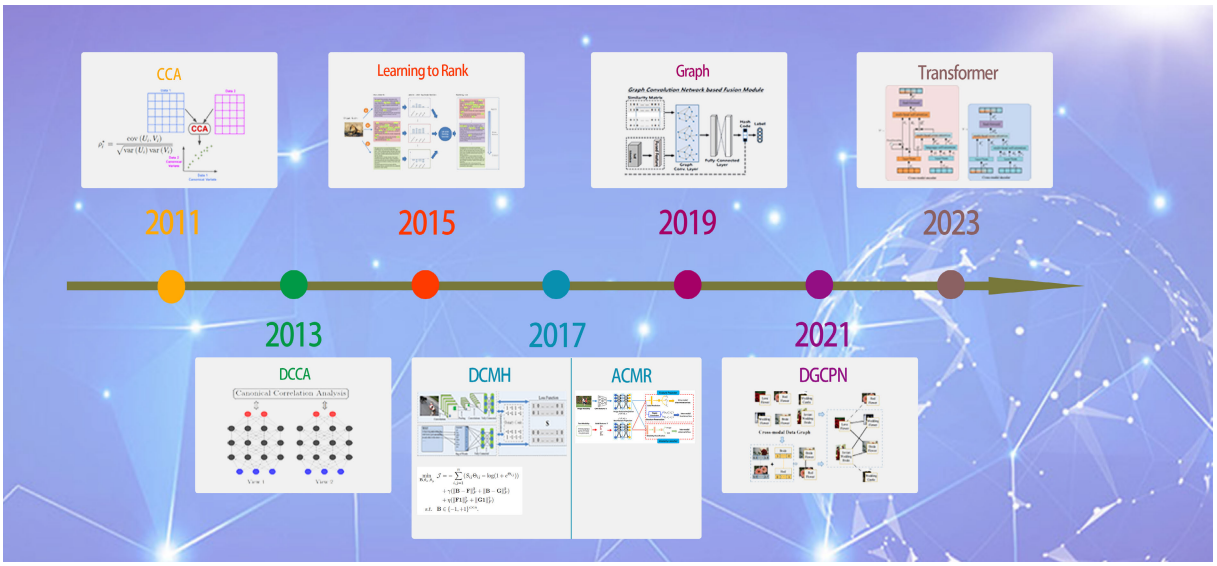

FIGURE 4. Real-valued representation learning methods' research evolution.

is distinguished by the simultaneous incorporation of both matched and unmatched image-text pairs during training. Grangier and Bengio [34] formalized the retrieval task as a ranking challenge, and introduced a learning method that optimizes a criterion linked to ranking performance. The model stands out because it does not depend on an intermediary image annotation task, setting it apart from prior research approaches. In fact, a certain modality of data may have more than one semantics; therefore, focusing solely on pairwise coupling is far from sufficient. The universal representation learned in this manner cannot fully preserve the potential cross-modal semantic structure in the data. Overall, none of the above methods utilize multi-label information. To accurately express multiple concepts in an image, it is necessary to fully consider multi-label information and accurately model the correlation between different modalities.

B. RNN MODELS

The RNN model is mainly used in cross-modal retrieval to process and understand sequential data (such as text or video), and fuse it with image features extracted by models such as CNN. RNN can effectively capture time series information in text or video, and achieve cross-modal retrieval by converting text sequences into vector representations and matching them with image or video features. Its advantage lies in its ability to process variable length sequence data, capture contextual information, and improve the matching accuracy between text, images, or videos. However, the disadvantages of RNN include high computational complexity, long training time, difficulty in capturing long-range dependencies, and the possibility of gradient vanishing or exploding when processing long sequences. Nevertheless, RNNs that combine attention mechanisms, such as LSTM or GRU, have to some extent alleviated these issues and improved the effectiveness of cross-modal retrieval. Zeng et al. proposed a Fine grained Iterative Attention Network (FIAT) for temporal language localization in videos. FIAT iteratively adjusts attention weights to accurately locate time periods in videos that match text descriptions. RNN is used to process time series information of videos and texts, while iterative attention mechanism is used to dynamically adjust the alignment relationship of different modalities, thereby improving the accuracy and robustness of localization [51]. Chen et al. proposed a Joint Visual Semantic Matching Embedding (JVSM) model for language based object retrieval tasks. This method encodes text using RNN, extracts features from images using CNN, and then matches them in a common embedding space. In order to further improve the matching effect, the author introduced Bidirectional RNN (Bi RNN) to capture the contextual information in the text [52]. Zhang et al. proposed a context aware cross-modal retrieval method that achieves accurate cross-modal retrieval through text and image embedding. The author uses RNN to process text sequences and combines image features extracted by CNN to enhance the quality of embedded representations using contextual information [53]. Neng et al. proposed an Attention Sequence to Sequence Model (ASSM) based on attention mechanism for video subtitle generation. Although this work mainly focuses on video subtitle generation tasks, the methods involved can be directly applied to video text retrieval. This model processes video frame sequences through RNN to generate corresponding text descriptions, while introducing attention mechanisms to capture the correspondence between keyframes and text segments in the video [54]. These methods that apply RNN to cross-modal retrieval tasks, combined with other techniques such as attention mechanisms and Transformers, have shown superior performance in cross-modal retrieval tasks, especially when dealing with complex queries and diverse image content.

C. GENERATIVE ADVERSARIAL NETWORK (GAN)

A Generative Adversarial Network (GAN) is an innovative architecture of deep learning, first proposed by Goodfellow et al. [35] in 2014. The basic idea is to learn the data distribution through two neural networks: a generator and a discriminator, competing with each other. Some scholars have applied GAN to the field of cross-modal retrieval, and the correlation between modalities and within modalities can be explored simultaneously in generative and discriminative models, thereby promoting cross-modal correlation learning through discriminative adversarial interactions. The ACMR model proposed by Wang et al. [36] in 2017 quickly became a benchmark for many subsequent methods. It combines common deep representation learning and adversarial learning for cross-modal retrieval. In this model, the construction of adversarial networks adopts a common-representation generation discrimination approach. In addition, semantic information is used to construct discriminative constraints for maintaining inter-modal and intra-modal data structures, thereby enhancing the ability to learn common representations. Xu et al. [37] introduced a novel model called Self-Supervised Ternary Adversarial Network (TANSS). This model leverages adversarial learning techniques to enhance the consistency and correlation of the semantic features across diverse modalities. It comprises three sub-networks that collectively form an end-to-end architecture, facilitating efficient iterative parameter optimization. Wu et al. proposed an approach known as model-specific and shared generative adversarial network (MS2GAN) [38], designed specifically for cross-modal retrieval. The generative model is trained to forecast the semantic labels of features, thereby capturing both inter-modal and intra-modal similarities with the aid of label information. This also ensures a distinction between modality-specific and modality-shared features. Simultaneously, the discriminative model learns to classify the feature modalities. These learned modality-specific and shared representations are jointly employed for retrieval purposes. Xu et al. [39] proposed a new method called Assembling AutoEncoder and Generative Adversarial Network (AAEGAN), which merges the capabilities of an autoencoder and a generative adversarial network. This combined approach facilitates simultaneous learning of a shared latent space, knowledge transfer, and feature synthesis to enable zero-shot cross-modal retrieval. He et al. [40] introduced a new approach called Category Alignment Adversarial Learning (CAAL) designed for cross-modal retrieval. They utilized two parallel generative adversarial networks that incorporated category information. These GANs generate artificial images and text features, which are then combined with pre-existing embeddings to reconstruct a shared representation. Finally, two joint discriminators are employed to minimize the disparity between the initial stage mapping and embadding of the subsequent stage. However, such methods, due to their greater focus on local information, have higher requirements for the size and quality of the dataset, as well as the refinement of the model, and are mostly not suitable for matching global information.

D. GRAPH REGULARIZATION

Graph regularization is a potent technique for cross-modal correlation learning due to its ability to capture diverse correlations within cross-modal data, encompassing semantic relevance, intra-modal similarities, and inter-modal similarities. Moreover, it naturally accommodates multiple media types within a cohesive framework. Wang et al. [41] suggested using visual and textual scene graphs—VSG and TSG—to represent images and text. Each graph captures objects and relationships within its respective modality. Liu et al. [42] introduced the Graph Structured Matching Network (GSMN), a new approach designed to capture intricate correspondences. GSMN explicitly represents objects, relations, and attributes as structured phrases. This unique representation enables the network to learn correspondences for each element separately while also facilitating the fine-grained understanding of structured phrases. Wang et al. [43] introduced an innovative cross-modal hashing method that leverages sparse graph structures. These structures are employed to utilize similarity information effectively, aiming to tackle the degradation issue commonly encountered in unsupervised algorithms. Cheng et al. [44] introduced the Cross-modal Graph Matching Network (CGMN), a graph-based approach that delves into both intra- and inter-relations without the need for network interaction. Han et al. [45] introduced a method called MGSGH, which utilizes the construction of global semantic graph and scene graph to deeply mine the coarse-grained and fine-grained semantic information of images and texts, respectively, providing an effective solution for cross-modal retrieval of hash methods. Pei et al. [46] proposed an approach that allows us to extract cross-modal features at both object and relationship levels. Assessing the similarity between images and text across these levels offers a more robust evaluation framework. This article introduces a new network, designed for matching images and text. SGSIN adeptly learns detailed semantic information in both visual and textual domains, aiming to reconcile differences between different modalities. However, constructing the graph typically incurs high time and space complexity, particularly in real-world scenarios involving extensive cross-modal datasets.

E. TRANSFORMER METHODS

In recent years, transformer architecture has been successfully applied in cross-modal retrieval, demonstrating favorable outcomes in various tasks. The transformer is a basic deep-learning model predominantly founded on the self-attention mechanism, allowing it to capture extensive dependencies within sequences and proficiently model contextual information. In 2020, Messina et al. presented the Transformer Encoder Reasoning and Alignment Network (TERAN) [47], aimed at producing comprehensive

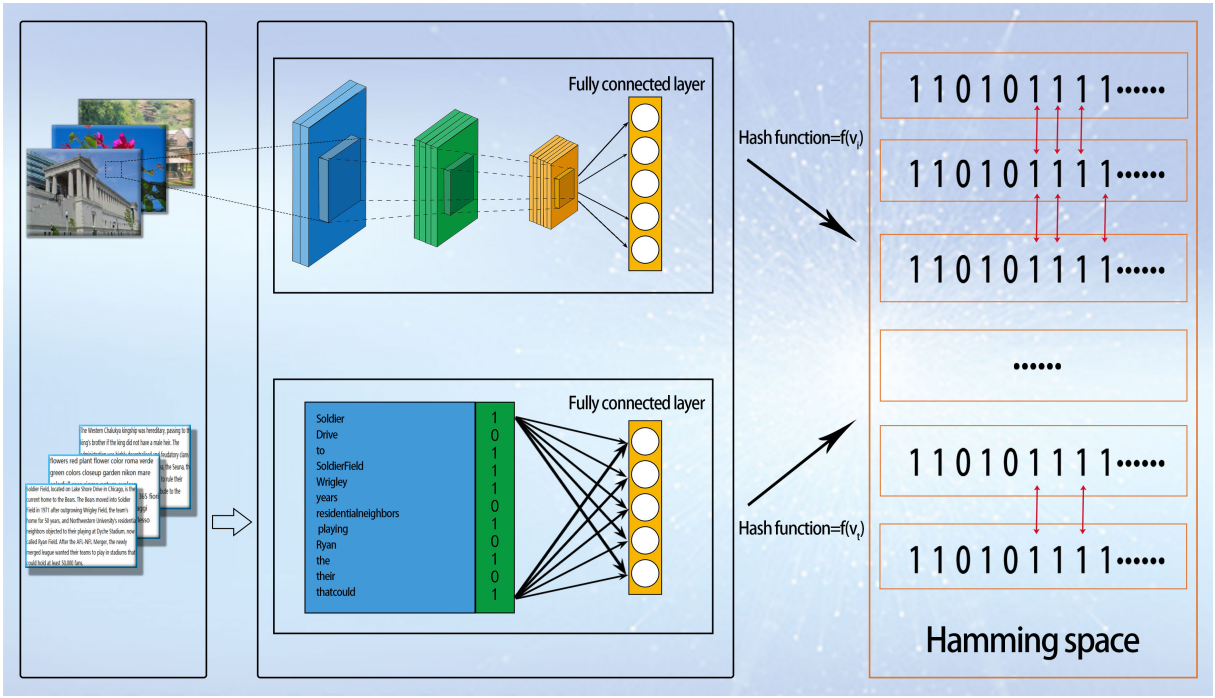

FIGURE 5. The framework of cross-modal retrieval hashing method.

region-word alignments to enhance the cross-modal information retrieval efficiency. Wang et al. [48] leveraged the structural semantics inherent in transformer architecture, enabling comprehensive cross-modal learning through unified global and fine-grained alignment. This approach effectively bridges the gap between heterogeneous modalities. Zhang et al. [49] introduced a pioneering framework for cross-modal retrieval characterized by pre-training within an initial interaction dataflow and decomposition into distinct encoders. This decomposed architecture ensures both high retrieval accuracy and rapid retrieval speeds. In 2023, Zhang et al. [50] unveiled the Cross-Modal Transformer (CMT), which incorporates a language query. This model adeptly fosters deep interactions between visual and linguistic modalities, thereby significantly boosting cross-modal capabilities. Although the transformer model has achieved many successes, we can continue to improve it in the following three directions. First, cross-modal feature fusion, the self-attention mechanism of a transformer, can be used to establish correlations between different modalities, allowing the model to focus on the correlations between different modalities when integrating features. Second, in cross-modal alignment, the transformer can help align data from different modalities in the embedding space, making it easier to compare and retrieve similarities by representing the data from different modalities in the same semantic space. Third, multimodal information fusion, through Transformer's multi-head self-attention mechanism, allows local and global information from different modal data to be effectively integrated, improving the model's ability to understand cross-modal data. In addition, transfer learning and pre-trained models leverage pre-trained transformer models to apply transfer learning to cross-modal retrieval tasks. This approach harnesses the semantic comprehension capabilities of pre-trained models to enhance overall model performance.

F. OTHER METHODS

There are also some deep learning based methods, such as learning to rank methods, metric learning methods, cross-modal reconstruction methods. They also adhere to the concept of projecting heterogeneous data into a shared space, enabling direct measurement of their similarities. For instance, the cross-modal reconstruction method focuses more on global information. This type of method is typically uses one modality information to reconstruct the corresponding modality while retaining the reconstruction information, which can enhance the consistency of the cross-modal features and semantic differentiation ability. In 2019, Wu et al. devised an alternating minimization approach for tackling the optimization problem of a multi-modal semantic autoencoder [55]. This approach entails an encoder responsible for transforming feature vectors into code vectors, while a decoder performs the reverse mapping, converting code vectors back into feature vectors. Such an encoder-decoder framework guarantees the preservation and maintenance of both feature-based and semantic information within the embeddings. Considering the challenges associated with acquiring semantic labels, Wu et al. [56] observed a consistent representation in the latent space among various view features belonging to samples within the same category.

TABLE 4. Taxonomy of binary hash learning methods.

Method	Supervision	Research Gap	References
Cross-modal Hash	Supervised	Has high computational efficiency, but weak modeling of complex semantic correlations.	[64] [65] [66] [67] [68] [69] [70] [71]
		Powerful non-linear semantic representation ability.	[72] [73] [74] [75] [76] [77] [78] [79] [80] [81] [82] [83] [84] [85]
	Unsupervised	Difficulty in effectively mining key semantic features and achieving cross-modal semantic alignment.	[86] [87] [88] [89] [90] [91] [92]
		There is a significant "heterogeneity gap" between annotated and unlabeled data.	[93] [94] [95] [96] [97] [98] [99] [44] [100] [101] [102] [103] [104] [105] [106] [107] [108] [109] [110] [111]
	Semi-Supervised	The complex correlations between multimodal data have not been fully explored.	[112] [113] [114] [115] [116] [117] [118]
		Supplementing the training set with rich unlabeled data.	[119] [120] [121] [122] [123] [124] [125] [126] [127]

Consequently, they introduced a reconstruction-based component aimed at restoring the original multimodal data to facilitate the cross-modal retrieval. Zhang et al. [57] introduced an autoencoder-based approach designed for cross-modal retrieval to obtain hash functions. This strategy aims to preserve richer and more valuable information, thereby enhancing the potency of the generated hash functions. Xie et al. [58] presented a novel methodology that integrates multi-similarity reconstruction with cluster-based contrastive hashing. This approach aims to bolster the semantic coherence of both inter-modal and intra-modal reconstructions within the cross-modal retrieval domain. The cross-modal reconstruction method utilizes deep encoders and other methods to effectively reduce the heterogeneity differences between modalities and enhance the semantic discrimination ability. This type of method does not require high training and scale requirements for datasets, has low annotation costs, is more suitable for small and medium-sized datasets, has scalability, and is often used for tasks such as image subtitle generation. However, such methods are prone to ignoring detailed information during model training, and their performance in integrating the target data for correlation is insufficient. Therefore, how to jointly learn the alignment of local text and image information while narrowing the statistical gap between modalities and dynamically adjusting the generation process between modalities based on this is currently a challenge faced by such methods. Learning to rank techniques utilize ranking information as training data, optimizing the ranking of retrieved outcomes directly, rather than focusing on similarities between pairwise data instances. Inspired by Bai et al. [59], some learning-to-rank techniques have been proposed, such as [60], [61], and [62]. They utilized ranking information as training data, optimizing the ranking of retrieved outcomes directly, rather than focusing on similarities between pairwise data instances. Wei et al. introduced a novel universal weighting metric learning framework tailored for cross-modal retrieval [63]. This framework effectively selects informative pairs and assigns suitable weight values to enhance performance. The assignment of weights is based on similarity scores, allowing for a nuanced approach in which distinct pairs benefit from varied penalty strengths.

IV. BINARY HASH METHOD

Cross-modal retrieval hashing techniques strive to compress data from various modalities into concise binary code. enabling an efficient and scalable retrieval in large-scale cross-modal databases. These methods utilize hash functions to map high-dimensional data, such as images and text, into binary codes, where similar items have codes that are close to each other. In this section, we explore various cross-modal retrieval hashing techniques. Table 4 summarizes the structure of our hash method. In both supervised and unsupervised hashing, hash functions are learned through an optimization process that minimizes the quantization error between the original data and binary codes. The learned hash functions can then be used to map new multi-modal data to the Hamming space for retrieval.

Figure 5 illustrates the basic process of the cross-modal retrieval hash method. First, the feature vectors VI and Vt of the image and text, respectively, are obtained through feature extraction. They are then mapped to the binary Hamming

space using a hash function to obtain the binary hash codes Hi and Ht of the image and text. In the Hamming space, the Hamming distance or cosine similarity is typically used to measure the similarity between two different modalities. The Hamming distance was used to calculate the difference between the digits of the two binary sequences. The smaller the difference in the digits of the sequence is, the more similar it is. The cosine similarity measures the cosine value of the angle between two vectors in the vector space; the closer the value is to 1, the higher the similarity.

Cross-modal hashing offers three key advantages over real-value representatio. First, based on binary expressions, cross-modal hashing allows the conversion of time-consuming distance calculations during retrieval into Hamming distance calculations between hash bits. Second, due to the substantially reduced storage space needed for binary hash codes in comparison to the features of the original multimedia data, such as text, image, or video features, cross-modal hashing substantially diminishes the storage overhead of multimedia data. Additionally, the length of the binary hash code acquired through learning is typically significantly shorter than the dimensionality of the original data, effectively alleviating the problem of the ''curse of dimensionality.''

Cross-modal hashing techniques can be classified into three main categories based on distinct learning methodologies: supervised, unsupervised, and semi-supervised cross-modal hashing approaches. The subsequent sections offer a detailed analysis of the current research status of these three types of cross-modal hashing methods, both domestically and internationally.

A. SUPERVISED

Supervised cross-modal hashing methods utilize labeled data to train the hash functions. These methods aim to preserve the semantic similarity between paired instances from different modalities while maximizing the Hamming distance between non-matching instances.

Supervised hash learning is one of the most common cross-modal hash learning methods. Its purpose is to use the semantic labels of data for learning, thereby mining the semantic discriminative features of multi-modal data. Over the last decade, numerous significant methodologies have emerged to enhance retrieval performance. Supervised cross-modal hashing techniques can be further categorized into shallow and deep model-based methods, based on the employed model type. Among shallow methods, methods based on eigenvalues or matrix factorization are widely used for learning cross-modal hash codes. Kumar et al. [64] constructed a set of hash functions for multimodal data and transformed the hash function learning problem into a solvable eigenvalue problem using a new relaxation method. Bronstein et al. [65] introduced an innovative approach for acquiring cross-modal hash functions by leveraging feature decomposition and boosting techniques. Their method not only mitigates redundancy but also enhanced the richness of the vocabulary used in the description. Tang et al. [66] concurrently addressed semantic label consistency and local geometric consistency within multimodal data. They employed a collaborative matrix decomposition method to compute the cross-modal hash code. This approach not only minimizes redundancy, but also enriches the lexicon employed in the explanation. Liu et al. [67] introduced an efficient cross-modal retrieval algorithm, termed Flexible Collaborative Matrix Factorization Hash (FS-CMFH). This method leverages label consistency across diverse modalities to maintain the semantic information within and between the modalities within a shared latent semantic representation space. Mandal et al. [68] produced latent factors corresponding to various modalities and employed a matrix factorization-based linear transformation technique to project multimodal data into a more discriminative label space. Li et al. [69] introduced a scalable discrete matrix factorization hashing approach. This method incorporates the collaborative matrix factorization of kernel features and labeled semantic embeddings to acquire latent semantic spaces, ensuring the preservation of semantic similarity across and within modalities. Wang et al. [70] presented a hash method based on label consistency matrix factorization. This approach enables the transformation of heterogeneous data into a latent semantic representation space, wherein multimodal data of the same category converge on a unified feature representation. Liu et al. [71] utilized matrix factorization to encode heterogeneous data with varying hash lengths. They introduced a matrix factorization hashing algorithm to extend cross-modal hash retrieval to varied and demanding scenarios. The method based on shallow models has high computational efficiency, but it mainly relies on manually extracted features, and the nonlinear representation ability of the model is limited. Therefore, modeling the complex semantic correlation of data is not outstanding. In recent years, with the booming development of deep learning technology, more research has begun to use deep models for end-to-end hash learning. Jiang et al. [72] proposed a cross-modal hashing model based on deep neural networks by merging multimodal feature learning and hash learning into a unified framework. Yang et al. [73] fused diverse types of paired constraints using deep models to augment the measurement of hash code similarity from both intra-modal and inter-modal perspectives. Chen et al. [74] introduced a tri-stage dual deep neural network cross-modal hashing approach, employing two deep networks to generate cross-modal hash codes. Deng et al. [75] developed a deep hash network based on triplets, utilizing triplet labels as supervised information to establish semantic connections between instances across modalities. Ma et al. [76] studied a cross-modal hashing method based on global and local semantic preservation, which maintains significant differences between different hash codes, thereby improving the discriminability of hash codes. Shi et al. [77] studied the relationship between semantic structure and discriminative behavior and proposed a discriminative hashing algorithm with equal guidance to construct a universal semantic

TABLE 5. Supervised hash learning approaches based on the MAP score on the Wikipedia dataset.

Methods	Year	Journals or conferences	Image Q Text				Text Q Image			
			16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
SMFH [66]	2016	IEEE Trans	0.257	0.276	0.286	0.291	0.578	0.604	0.617	0.622
DCMH [72]	2017	CVPR	-	-	-	-	-	-	-	-
PRDH [73]	2017	AAAI	-	-	-	-	-	-	-	-
LCMFH [68]	2017	ICIP	0.338	0.366	0.373	0.378	0.729	0.744	0.753	0.755
FS-CMFH [69]	2018	Multimed Tools Appl	0.301	0.309	0.305	0.310	0.662	0.675	0.691	0.693
DDCMH [74]	2018	AAAI	-	-	-	-	-	-	-	-
TDH [75]	2018	IEEE Trans	-	-	-	-	-	-	-	-
GLSPH [76]	2018	Neurocomputing	-	-	-	-	-	-	-	-
EGDH [77]	2019	IJCAI	-	-	-	-	-	-	-	-
DBRH [78]	2019	IEEE Trans	0.253	0.265	0.269	0.288	0.544	0.538	0.548	0.549
LCMFH [70]	2019	IEEE Trans	0.410	0.414	0.426	0.439	0.637	0.639	0.657	0.671
SCRATCH [69]	2020	IEEE Trans	-	-	-	-	-	-	-	-
MTFH [71]	2021	IEEE Trans	0.341	0.353	0.351	0.335	0.702	0.713	0.734	0.737
IRGRH [82]	2022	Knowledge-Based Systems	-	-	-	-	-	-	-	-
DAZSH [83]	2022	Neurocomputing	0.403	0.420	-	-	0.476	0.468	-	-
ROHLSE [84]	2023	Pattern Recognition	-	-	-	-	-	-	-	-
TA-DCH [85]	2023	IEEE Trans	0.434	0.418	0.422	0.463	0.591	0.627	0.555	0.605

structure preserving classification. Hu et al. [78] analyzed the effectiveness of multimodal networks in maintaining inter modal and intra modal consistency, and proposed a multimodal deep binary reconstruction model. Xia et al. [79] proposed an efficient multimodal learning module that incorporates a fusion converter encoder supervised by contrastive loss. This module enhances the interaction between modalities and concurrently refines the semantic representations of the individual modalities. In adition, a dedicated contrastive hash learning module was designed to produce high-quality hash codes associated with each modality. Xie et al. [80] introduced an adversarial hashing method that focused on multi-task consistency preservation. This approach thoroughly investigated multimodal semantic consistency and correlation, thereby accomplishing efficient cross-modal retrieval. Nie et al. introduced a deep multi-scale fusion hashing method [81], which employs a multi-scale fusion model to explore cross-modal semantic correlations through multi-scale semantic fusion. Hou et al. [82] introduced a reasoning approach centered on multiple instance relation graphs in their research. Through the construction of similarity matrices, both global and local instance relation graphs were formulated to leverage fine-grained relations among instances comprehensively. Additionally, to address the characteristics inherent to both the image and text modalities, a progressive training strategy was employed to train the proposed neural network model. Shu et al. introduced the DAZSH method [83], that incorporates data features and class attributes to generate a semantic category representation for each category. This method adeptly captures the relationships between seen and unseen classes through the acquisition of a category representation vector for each instance. Consequently, it enables the transfer of supervised knowledge from seen classes to unseen ones. Li et al. [84] introduced the ROHLSE method, which utilizes sample representations in the feature space to predict labels based on the dependencies between the sample instances and labels. Wang et al. devised a semantic adaptation network [85] to produce target prototype code and reconstruct category labels. This facilitates semantic interactions between the target semantics and implicit semantics within the targeted model. Moreover, the authors incorporated a discriminator into adversarial training, with the goal of enhancing the visual realism and category discrimination of adversarial examples, thereby improving targeted attack performance. Tables 5, 6, and 7 respectively show supervised hash learning approaches based on the MAP score on the Wikipedia, NUS WIDE, and MIRFlickr datasets. These data are all from the source paper, and the "-" in the table indicates that the method was not tested on this dataset or at this hash code length. The supervised hash method based on deep learning models not only has a strong nonlinear semantic representation ability, but also achieves end-to-end semantic feature learning and hash code generation, significantly improving the accuracy of cross-modal hash retrieval. First, fine-grained semantic mining was insufficient. Although existing fine-grained semantic mining fully considers objects and their relationships, the degree to

TABLE 6. Supervised hash learning approaches based on the MAP score on the NUS-WIDE dataset.

Methods	Year	Journals or conferences	Image Q Text				Text Q Image			
			16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
SMFH [66]	2016	IEEE Trans	0.455	0.462	0.466	0.467	0.503	0.506	0.507	0.508
LCMFH [68]	2017	ICIP	0.718	0.743	0.765	0.782	0.842	0.853	0.845	0.855
DCMH [72]	2017	CVPR	0.590	0.603	0.609	-	0.639	0.651	0.657	-
PRDH [73]	2017	AAAI	0.635	0.653	0.651	-	0.681	0.696	0.694	-
DDCMH [74]	2018	AAAI	0.802	0.827	0.832	-	0.687	0.701	0.741	-
TDH [75]	2018	IEEE Trans	0.639	0.663	0.675	-	0.665	0.676	0.685	-
GLSPH [76]	2018	Neurocomputing	-	-	-	-	-	-	-	-
EGDH [77]	2019	IJCAI	0.319	0.319	0.319	0.319	0.319	0.319	0.319	0.319
DBRH [78]	2019	IEEE Trans	0.394	0.409	0.417	0.417	0.425	0.429	0.438	0.443
FS-CMFH [67]	2018	Multimed Tools Appl	0.601	0.605	0.606	0.614	0.609	0.605	0.619	0.611
LCMFH [70]	2019	IEEE Trans	0.922	0.939	0.948	0.954	0.868	0.878	0.893	0.900
SCRATCH [69]	2020	IEEE Trans	0.643	0.649	0.670	0.673	0.788	0.807	0.827	0.832
MTCAH [80]	2020	IEEE Trans	-	-	-	-	-	-	-	-
DMFH [81]	2021	IEEE Trans	0.630	0.647	0.680	-	0.607	0.621	0.640	-
MTFH [71]	2021	IEEE Trans	0.655	0.659	0.676	0.675	0.757	0.780	0.795	0.804
IRGRH [82]	2022	Knowledge-Based Systems	0.756	0.793	0.816	0.839	0.750	0.783	0.804	0.817
DAZSH [83]	2022	Neurocomputing	-	-	-	-	-	-	-	-
UCMFH [79]	2023	Information Fusion	0.849	882	0.894	0.897	0.859	0.886	0.901	0.900
ROHLSE [84]	2023	Pattern Recognition	0.618	0.619	0.632	0.633	0.716	0.724	0.735	0.742
TA-DCH [85]	2023	IEEE Trans	0.751	0.786	0.805	0.836	0.780	0.796	0.812	0.804

which different combinations of objects and relationships are mined is not sufficiently deep. Second, existing research has not considered the organic combination of global semantic information and local semantic information; that is, the fusion of coarse-grained semantic information and fine-grained semantics in multi-source heterogeneous data.

B. UNSUPERVISED

Unsupervised cross-modal hashing methods do not rely on labeled data and focus on learning hash functions solely from data distribution. These methods typically leverage the correlation between modalities to learn binary codes that capture the underlying cross-modal relationships.

The good learning performance of supervised learning methods relies heavily on a large amount of labeled data. However, data labeling relies mainly on manual annotation. A large amount of manual annotation requires considerable manpower, material, and financial resources, which are expensive and unrealistic. Therefore, the vast majority of the existing heterogeneous data from multiple sources have not been manually annotated. to fully utilize the massive amount of unlabeled data, many scholars have focused on unsupervised cross-modal hash learning. Unsupervised cross-modal hash learning does not rely on the semantic labels of data, and its purpose is to deeply explore the potential association relationships between multi-modal and multi-view data. Existing research has mainly focused on mining the association relationships between paired views, with few studies considering the association relationships between all modalities and all views simultaneously.

Hu et al. [86] proposed the Iterative Multi-View Hashing algorithm (IMVH), which learns optimal alignment in encoding schemes to maintain similarity between views. Irie et al. [87] proposed an unsupervised hash method called alternating common quantization, which alternately searches for binary quantizers for each modal space by connecting with multimodal data, ensuring minimal quantization errors while maintaining data similarity. Wu et al. [88] constructed the underlying relationships between domains of the same object by maximizing the correlation between cross domain hash codes, and the multimodal objective function was transformed into a single modal formalization. Ding et al. [89] introduced a collaborative matrix factorization hash algorithm that seeks to acquire unified hash code for a multimodal instance within a common latent semantic space. Rafailidis and Crestani [90] introduced a method for cross-modal hash learning that utilizes a clustering-based joint matrix factorization strategy. This approach enables the calculation of inter-modal similarity, intra-modal similarity, and clustering-based similarities in a unified representation space. Fang et al. [91] introduced a method for multimodal graph-regularized smooth matrix factorization hashing. This method ensures the sparsity of the learned dictionary and common features, thereby effectively

TABLE 7. Supervised hash learning approaches based on the MAP score on the MIRFlickr dataset.

Methods	Year	Journals or conferences	Image Q Text				Text Q Image			
			16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
SMFH [66]	2016	IEEE Trans	-	-	-	-	-	-	-	-
LCMFH [68]	2017	ICIP	0.670	0.689	0.857	0.869	0.830	0.861	0.863	0.896
DCMH [72]	2017	CVPR	0.741	0.747	0.749	-	0.783	0.790	0.793	-
PRDH [73]	2017	AAAI	0.713	0.713	0.720	-	0.747	0.754	0.751	-
FS-CMFH [67]	2018	Multimed Tools Appl	0.626	0.633	0.640	0.645	0.617	0.623	0.623	0.626
DDCMH [74]	2018	AAAI	0.821	0.843	0.855	-	0.773	0.777	0.791	-
TDH [75]	2018	IEEE Trans	0.711	0.723	0.723	-	0.742	0.750	0.755	-
GLSPH [76]	2018	Neurocomputing	0.746	0.771	0.793	0.799	0.754	0.777	0.803	0.805
EGDH [77]	2019	IJCAI	0.757	0.773	0.796	0.790	0.779	0.794	0.799	0.801
DBRH [78]	2019	IEEE Trans	0.587	0.590	0.590	0.591	0.588	0.596	0.596	0.598
LCMFH [70]	2019	IEEE Trans	0.787	0.816	0.825	0.838	0.905	0.921	0.931	0.934
6SCRATCH [69]	2020	IEEE Trans	0.723	0.741	0.766	0.776	0.798	0.818	0.842	0.851
MTCAH [80]	2020	IEEE Trans	0.789	0.796	0.795	-	0.778	0.786	0.785	-
DMFH [81]	2021	IEEE Trans	0.798	0.810	0.810	-	0.780	0.792	0.795	-
MTFH [71]	2021	IEEE Trans	0.747	0.761	0.765	0.768	0.804	0.815	0.817	0.835
IRGRH [82]	2022	Knowledge-Based Systems	0.831	0.855	0.877	0.894	0.825	0.877	0.882	0.892
DAZSH [83]	2022	Neurocomputing	-	-	-	-	-	-	-	-
UCMFH [79]	2023	Information Fusion	0.923	0.953	0.964	0.965	0.920	0.947	0.959	0.961
ROHLSE [84]	2023	Pattern Recognition	0.723	0.731	0.734	0.737	0.789	0.804	0.808	0.812
TA-DCH [85]	2023	IEEE Trans	0.820	0.852	0.848	0.858	0.775	0.795	0.819	0.823

minimizing quantization loss. Cheng et al. [92] proposed a robust unsupervised cross-modal hashing method that preserves the original feature information as much as possible by exploring the partial or incomplete correspondence between modalities. In recent years, the performance of unsupervised cross-modal hashing, particularly that based on deep learning, has been substantially enhanced owing to the potent representational capabilities of deep neural networks. Liong et al. [93] crafted a deep fusion neural network to acquire the nonlinear transformations of image-text pairs. The model generates a unified modal hash codes through the application of classification hinge loss criteria. Wu et al. [94] investigated the unsupervised deep cross-modal hashing method, which is innovative in that it combines deep learning with matrix decomposition, and cleverly utilizes the binary latent factor model to facilitate cross-modal hash learning. Su et al. [95] introduced an unsupervised deep joint semantic reconstruction hash (DJSRH) technique. This method integrates original neighborhood information from various modalities through a semantic similarity matrix, capturing intrinsic semantic similarities. Huang et al. [96] presented an unsupervised cross-modal hashing framework that employs data fusion to capture the underlying manifolds across modalities. This approach mitigates issues related to maximizing the intra-modal and inter-modal similarity constraints. Yang et al. [97] investigated a deep semantic alignment hash algorithm (DSAH), that intelligently aligns the similarity between features and hash codes by incorporating a semantic alignment loss function. Liu et al. [98] introduced a similarity hashing method (JDSH) rooted in joint modal distribution. The proposed approach employs an unsupervised learning algorithm, specifically based on Distributed Similarity Decision and Weighting (DSDW), to generate hash codes with enhanced discriminative properties. Hoang et al. [99] presented an algorithm grounded in spectral embedding, aiming to learn single-modal and binary cross-modal representations concurrently. This method effectively preserves the local structure of each modality and captures the hidden patterns across all modalities. Wang et al. [43] designed an unsupervised method called Semantic-based Cross-modal Hashing (SRCH). The method accomplishes the computation of cross-modal similarity by modeling the set recombination to update the training data. Wang et al. [100] introduced an unsupervised deep cross-pattern hashing algorithm, referred to as UDCH-VLR, which leverages virtual label regression. This method focuses on learning a unified hash code through collaborative matrix decomposition, ensuring the preservation of shared semantics across multiple modalities. In addition, in order to effectively reduce the heterogeneity between modalities, some studies have integrated deep adversarial learning methods into cross modal hash learning models to improve the quality of hash code generation. For instance, He et al. introduced an unsupervised cross-modal retrieval approach based on

TABLE 8. UnSupervised hash learning approaches based on the MAP score on the Wikipedia dataset.

Methods	Year	Journals or conferences	Image Q Text				Text Q Image			
			16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
QCH [88]	2015	IJCAI	0.234	0.248	-	-	0.303	0.317	-	-
SCMFH [89]	2016	TransIP	0.259	0.264	0.280	0.288	0.632	0.640	0.657	0.664
CMDVH [93]	2017	ICCV	0.424	0.443	0.452	0.444	0.727	0.733	0.738	0.737
UDCMH [94]	2018	IJCAI	0.309	0.318	0.329	0.346	0.622	0.633	0.645	0.658
UGACH [102]	2018	AAAI	-	-	-	-	-	-	-	-
DJSRH [95]	2019	ICCV	0.388	0.403	0.412	0.421	0.611	0.635	0.646	0.658
MSFH [91]	2019	Knowledge-Based Systems	0.259	0.267	0.273	0.273	0.624	0.642	0.658	0.655
DCSH [99]	2020	TransIP	-	-	-	-	-	-	-	-
SRCH [43]	2020	IJCAI	0.374	0.380	0.391	-	0.377	0.401	0.406	-
UDCH-VLR [100]	2020	Neurocomputing	-	-	-	-	-	-	-	-
AGSH [105]	2021	ICME	0.397	0.434	0.446	-	0.431	0.443	0.453	-
CDAH [106]	2021	Neurocomputing	0.395	0.406	0.415	-	0.433	0.453	0.478	-
DAEH [108]	2022	IEEE Trans	-	-	-	-	-	-	-	-
CMIMH [109]	2023	IEEE Trans	-	-	-	-	-	-	-	-
UCCH [110]	2023	IEEE Trans	-	-	-	-	-	-	-	-
CIRH [111]	2023	IEEE Trans	-	-	-	-	-	-	-	-

adversarial learning [101]. Integrating a modal classifier that predicts the modality of the transformed features ensures their statistical indistinguishability. Zhang et al. [102] maximized the unsupervised representation learning capability inherent in Generative Adversarial Networks (GANs) to extract the manifold structure within cross-modal data. They formulated an unsupervised cross-modal hash model based on GANs. Li et al. investigated an unsupervised coupled cyclic generative adversarial hash network (UCH) [103]. This approach employs an outer recurrent network to learn shared representations and an inner recurrent network to generate reliable hash codes. Zhang et al. introduced a multipath generative adversarial network model for unsupervised cross-modal hash modeling [104]. Shen et al. [105] presented a novel network designed to pass extracted features through an attention module, effectively encoding rich and relevant features. Simultaneously, the network is capable of generating hash codes through self-supervision, facilitated by the proposed attention-aware semantic fusion matrix. Shen et al. presented Clustering-driven Deep Adversarial Hashing (CDAH) [106], that integrates soft clustering into approximates and iteratively optimizes the clustering centers for each modality. This approach aims to enhance the accuracy of capturing the semantic similarities within each modality. Mikriukov et al. [107] presented an innovative unsupervised cross-modal hashing approach, referred to as CHNR, specifically designed to handle noisy image-text correspondences. The proposed method employs a multi-term noise-robust contrastive loss function for unsupervised learning of cross-modal hash codes. Shi et al. proposed a method called DAEH [108], which combines information mixture similarity estimation with integrated distance analysis to model image text similarity discrimination. The utilization of IMSE is aimed at mitigating redundant information within similarity guidance during the optimization of hash functions. Hoang et al. [109] utilize a strategy that maximizes mutual information (MI) to tackle the challenge of unsupervised learning for binary hash codes, with a specific emphasis on improving efficiency in cross-modal retrieval. Hu et al. [110] introduced a novel momentum optimizer designed for learnable hashing operations in the context of contrastive learning. This innovative approach facilitates on-the-shelf deep cross-modal hashing. Zhu et al. [111] presented a novel unsupervised cross-modal retrieval method termed correlation-identity reconstruction hashing. The proposed approach incorporates a reconstruction strategy that allows for simultaneous preservation of multi-modal correlation and identity semantics within binary hash codes. Tables 8, 9, and 10 respectively show unsupervised hash learning approaches based on the MAP score on the Wikipedia, NUS WIDE, and MIRFlickr datasets. Currently, unsupervised methods have two major limitations. Unsupervised cross-modal hashing methods learn potential correlations between multimodal data through multi-view methods or deep models, the complex correlations between multimodal data have not been fully explored. Although there have been studies attempting to mine the correlations between multiple modalities of data, most of them are focused on learning for two modalities or are based on pairwise correlations, making it difficult to simultaneously consider the correlations between all modalities.In addition, correlation analysis methods based on tensor decomposition cannot effectively characterize complex nonlinear semantic correlations. However, owing to the lack of semantic labels

TABLE 9. UnSupervised hash learning approaches based on the MAP score on the NUS-WIDE dataset.

Methods	Year	Journals or conferences	Image Q Text				Text Q Image			
			16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
QCH [88]	2015	IJCAI	0.540	0.549	-	-	0.557	0.574	-	-
SCMFH [89]	2016	TransIP	0.552	0.583	0.597	0.610	0.645	0.698	0.727	0.734
CMDVH [93]	2017	ICCV	0.850	0.876	0.880	0.891	0.827	0.833	0.840	0.878
UGACH [102]	2018	AAAI	0.613	0.623	0.628	0.631	0.603	0.614	0.640	0.641
UDCMH [94]	2018	IJCAI	0.511	0.519	0.524	0.588	0.637	0.653	0.695	0.716
DJSRH [95]	2019	ICCV	0.724	0.773	0.798	0.817	0.712	0.744	0.771	0.789
MSFH [91]	2019	Knowledge-Based Systems	0.468	0.460	0.453	0.434	0.471	0.465	0.458	0.451
DCSH [99]	2020	TransIP	0.604	0.620	-	-	0.607	0.621	-	-
SRCH [43]	2020	IJCAI	0.544	0.557	0.567	-	0.553	0.567	0.575	-
UDCH-VLR [100]	2020	Neurocomputing	0.557	0.598	0.632	0.645	0.605	0.656	0.674	0.735
AGSH [105]	2021	ICME	0.543	0.552	0.562	-	0.543	0.567	0.570	-
CDAH [106]	2021	Neurocomputing	0.558	0.566	0.574	-	0.559	0.576	0.581	-
DAEH [108]	2022	IEEE Trans	0.731	0.754	0.773	0.779	0.713	0.734	0.749	0.751
CMIMH [109]	2023	IEEE Trans	0.739	0.764	-	-	0.728	0.750	-	-
UCCH [110]	2023	IEEE Trans	0.698	0.708	0.737	0.742	0.701	0.724	0.745	0.750
CIRH [111]	2023	IEEE Trans	0.815	0.836	0.854	0.862	0.774	0.803	0.810	0.817

in the data, unsupervised hash models find it difficult to effectively mine key semantic features and achieve cross-modal semantic alignment.

C. SEMI-SUPERVISED

To address the challenge of accurately capturing cross-modal semantic correlations in unsupervised cross-modal hashing methods, some studies have introduced semi-supervised learning methods into cross-modal hashing tasks, supplementing the training set with rich unlabeled data to improve supervised learning performance. Wang et al. [112] proposed a semi-supervised semantic factor decomposition hash algorithm (S3FH), which optimizes a joint framework consisting of three interactive parts: semantic decomposition, multi-graph learning, and multimodal correlation learning. Wang et al. [113] proposed a semi-supervised deep quantization (SSDQ) model, which is innovative in that it integrates supervised information (pairwise similarity + class labeling) and unsupervised information in a unified framework. Employing composite quantization, this model facilitates precise and efficient cross-domain queries. Zhang et al. [114] designed a semi-supervised cross-modal hash algorithm (SS-LPDP) based on label prediction and distance preservation. This method employs a shared objective function with distance preservation constraints, which effectively classifies data and minimizes interference during retrieval. Wang et al. [115] amalgamated labeled and unlabeled data, undertaking the learning of latent subspaces for both types by integrating cross-modal relaxation latent subspace learning and semantic preservation regularization. This approach is based on adversarial learning techniques. Zhang et al. [116] introduced a model, namely the cross-modal hash generation adversarial network, and applied a reinforcement learning-based algorithm to facilitate the training of the model. Mandal et al. [117] learned complementary information from different modalities to predict the class labels of unlabeled data. Their proposed method can be used as a baseline method to effectively improve the performance, even in situations with limited labeled data. Shen et al. [118] proposed a semi-supervised graph convolutional hash network (SGCH) method that learns a common cross-modal Hamming space through end-to-end neural networks. Owing to the lack of use of semantic labels in unsupervised learning, it is difficult for unsupervised learning to capture rich semantic discriminative information and effectively narrow the semantic gap. To fully utilize massive amounts of unlabeled data and improve the mining ability of unsupervised learning for semantic discriminative information, some scholars have turned their attention to domain adaptive learning based on transfer learning. Transfer learning, an innovative machine learning paradigm, leverages pre-existing knowledge to address challenges in diverse yet related domains. This method relaxes the two key assumptions found in traditional machine learning, aiming to apply existing knowledge to address learning tasks in the target domain, even in situations with scarce labeled data. Domain adaptive learning methods facilitate the transition from the source domain to the target domain via four consistency constraints: structural, domain, semantic, and modal. For example, Huang et al. introduced the Modal Confrontation Hybrid Transfer Network (MHTN) [119], which is specifically designed for transferring knowledge from a single-modal source domain to a cross-modal target domain. This model enabled the acquisition of a unified cross-modal representation. In anther study [120], a two-stage progressive cross modal knowledge transfer (TPCKT) method was

TABLE 10. UnSupervised hash learning approaches based on the MAP score on the MIRFlickr dataset.

Methods	Year	Journals or conferences	Image Q Text				Text Q Image			
			16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
QCH [88]	2015	IJCAI	-	-	-	-	-	-	-	-
SCMFH [89]	2016	TransIP	0.622	0.630	0.637	0.643	0.652	0.664	0.672	0.679
CMDVH [93]	2017	ICCV	-	-	-	-	-	-	-	-
UGACH [102]	2018	AAAI	0.685	0.693	0.704	0.702	0.673	0.676	0.686	0.690
UDCMH [94]	2018	IJCAI	0.689	0.698	0.714	0.717	0.692	0.704	0.718	0.733
DJSRH [95]	2019	ICCV	0.810	0.843	0.862	0.876	0.786	0.822	0.835	0.847
MSFH [91]	2019	Knowledge-Based Systems	0.594	0.588	0.583	0.580	0.615	0.610	0.604	0.597
DCSH [99]	2020	TransIP	0.732	0.743	-	-	0.736	0.741	-	-
SRCH [43]	2020	IJCAI	0.681	0.692	0.700	-	0.697	0.708	0.715	-
UDCH-VLR [100]	2020	Neurocomputing	0.740	0.741	0.746	0.749	0.758	0.761	0.762	0.770
AGSH [105]	2021	ICME	0.679	0.691	0.698	-	0.674	0.689	0.693	-
CDAH [106]	2021	Neurocomputing	0.698	0.704	0.713	-	0.709	0.714	0.727	-
DAEH [108]	2022	IEEE Trans	0.783	0.794	0.800	0.805	0.761	0.768	0.774	0.781
CMIMH [109]	2023	IEEE Trans	0.807	0.819	-	-	0.798	0.814	-	-
UCCH [110]	2023	IEEE Trans	0.739	0.744	0.754	0.760	0.725	0.725	0.743	0.747
CIRH [111]	2023	IEEE Trans	0.901	0.913	0.929	0.937	0.867	0.885	0.900	0.901

TABLE 11. Semi-Supervised hash learning approaches based on the MAP score on the Wikipedia dataset.

Methods	Year	Journals or conferences	Image Q Text				Text Q Image			
			16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
S3FH [112]	2017	Multimed Tools Appl	0.331	0.362	0.367	0.366	0.388	0.408	0.435	0.424
SSDQH [113]	2019	ACM Conf. on Multimedia	-	-	-	-	-	-	-	-
S3PH [115]	2019	ICME	-	-	-	-	-	-	-	-
SCH-GAN [116]	2020	IEEE Trans	0.527	0.528	0.550	0.542	0.858	0.877	0.885	0.890
MGCH [124]	2022	Information Sciences	0.368	0.454	0.491	-	0.489	0.573	0.598	-
MCGCNH [125]	2022	Pattern Recognition	0.544	0.638	0.654	-	0.553	0.641	0.653	-
SSCH [126]	2023	IEEE Trans	0.230	0.240	0.245	0.249	0.228	0.237	0.242	0.246
TS3H [127]	2023	IEEE Trans	-	-	-	-	-	-	-	-

proposed, with parameter shared domain discriminators set up at the modality specific layer and modality shared layer respectively to maintain cross modal consistency of transmission. Wen et al. introduced the Cross-Modal Similarity Transfer (CMST) method [121] designed for unsupervised learning and preservation of semantic relationships between unpaired items. Zhen et al. [122] proposed a method called Deep Multimodal Transfer Learning (DMTL), the highlight of which is the transfer of semantic category information from the source domain to the target domain. Peng et al. presented a domain adaptation technique called scene graph-based domain adaptation (DASG) [123]. This method utilizes the Visual Genome as the source domain for knowledge transfer, with the goal of improving the cross-modal retrieval performance in the target domain. Shen et al. [124] made a significant contribution by introducing an approach that utilizes multi-view graphs to establish connections between labeled and unlabeled data from various perspectives. This innovative method effectively filters and highlights the significant features in the context of this study. Wu et al. proposed a new semi-supervised cross-modal hashing approach, called MCGCN [125]. This method optimizes the utilization of label and structural information from both labeled and unlabeled samples to enhance the propagation of semantic information and the acquisition of discriminative hash codes. Zhang et al. [126] introduced a flexible modality-specific approach that efficiently addressed label completion for unsupervised data, regardless of alignment. The proposed method uses label regression to expedite and enhance the precision of hash code learning for diverse data. In contrast to conventional semi-supervised approaches that jointly learn pseudolabels, hash codes, and hash functions, Fan et al. [127] proposed a novel approach that decomposes into three distinct stages. As implied by its name, each stage is

TABLE 12. Semi-Supervised hash learning approaches based on the MAP score on the NUS-WIDE dataset.

Methods	Year	Journals or conferences	Image Q Text				Text Q Image			
			16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
S3FH [112]	2017	Multimed Tools Appl	0.360	0.385	0.412	0.420	0.345	0.327	0.378	0.391
SSDQH [113]	2019	ACM Conf. on Multimedia	0.709	0.765	-	-	0.763	0.832	-	-
S3PH [115]	2019	ICME	0.718	0.741	0.729	0.726	0.837	0.846	0.840	0.856
SCH-GAN [116]	2020	IEEE Trans	0.713	0.724	0.732	0.749	0.738	0.742	0.769	0.782
MGCH [124]	2022	Information Sciences	0.664	0.669	0.684	-	0.679	0.693	0.690	-
MCGCNH [125]	2022	Pattern Recognition	0.753	0.781	0.782	-	0.777	0.780	0.781	-
SSCH [126]	2023	IEEE Trans	0.506	0.507	0.515	0.520	0.483	0.490	0.492	0.500
TS3H [127]	2023	IEEE Trans	0.472	0.498	0.512	0.521	0.559	0.598	0.622	0.637

TABLE 13. Semi-Supervised hash learning approaches based on the MAP score on the MIRFlickr dataset.

Methods	Year	Journals or conferences	Image Q Text				Text Q Image			
			16bits	32bits	64bits	128bits	16bits	32bits	64bits	128bits
S3FH [112]	2017	Multimed Tools Appl	-	-	-	-	-	-	-	-
SSDQH [113]	2019	ACM Conf. on Multimedia	0.839	0.855	-	-	0.835	0.872	-	-
S3PH [115]	2019	ICME	0.779	0.798	0.809	0.815	0.892	0.908	0.912	0.913
SCH-GAN [116]	2020	IEEE Trans	0.738	0.745	0.757	0.768	0.771	0.790	0.793	0.804
MGCH [124]	2022	Information Sciences	0.735	0.738	0.751	-	0.738	0.749	0.757	-
MCGCNH [125]	2022	Pattern Recognition	-	-	-	-	-	-	-	-
SSCH [126]	2023	IEEE Trans	0.632	0.643	0.644	0.648	0.620	0.631	0.634	0.640
TS3H [127]	2023	IEEE Trans	0.731	0.736	0.737	0.741	0.793	0.800	0.803	0.804

conducted individually, contributing to a more cost-effective and precise optimization process. Tables 11, 12, and 13 respectively show semi-supervised hash learning approaches based on the MAP score on the Wikipedia, NUS WIDE, and MIRFlickr datasets. However, existing knowledge transfer or domain adaptation methods mainly focus on cross-modal real value representation retrieval, and there is little research on large-scale cross-modal hashing. On the other hand, existing work usually only selects one or two consistency constraints for research, and there is no research that comprehensively considers all consistency constraints in the learning process of hash functions. Meanwhile, in terms of research on structural consistency loss, existing studies have only used a single spatial metric to describe the structural relationships between data in different fields, without fully exploring the deeper structural relationships between data in different fields brought about by the combination of multiple spatial metrics. Therefore, combining cross-domain knowledge transfer methods with efficient cross-modal hash code generation methods, transfering rich semantic information from different fields to the target domain, and improving the semantic discrimination ability of cross-modal hashing will become the key to breaking through the performance bottleneck of unsupervised cross-modal hashing. In the era of big data, cross-modal hashing has gained widespread application in cross-modal retrieval owing to its high efficiency, low dimensionality, low storage overhead, and effectiveness in representing the consistency of high-level features within multi-source heterogeneous data. However, current cross-modal hashing approaches have not fully harnessed the rich semantic information embedded in extensive multi-source heterogeneous datasets. Consequently, a significant challenge in cross-modal hash retrieval research is the effective mining of diverse information from multi-source heterogeneous data to guide the training of cross-modal deep hash models and enhance the quality of the generated hash codes. This project aims to start with the deep mining of different granularity data, different modal data, and the inherent correlation relationships between different data sources. It aims to study cross-modal hashing for multi-granularity semantic fusion representation, multi-view correlation cross-modal hashing for multi-modal data, and domain adaptive learning cross-modal hashing for knowledge transfer, to enhance the representation ability of cross-modal hashing for multi-source heterogeneous data. Therefore, the research on cross-modal hash retrieval for heterogeneous data from multiple sources has innovative and significant practical value.

Investigating cross-modal hash retrieval for heterogeneous data from multiple sources is of significant theoretical and

practical importance. Cross-modal hashing is extensively utilized in retrieval due to its efficiency, low dimensionality, and minimal storage overheads. Although existing research has made progress, numerous challenges remain unresolved and require urgent attention. In a big data environment, the correlation between different granularity data, different modality data, and different data source data in multi-source heterogeneous data has not been fully explored. This may lead to a decrease in the representation ability of data through cross-modal hash learning hash codes and reduce the efficiency of cross-modal hash retrieval. Despite their advantages, cross-modal retrieval hashing methods have several challenges. One challenge is the semantic gap between modalities, because binary codes may not fully capture the semantic relationships. Another challenge is the scalability of the hashing methods for handling large-scale cross-modal databases efficiently. Additionally, the design of effective hash functions and selection of appropriate training strategies remain active research areas.

V. DATASET AND EVALUATION METRICS

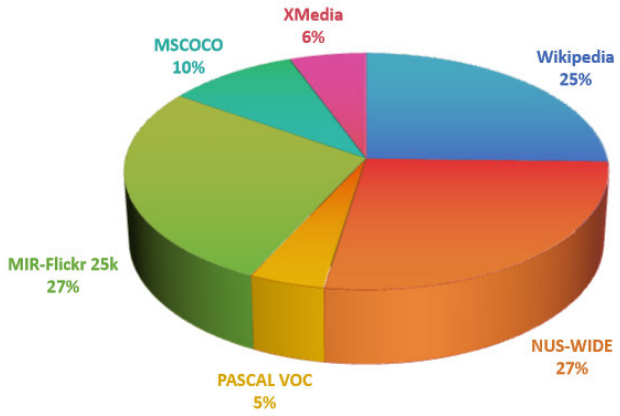

FIGURE 6. The percentage of usage frequency for each dataset.

A. DATASET

For the cross-modal information processing of images and text, datasets are generally required for evaluation. High quality datasets can enable neural networks to fully learn various potential knowledge while avoiding over fitting and other problems. Currently, there are several commonly used datasets for cross-modal image and text retrieval. Figure 7 shows some examples of the Wikipedia, NUS-WIDE, and MIR-Flickr datasets. We have summarized all the references in this paper, conducted statistical analysis on the datasets used in their paper, and obtained a pie chart, as shown in Figure 6, which shows the percentage of usage frequency for each dataset, and Table 14 shows the differences between different datasets.

1) WIKIPEDIA

The Wikipedia [15] dataset was collected from Wikipedia and is the most commonly used dataset for cross-modal retrieval research. It consists of 2866 document corpora with relevant image text pairs that describe images in short paragraphs. In addition, each pair was assigned a label from one of the ten semantic classes. The text is presented with 10 dimensions derived from a recent Dirichlet allocation model [128], and the images are presented by 128 dimensional SIFT descriptor histograms [129]. Because to the small size of this dataset and the abstract categories of image text pairs, misunderstandings can easily occurt, resulting in many methods not performing well on this dataset. http://www.svcl.ucsd.edu/projects/crossmodal/ Readers can directly download 2866 multimedia documents (image+text) and features (Matlab format) from this link for experimental research.

2) NUS-WIDE

NUS-WIDE [130] is a network image dataset created by the Multimedia Retrieval Laboratory of the National University of Singapore in 2009, and its images were mainly sourced from the Flickr website. This dataset includes 269 648 images, that contain only two modalities: image and text. It is one of the most commonly used datasets for cross-modal retrieval. The dataset, encoding low level features, ground truth, tags, concept lists, image lists, and original URLs were acquired through the links https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDEHTML.

3) PASCAL VOC

PASCAL VOC [131] This dataset contains 5 011 and 4 952 training and testing image phrase label data pairs respectively, totaling 20 semantic annotations. Each image has text semantic annotations, and the annotation file is an object class label for one of the 20 classes present in the image. The PASCAL VOC 2007 dataset was divided into two parts: training and validation set training, and testing set test, each accounting for approximately 50 precent of the total data. The dataset can be obtained directly from the following link. http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html.

4) MIR FLICKR 25K

The MIR Flickr 25k [132] contains approximately 25000 images from the Flickr platform. Each image had five annotations, which where natural language descriptions provided by human annotators to describe the main content of the image. This dataset is commonly used to train and evaluate image annotation models, helping the model understand images and generate relevant semantic descriptions. The Flickr 8k is a scaled down version. Flickr 30k is an expanded version of it, and these datasets provide images and associated manual text descriptions for image annotation tasks. The primary distinction lies in the dataset size, which encompasses both the quantity of images and the number of annotations associated with each image. Readers can obtain dataset information from the following link https://press.liacs.nl/mirflickr/.

TABLE 14. Differences between different datasets.

Dataset	Model	Sample size	categories	Source
Wikipedia	Image/Text	2866	10	Wikipedia website
NUS-WIDE	Images/Tags	269648	81	National University of Singapore
PASCAL VOC	Image/Phrase	9963	20	The PASCAL Visual Object Classes Challenge 2007
MIR-Flickr 25k	Images/Tags	250000	5	Flickr platform
MSCOCO	Images/Tags	120000	80	Microsoft
XMedia	Text/Image/Video/Audio/3D models	more than 1200,0	20	Peking University

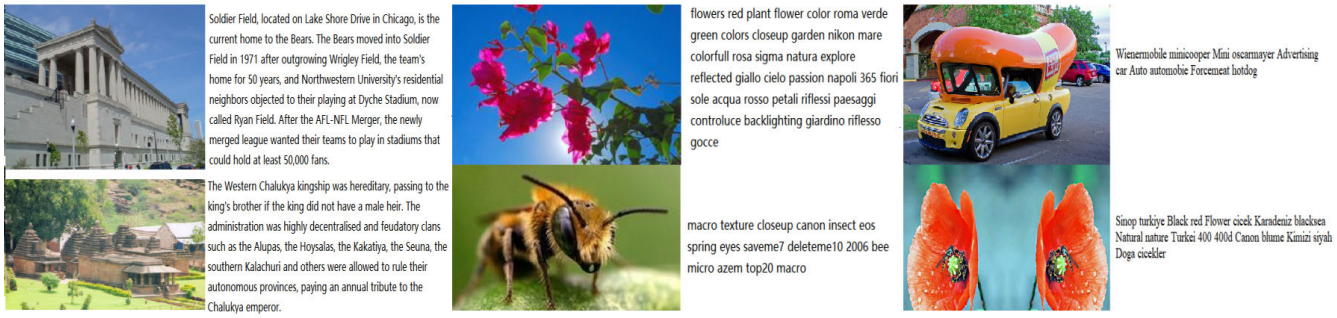

FIGURE 7. Some example on Wikipedia, NUS-WIDE and MIR-Flickr dataset.

5) MSCOCO

MSCOCO [133] This dataset contains over 120000 images from different scenes and contexts, each with multiple manually annotated descriptions. It is primarily used for tasks such as image annotation, object detection, and instance segmentation, covering 80 different object categories. The rich annotations and diversity of this dataset make it an important resource for research in the fields of computer vision as well as natural language processing, providing strong support for the training and evaluation of image to text generation models. The download link is http://cocodataset.org/.

6) XMEDIA DATASET

The XMedia dataset [134] developed by the Multimedia Computing Laboratory of Peking University through Wikipedia, Flickr, and You Tube sources. It is a large-scale dataset comprising five modalities, including text, image, audio, video, and 3-D models, with a total of 100,000 instances. These instances were categorized into 200 semantic groups, encompassing 153 artifact species and 47 animal species. The largest amount of data and the largest number of modalities. This is the largest data set with the most modalities in a cross-modal search. The download link is: http: / /www. icst. pku. edu. cn /mipl/XMediaNet.

B. EVALUATION METRICS

1) PRECISION

There are two methods for detecting the performance indicators in CMR models. The first approach uses images as input and retrieves relevant text from the dataset as output; and the second approach uses text as input and retrieves relevant images from the dataset as output. Specifically, they can be classified into two operational methods. The first method is to detect correlation evaluation between query and output; the second type examining cross-modal release for image text pairs for the first method, using the precision [135] method, which is a commonly used performance metric in information retrieval and classification tasks, used to measure the proportion of true positives in the results returned by the system. Accuracy mainly focuses on the number of results returned by the model that are truly relevant.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

Among them: True Positions (TP) refers to the number of relevant instances correctly returned by the system. False Positions (FP) refers to the number of unrelated instances returned by system errors.

2) AVERAGE PRECISION

Average Precision (AP) [136] and Mean Average Precision (MAP) [137] are two other important indicators used in the field of information retrieval to evaluate the performance of retrieval systems. Average accuracy is an indicator used to measure system performance across different queries and is, particularly suitable for binary classification tasks, such as relevance judgment in information retrieval. The calculation process is as follows. (1) For each query, the Precision Recall curve is calculated. (2) On the Recall axis, calculate the average Precision, which is the AP. (3) The APs of all queries are averaged to obtain the MAP. Its advantage is that it considers the ranking order of search results and is sensitive to the sorting of the returned results. The mean average accuracy is the average accuracy calculated across

multiple queries and is used to comprehensively evaluate the performance of the entire retrieval system. The calculation process was as follows: (1) Calculate the average accuracy (AP) for each query. (2) The average of all query APs is used to obtain the MAP. Its advantage is that it considers the performance of the system for different queries, providing a more comprehensive evaluation of the overall performance of the retrieval system. Overall, the AP is mainly used for individual queries, emphasizing the performance of the model on a particular query. MAP integrates the performance of multiple queries and provides a more comprehensive evaluation of the entire retrieval system. These two indicators are commonly used to evaluate performance in fields such as information retrieval, text retrieval, and image retrieval, particularly in tasks that require consideration of ranking and ranking quality.

$$\mathrm{AP} = \frac{1}{N}\sum_{k=1}^{N} P(k)\cdot \mathrm{rel}(k) \tag{9}$$

where, N is the total number of search results, $P(k)$ is the accuracy of the first k results, and $\mathrm{rel}(k)$ is the number of truly relevant instances in the first k results.

$$\mathrm{MAP} = \frac{1}{Q}\sum_{q=1}^{Q} \mathrm{AP}(q) \tag{10}$$

Among them, Q is the total number of queries, and $AP(q)$ is the average accuracy of the $q-th$ query.

3) PRECISION AT k (P@k)

Precision at k (P@k) is another commonly used metric for cross-modal retrieval evaluation. It measures the precision of top k retrieved items. P@k is calculated by dividing the number of relevant items among the top k retrieved items by k. P@k provides a measure of the retrieval accuracy at the top k positions, which is particularly important when users are only interested in a limited number of results.

4) SSIM AND MULTI-SSIM

The Structural Similarity Index (SSIM) is an indicator used to measure the similarity between two images, with a particular focus on three aspects: brightness, contrast, and structure. The value range of SSIM is between 0 and 1, where 1 indicates that the two images are exactly the same and 0 indicates that they are completely different. SSIM is calculated using the following formula:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{11}$$

Among them, u_x and u_y are the average values of images x and y. σ_x and σ_y are the variances of images x and y. σ_{xy} is the covariance of images x and y. C_1 and C_2 are constants used for stability.

The Multi-Scale-SSIM (MS-SSIM) is an extended version of SSIM, which considers the structural information of images at different scales to comprehensively evaluate the quality of images. MS-SSIM calculates SSIM at different resolutions and calculates the final similarity index by weighting and summing the results. By considering multi-scale information, MS-SSIM can better simulate the human visual system's perception of image quality at different resolutions, especially in the presence of image distortion, providing more accurate quality assessment.

5) PSNR AND MULTI-PSNR

Peak Signal to Noise Ratio (PSNR) is an indicator used to measure the quality of image compression, mainly focusing on the ratio between the peak signal and noise of the image. PSNR is usually expressed in decibels (dB), with higher values indicating better image quality. The calculation of PSNR is based on mean square error (MSE), and the first step is to calculate the average square of the difference between the original image and the reconstructed image. The formula is as follows:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2 \tag{12}$$

$$PSNR = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{13}$$

Among them, I and K are the original image and reconstructed image, respectively, and m and n are the width and height of the image. MAX_I^2 is the maximum possible pixel value of an image (usually 255 for 8-bit images). The higher the PSNR, the closer the reconstructed image is to the original image. PSNR is mainly used in engineering and research because it is computationally simple and related to mean square error (MSE). However, it should be noted that PSNR cannot fully reflect the perceptual quality of the human visual system.

Multi-Scale Peak Signal to Noise Ratio (MS-PSNR) is an extension of PSNR used to evaluate image quality at different scales. Unlike PSNR, which only evaluates image quality at a single resolution, MS-PSNR considers the detailed information of images at multiple resolutions. By considering multi-scale information, MS-PSNR can provide a more comprehensive and robust image quality assessment than single scale PSNR. MS-PSNR is more commonly used in some advanced image processing applications because it can better reflect the quality changes of images at different scales.

In summary, PSNR is a fundamental indicator for measuring image quality and is suitable for simple image quality comparisons. MS-PSNR provides more comprehensive and accurate quality measurements through multi-scale evaluation, making it particularly suitable for complex image processing tasks and advanced image quality evaluation.

VI. FUTURE WORK

Despite significant progress made in cross-modal retrieval, several challenges and open research questions remain. In this section, we summarize a few points about the future direction of cross-modal retrieval.

A. HIERARCHICAL SEMANTIC ALIGNMENT

It is crucial to achieve fine-grained hierarchical semantic alignment between different modalities, as well as effective integration of coarse-grained and fine-grained semantics. In the process of multi-granularity semantic fusion representation learning, we need to mine fine-grained semantic features from multi-modal data, identify hierarchical semantic relationships composed of ''semantic objects relationships scenes'' from multi-modal data, achieve cross-modal fine-grained hierarchical semantic alignment, and effectively improve the discriminability of cross-modal semantic representation. Therefore, achieving fine-grained hierarchical semantic alignment between different modalities is particularly important for improving hash code generation quality.

B. MULTI-LEVEL CORRELATION LEARNING

In multi-view correlation mining, deep representation learning and correlation analysis methods are combined to synchronously mine nonlinear correlations between views. The existing unsupervised multi-view correlation learning methods mainly focus on pairwise view correlation mining, and are mostly implemented by shallow models, making it difficult to achieve synchronous mining of nonlinear semantic correlations in multiple views. Meanwhile, in the process of multi-view feature learning, existing methods have not considered the mining of view invariant features at different levels, which are crucial for the representation of multi-view semantic correlations. Therefore, organically combining deep representation models with multi-view correlation learning methods, mining multi-level key features, and improving the performance of unsupervised cross-modal hash learning is another key issue that needs further research in cross-modal retrieval in the future.

C. COMPLETE CONSISTENCY CONSTRAINT

In the process of cross domain knowledge transfer learning, research cross-domain and cross-modal consistency constraints to narrow the ''heterogeneous gap'' between different domains and modes. Usually, data from different fields exhibit significant heterogeneity in terms of structural information, feature distribution, semantic categories, and modalities, making it difficult to achieve effective and accurate semantic knowledge transfer between different fields. Existing domain adaptation methods mainly focus on maintaining consistency in one or two aspects, without simultaneously considering structural heterogeneity, domain heterogeneity, semantic heterogeneity, and modal heterogeneity. This makes it difficult for existing methods to effectively address cross-modal and cross-domain semantic knowledge transfers, making it difficult to provide effective support for unsupervised cross-modal hash learning. Therefore, how to conduct effective theoretical analysis on the four ''heterogeneity'' mentioned above, provide complete consistency constraints for cross-domain and cross-modal knowledge transfer learning, and effectively narrow the ''heterogeneity gap'', is a scientific problem that requires further research focus in cross-modal retrieval in the future.

D. DESIGN INTERACTION AND FEEDBACK MECHANISMS

By utilizing advanced natural language processing technology, design an intuitive and user-friendly user interface that enables users to describe their needs through natural language, enabling them to easily perform cross-modal retrieval. By analyzing user behavior (such as clicks, dwell time, scrolling, etc.), the system can infer user interests and needs, thereby optimizing search results. Allow users to provide feedback and evaluation on search results. Through continuous user feedback, the system can adaptively adjust the weights of different modal features, optimize search algorithms and model parameters, and ensure continuous improvement and optimization of the system.

E. THE APPLICATION OF CROSS-MODAL RETRIEVAL IN VARIOUS FIELDS

Future research will further enhance the accuracy, efficiency, and intelligence level of cross-modal retrieval technology, providing more intelligent and efficient information retrieval solutions for various industries through its wide applicability and powerful functions. For example, medical researchers can quickly find relevant images and text reports from a large case library through cross-modal retrieval systems for scientific research and teaching. The monitoring system can input text descriptions, such as ''searching for running people,'' and retrieve relevant clips from the surveillance video to improve the efficiency of public safety monitoring. The teaching system utilizes cross-modal retrieval technology to provide an interactive learning experience, such as retrieving relevant video or image explanations based on student questions.

F. PRIVACY AND SECURITY

With the increasing diversity and complexity of data, how to achieve efficient cross-modal retrieval while ensuring user privacy and system security has become an important research direction. Privacy protection and security are key issues that must be taken seriously. By implementing various technical means such as data anonymization, encryption technology, access control, anomaly detection, and complying with relevant laws and regulations, the privacy protection and security of the system can be effectively improved, ensuring the security of user data and the stable operation of the system. Future research will continue to explore more advanced and intelligent privacy protection and security technologies, utilizing machine learning and big data analysis techniques to detect and identify abnormal behaviors in the

system, such as abnormal login and data leakage, and taking timely measures to prevent security risks, providing solid support for the development of cross-modal retrieval systems.

VII. CONCLUSION

In this paper, we provide a comprehensive review of cross-modal retrieval, covering traditional CCA-based methods, deep learning approaches, and cross-modal retrieval hashing methods. We discussed the basic models and implementation processes of these methods, enabling readers to grasp the fundamental principles of cross-modal retrieval. We hope to facilitate adoption and further research in this area by presenting the basic concepts, techniques, and evaluation metrics. Moreover, we explored emerging trends and future directions, including multi-modal fusion techniques, deep metric learning, cross-modal adversarial learning, deep reinforcement learning, and explainable cross-modal retrieval. The aim of this review is to serve as a valuable reference for researchers and practitioners in the realm of cross-modal retrieval, sparking further advancements in this dynamic field.

REFERENCES

[1] Y. Zeng, Y. Wang, D. Liao, G. Li, W. Huang, J. Xu, D. Cao, and H. Man, "Keyword-based diverse image retrieval with variational multiple instance graph," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10528–10537, Dec. 2023.
[2] C. Zhang and T. Chen, "An active learning framework for content-based information retrieval," *IEEE Trans. Multimedia*, vol. 4, no. 2, pp. 260–268, Jun. 2002.
[3] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.
[4] L. Zhang, X. Zhang, and J. Pan, "Hierarchical cross-modality semantic correlation learning model for multimodal summarization," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 10, pp. 11676–11684.
[5] C. Zhang, J. Song, X. Zhu, L. Zhu, and S. Zhang, "HCMSL: Hybrid cross-modal similarity learning for cross-modal retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1s, pp. 1–22, Jan. 2021.
[6] Z. Al-Halah, T. Gehrig, and R. Stiefelhagen, "Learning semantic attributes via a common latent space," in *Proc. Int. Conf. Comput. Vis. Theory Appl. (VISAPP)*, vol. 2, Jan. 2014, pp. 48–55.
[7] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," 2016, *arXiv:1607.06215*.
[8] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: A review," *Comput. Sci. Rev.*, vol. 39, Feb. 2021, Art. no. 100336.
[9] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[10] W. Burger and M. J. Burge, "Scale-invariant feature transform (SIFT)," in *Digital Image Processing: An Algorithmic Introduction*. Cham, Switzerland: Springer, 2022, pp. 709–763.
[11] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, *arXiv:1511.08458*.
[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[13] W. Brendel and M. Bethge, "Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet," 2019, *arXiv:1904.00760*.
[14] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, Jun. 1999, pp. 1150–1157.
[15] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 251–260.
[16] S. J. Hwang and K. Grauman, "Learning the relative importance of objects from tagged images for retrieval and cross-modal search," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 134–153, Nov. 2012.
[17] G. Andrew, R. Arora, and J. Bilmes, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[18] V. Ranjan, N. Rasiwasia, and C. V. Jawahar, "Multi-label cross-modal retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4094–4102.
[19] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *Proc. Int. Stat. Optim. Perspect. Workshop Subspace, Latent Struct. Feature Selection*, 2005, pp. 34–51.
[20] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Comput.*, vol. 12, no. 6, pp. 1247–1283, Jun. 2000.
[21] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia MULTIMEDIA*, 2003, pp. 604–611.
[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[26] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.
[27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
[28] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.
[29] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.
[30] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 2222–2230.
[31] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 7–16.
[32] F. Wu, X. Lu, J. Song, S. Yan, Z. M. Zhang, Y. Rui, and Y. Zhuang, "Learning of multimodal representations with random walks on the click graph," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 630–642, Feb. 2016.
[33] Y. He, S. Xiang, C. Kang, J. Wang, and C. Pan, "Cross-modal retrieval via deep and bidirectional representation learning," *IEEE Trans. Multimedia*, vol. 18, no. 7, pp. 1363–1377, Jul. 2016.
[34] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1371–1384, Aug. 2008.
[35] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[36] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 154–162.
[37] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2400–2413, Jun. 2020.
[38] F. Wu, X.-Y. Jing, Z. Wu, Y. Ji, X. Dong, X. Luo, Q. Huang, and R. Wang, "Modality-specific and shared generative adversarial network for cross-modal retrieval," *Pattern Recognit.*, vol. 104, Aug. 2020, Art. no. 107335.
[39] X. Xu, J. Tian, K. Lin, H. Lu, J. Shao, and H. T. Shen, "Zero-shot cross-modal retrieval by assembling AutoEncoder and generative adversarial network," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 1s, pp. 1–17, Jan. 2021.
[40] S. He, W. Wang, Z. Wang, X. Xu, Y. Yang, X. Wang, and H. T. Shen, "Category alignment adversarial learning for cross-modal retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 4527–4538, May 2023.

[41] S. Wang, R. Wang, Z. Yao, S. Shan, and X. Chen, "Cross-modal scene graph matching for relationship-aware image-text retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1497–1506.

[42] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10918–10927.

[43] W. Wang, Y. Shen, H. Zhang, Y. Yao, and L. Liu, "Set and rebase: Determining the semantic graph connectivity for unsupervised cross-modal hashing," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 853–859.

[44] Y. Cheng, X. Zhu, J. Qian, F. Wen, and P. Liu, "Cross-modal graph matching network for image-text retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 4, pp. 1–23, Nov. 2022.

[45] Z. Han, A. B. Azman, F. B. Khalid, and M. R. B. Mustaffa, "Multi-granularity semantic information integration graph for cross-modal hash retrieval," *IEEE Access*, vol. 12, pp. 44682–44694, 2024.

[46] J. Pei, K. Zhong, Z. Yu, L. Wang, and K. Lakshmanna, "Scene graph semantic inference for image and text matching," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 22, no. 5, pp. 1–23, May 2023.

[47] N. Messina, G. Amato, A. Esuli, F. Falchi, C. Gennaro, and S. Marchand-Maillet, "Fine-grained visual textual alignment for cross-modal retrieval using transformer encoders," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 17, no. 4, pp. 1–23, Nov. 2021.

[48] J. Wang et al., "Hugs are better than handshakes: Unsupervised cross-modal transformer hashing with multi-granularity alignment," in *Proc. 33rd Brit. Mach. Vis. Conf.*, 2022, p. 1035.

[49] L. Zhang, H. Wu, Q. Chen, Y. Deng, J. Siebert, Z. Li, Y. Han, D. Kong, and Z. Cao, "VLDeformer: Vision–language decomposed transformer for fast cross-modal retrieval," *Knowl.-Based Syst.*, vol. 252, Sep. 2022, Art. no. 109316.

[50] W. Zhang, Q. Tan, P. Li, Q. Zhang, and R. Wang, "Cross-modal transformer with language query for referring image segmentation," *Neurocomputing*, vol. 536, pp. 191–205, Jun. 2023.

[51] X. Qu, P. Tang, Z. Zou, Y. Cheng, J. Dong, P. Zhou, and Z. Xu, "Fine-grained iterative attention network for temporal language localization in videos," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4280–4288.

[52] Y. Chen and L. Bazzani, "Learning joint visual semantic matching embeddings for language-guided retrieval," in *Proc. Eur. Conf. Comput. Vis.*, vol. 12367, Nov. 2020, pp. 136–152.

[53] Q. Zhang, Z. Lei, Z. Zhang, and S. Z. Li, "Context-aware attention network for image-text retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3533–3542.

[54] K. Ning, M. Cai, D. Xie, and F. Wu, "An attentive sequence to sequence translator for localizing video clips by natural language," *IEEE Trans. Multimedia*, vol. 22, no. 9, pp. 2434–2443, Jun. 2020.

[55] Y. Wu, S. Wang, and Q. Huang, "Multi-modal semantic autoencoder for cross-modal retrieval," *Neurocomputing*, vol. 331, pp. 165–175, Feb. 2019.

[56] J. Wu, X. Xie, L. Nie, Z. Lin, and H. Zha, "Reconstruction regularized low-rank subspace learning for cross-modal retrieval," *Pattern Recognit.*, vol. 113, May 2021, Art. no. 107813.

[57] D. Zhang and X.-J. Wu, "Robust and discrete matrix factorization hashing for cross-modal retrieval," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108343.

[58] C. Xie, Y. Gao, Q. Zhou, and J. Zhou, "Multi-similarity reconstructing and clustering-based contrastive hashing for cross-modal retrieval," *Inf. Sci.*, vol. 647, Nov. 2023, Art. no. 119543.

[59] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger, "Learning to rank with (a lot of) word features," *Inf. Retr.*, vol. 13, no. 3, pp. 291–314, Jun. 2010.

[60] X. Jiang, F. Wu, X. Li, Z. Zhao, W. Lu, S. Tang, and Y. Zhuang, "Deep compositional cross-modal learning to rank via local–global alignment," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 69–78.

[61] Q. Xu, M. Li, and M. Yu, "Learning to rank with relational graph and pointwise constraint for cross-modal retrieval," *Soft Comput.*, vol. 23, no. 19, pp. 9413–9427, Oct. 2019.

[62] Y. Wu, S. Wang, and Q. Huang, "Online fast adaptive low-rank similarity learning for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 5, pp. 1310–1322, May 2020.

[63] J. Wei, Y. Yang, X. Xu, X. Zhu, and H. T. Shen, "Universal weighting metric learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6534–6545, Oct. 2022.

[64] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, vol. 2, 2011, pp. 1360–1365.

[65] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3594–3601.

[66] J. Tang, K. Wang, and L. Shao, "Supervised matrix factorization hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 3157–3166, Jul. 2016.

[67] X. Liu, A. Li, J.-X. Du, S.-J. Peng, and W. Fan, "Efficient cross-modal retrieval via flexible supervised collective matrix factorization hashing," *Multimedia Tools Appl.*, vol. 77, no. 21, pp. 28665–28683, Nov. 2018.

[68] D. Mandai and S. Biswas, "Label consistent matrix factorization based hashing for cross-modal retrieval," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2901–2905.

[69] C.-X. Li, Z.-D. Chen, P.-F. Zhang, X. Luo, L. Nie, W. Zhang, and X.-S. Xu, "SCRATCH: A scalable discrete matrix factorization hashing for cross-modal retrieval," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1–9.

[70] D. Wang, X. Gao, X. Wang, and L. He, "Label consistent matrix factorization hashing for large-scale cross-modal similarity search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2466–2479, Oct. 2019.

[71] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, "MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 964–981, Mar. 2021.

[72] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3270–3278.

[73] E. Yang et al., "Pairwise relationship guided deep hashing for cross-modal retrieval," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1618–1625.

[74] Z. D. Chen et al., "Dual deep neural networks cross-modal hashing," in *Proc. 32nd AAAI Conf. Artif. Intell., 30th Innov. Appl. Artif. Intell. Conf., 8th AAAI Symp. Educ. Adv. Artif. Intell.*, 2018, pp. 274–281.

[75] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.

[76] L. Ma, H. Li, F. Meng, Q. Wu, and K. N. Ngan, "Global and local semantics-preserving based deep hashing for cross-modal retrieval," *Neurocomputing*, vol. 312, pp. 49–62, Oct. 2018.

[77] Y. Shi, X. You, F. Zheng, S. Wang, and Q. Peng, "Equally-guided discriminative hashing for cross-modal retrieval," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4767–4773.

[78] D. Hu, F. Nie, and X. Li, "Deep binary reconstruction for cross-modal hashing," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 973–985, Apr. 2019.

[79] X. Xia, G. Dong, F. Li, L. Zhu, and X. Ying, "When CLIP meets cross-modal hashing retrieval: A new strong baseline," *Inf. Fusion*, vol. 100, Dec. 2023, Art. no. 101968.

[80] D. Xie, C. Deng, C. Li, X. Liu, and D. Tao, "Multi-task consistency-preserving adversarial hashing for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 29, pp. 3626–3637, 2020.

[81] X. Nie, B. Wang, J. Li, F. Hao, M. Jian, and Y. Yin, "Deep multiscale fusion hashing for cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 401–410, Jan. 2021.

[82] C. Hou, Z. Li, Z. Tang, X. Xie, and H. Ma, "Multiple instance relation graph reasoning for cross-modal hash retrieval," *Knowl.-Based Syst.*, vol. 256, Nov. 2022, Art. no. 109891.

[83] Z. Shu, K. Yong, J. Yu, S. Gao, C. Mao, and Z. Yu, "Discrete asymmetric zero-shot hashing with application to cross-modal retrieval," *Neurocomputing*, vol. 511, pp. 366–379, Oct. 2022.

[84] L. Li, Z. Shu, Z. Yu, and X.-J. Wu, "Robust online hashing with label semantic enhancement for cross-modal retrieval," *Pattern Recognit.*, vol. 145, Jan. 2024, Art. no. 109972.

[85] T. Wang, L. Zhu, Z. Zhang, H. Zhang, and J. Han, "Targeted adversarial attack against deep cross-modal hashing retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 6159–6172, Oct. 2023.

[86] Y. Hu, Z. Jin, H. Ren, D. Cai, and X. He, "Iterative multi-view hashing for cross media indexing," in *Proc. 22nd ACM Int. Conf. Multimedia*, Nov. 2014, pp. 527–536.

[87] G. Irie, H. Arai, and Y. Taniguchi, "Alternating co-quantization for cross-modal hashing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1886–1894.
[88] B. Wu, Q. Yang, and W. S. Zheng, "Quantized correlation hashing for fast cross-modal search," in *Proc. 24th Int. Conf. Artif. Intell.*, 2015, pp. 3946–3952.
[89] G. Ding, Y. Guo, J. Zhou, and Y. Gao, "Large-scale cross-modality search via collective matrix factorization hashing," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5427–5440, Nov. 2016.
[90] D. Rafailidis and F. Crestani, "Cluster-based joint matrix factorization hashing for cross-modal retrieval," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2016, pp. 781–784.
[91] Y. Fang, H. Zhang, and Y. Ren, "Unsupervised cross-modal retrieval via multi-modal graph regularized smooth matrix factorization hashing," *Knowl.-Based Syst.*, vol. 171, pp. 69–80, May 2019.
[92] M. Cheng, L. Jing, and M. K. Ng, "Robust unsupervised cross-modal hashing for multimedia retrieval," *ACM Trans. Inf. Syst.*, vol. 38, no. 3, pp. 1–25, Jul. 2020.
[93] V. E. Liong, J. Lu, Y.-P. Tan, and J. Zhou, "Cross-modal deep variational hashing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4097–4105.
[94] G. Wu, Z. Lin, and J. Han, "Unsupervised deep hashing via binary latent factor models for large-scale cross-modal retrieval," in *Proc. IJCAI*, 2018, pp. 2854–2860.
[95] S. Su, Z. Zhong, and C. Zhang, "Deep joint-semantics reconstructing hashing for large-scale unsupervised cross-modal retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3027–3035.
[96] J. Huang, C. Min, and L. Jing, "Unsupervised deep fusion cross-modal hashing," in *Proc. Int. Conf. Multimodal Interact.*, Oct. 2019, pp. 358–366.
[97] D. Yang, D. Wu, W. Zhang, H. Zhang, B. Li, and W. Wang, "Deep semantic-alignment hashing for unsupervised cross-modal retrieval," in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2020, pp. 44–52.
[98] S. Liu, S. Qian, Y. Guan, J. Zhan, and L. Ying, "Joint-modal distribution-based similarity hashing for large-scale unsupervised deep cross-modal retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1379–1388.
[99] T. Hoang, T.-T. Do, T. V. Nguyen, and N.-M. Cheung, "Unsupervised deep cross-modality spectral hashing," *IEEE Trans. Image Process.*, vol. 29, pp. 8391–8406, 2020.
[100] T. Wang, L. Zhu, Z. Cheng, J. Li, and Z. Gao, "Unsupervised deep cross-modal hashing with virtual label regression," *Neurocomputing*, vol. 386, pp. 84–96, Apr. 2020.
[101] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, and H. T. Shen, "Unsupervised cross-modal retrieval through adversarial learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 1153–1158.
[102] J. Zhang, Y. Peng, and M. Yuan, "Unsupervised generative adversarial cross-modal hashing," in *Proc. 32nd AAAI Conf. Artif. Intell., 30th Innov. Appl. Artif. Intell. Conf., 8th AAAI Symp. Educ. Adv. Artif. Intell.*, 2018, pp. 539–546.
[103] C. Li, C. Deng, and L. Wang, "Coupled CycleGAN: Unsupervised hashing network for cross-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 176–183.
[104] J. Zhang and Y. Peng, "Multi-pathway generative adversarial hashing for unsupervised cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 174–187, Jan. 2020.
[105] X. Shen, H. Zhang, L. Li, and L. Liu, "Attention-guided semantic hashing for unsupervised cross-modal retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2021, pp. 1–6.
[106] X. Shen, H. Zhang, L. Li, Z. Zhang, D. Chen, and L. Liu, "Clustering-driven deep adversarial hashing for scalable unsupervised cross-modal retrieval," *Neurocomputing*, vol. 459, pp. 152–164, Oct. 2021.
[107] G. Mikriukov, M. Ravanbakhsh, and B. Demir, "An unsupervised cross-modal hashing method robust to noisy training image-text correspondences in remote sensing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 2556–2560.
[108] Y. Shi, Y. Zhao, X. Liu, F. Zheng, W. Ou, X. You, and Q. Peng, "Deep adaptively-enhanced hashing with discriminative similarity guidance for unsupervised cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 7255–7268, Oct. 2022.
[109] T. Hoang, T.-T. Do, T. V. Nguyen, and N.-M. Cheung, "Multimodal mutual information maximization: A novel approach for unsupervised deep cross-modal hashing," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6289–6302, Sep. 2023.
[110] P. Hu, H. Zhu, J. Lin, D. Peng, Y.-P. Zhao, and X. Peng, "Unsupervised contrastive cross-modal hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3877–3889, Mar. 2023.
[111] L. Zhu, X. Wu, J. Li, Z. Zhang, W. Guan, and H. T. Shen, "Work together: Correlation-identity reconstruction hashing for unsupervised cross-modal retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 9, pp. 8838–8851, Sep. 2023.
[112] J. Wang, G. Li, P. Pan, and X. Zhao, "Semi-supervised semantic factorization hashing for fast cross-modal retrieval," *Multimedia Tools Appl.*, vol. 76, no. 19, pp. 20197–20215, Oct. 2017.
[113] X. Wang, W. Zhu, and C. Liu, "Semi-supervised deep quantization for cross-modal search," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1730–1739.
[114] X. Zhang, X. Tian, B. Yang, Z. Zhang, and Y. Li, "Semi-supervised cross-modal hashing based on label prediction and distance preserving," in *Proc. IEEE 31st Int. Conf. Tools With Artif. Intell. (ICTAI)*, Nov. 2019, pp. 324–330.
[115] X. Wang, X. Liu, Z. Hu, N. Wang, W. Fan, and J.-X. Du, "Semi-supervised semantic-preserving hashing for efficient cross-modal retrieval," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2019, pp. 1006–1011.
[116] J. Zhang, Y. Peng, and M. Yuan, "SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network," *IEEE Trans. Cybern.*, vol. 50, no. 2, pp. 489–502, Feb. 2020.
[117] D. Mandal, P. Rao, and S. Biswas, "Label prediction framework for semi-supervised cross-modal retrieval," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2311–2315.
[118] Z. Shen, D. Zhai, X. Liu, and J. Jiang, "Semi-supervised graph convolutional hashing network for large-scale cross-modal retrieval," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2366–2370.
[119] X. Huang, Y. Peng, and M. Yuan, "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1047–1059, Mar. 2020.
[120] X. Huang and Y. Peng, "TPCKT: Two-level progressive cross-media knowledge transfer," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2850–2862, Nov. 2019.
[121] X. Wen, Z. Han, X. Yin, and Y.-S. Liu, "Adversarial cross-modal retrieval via learning and transferring single-modal similarities," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 478–483.
[122] L. Zhen, P. Hu, X. Peng, R. S. M. Goh, and J. T. Zhou, "Deep multimodal transfer learning for cross-modal retrieval," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 798–810, Feb. 2022.
[123] Y. Peng and J. Chi, "Unsupervised cross-media retrieval using domain adaptation with scene graph," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4368–4379, Nov. 2020.
[124] X. Shen, H. Zhang, L. Li, W. Yang, and L. Liu, "Semi-supervised cross-modal hashing with multi-view graph representation," *Inf. Sci.*, vol. 604, pp. 45–60, Aug. 2022.
[125] F. Wu, S. Li, G. Gao, Y. Ji, X.-Y. Jing, and Z. Wan, "Semi-supervised cross-modal hashing via modality-specific and cross-modal graph convolutional networks," *Pattern Recognit.*, vol. 136, Apr. 2023, Art. no. 109211.
[126] X. Zhang, X. Liu, X. Nie, X. Kang, and Y. Yin, "Semi-supervised semi-paired cross-modal hashing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 6517–6529, Jul. 2024.
[127] W. Fan, C. Zhang, H. Li, X. Jia, and G. Wang, "Three-stage semisupervised cross-modal hashing with pairwise relations exploitation," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Apr. 6, 2023, doi: 10.1109/TNNLS.2023.3263221.
[128] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[129] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
[130] T. S. Chua et al., "NUS-WIDE: A real-world web image database from national University of Singapore," in *Proc. ACM Int. Conf. Image Video Retr.*, 2009, pp. 1–9.
[131] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.
[132] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proc. 1st ACM Int. Conf. Multimedia Inf. Retr.*, Oct. 2008, pp. 39–43.
[133] T. Lin, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[134] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 583–596, Mar. 2016.
[135] Y. Xia, Y. Wu, and J. Feng, "Cross-media retrieval using probabilistic model of automatic image annotation," *Int. J. Signal Process., Image Process. Pattern Recognit.*, vol. 8, no. 4, pp. 145–154, Apr. 2015.
[136] B. Jiang, J. Yang, Z. Lv, K. Tian, Q. Meng, and Y. Yan, "Internet cross-media retrieval based on deep learning," *J. Vis. Commun. Image Represent.*, vol. 48, pp. 356–366, Oct. 2017.
[137] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 39, no. 1, pp. 1–22, Sep. 1977.

ZHICHAO HAN was born in Hebei, China. He received the M.Sc. degree from Guangxi Normal University, China, in 2018. He is currently pursuing the Ph.D. degree with the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia. His research interests include machine learning, deep learning, and cross-modal retrieval.

AZREEN BIN AZMAN (Member, IEEE) received the Diploma degree in software engineering from the Institute of Telecommunication and Information Technology, in 1997, the Bachelor of Information Technology degree in information systems engineering from Multimedia University, Malaysia, in 1999, and the Ph.D. degree in computing science (information retrieval) from the University of Glasgow, Scotland, in September 2007. Before joining the Ph.D. degree, he served in the industry for a few years. He is currently an Associate Professor with Universiti Putra Malaysia. His current research interests include information retrieval, text mining, natural language processing, and intelligent systems. He serves as a Committee Member for Malaysian Society of Information Retrieval and Knowledge Management (PECAMP) and Malaysian Information Technology Society (MITS).

MAS RINA BINTI MUSTAFFA (Senior Member, IEEE) received the Ph.D. degree in multimedia systems from Universiti Putra Malaysia, Malaysia. She is currently an Associate Professor with Universiti Putra Malaysia. Her current research interests include multimedia information retrieval, computer vision, pattern recognition, image processing, multimedia computing, and multimedia systems and applications.

FATIMAH BINTI KHALID (Member, IEEE) received the B.Sc. degree in computer science from the University of Technology Malaysia (UTM), in 1992, and the master's and Ph.D. degrees in system science and management from Universiti Kebangsaan Malaysia (UKM), in 1997 and 2008, respectively. From 1993 to 1995, she was a System Analyst with UKM. After the master's degree, she started involved in teaching with the Sal College, until 1999, and continued with Universiti Putra Malaysia, in June 1999. In January 2016, she was a Secondment with the Computer Science Department, Tabuk University, Saudi Arabia, for one and a half years. Currently, she is a Lecturer and an Associate Professor with the Faculty of Computer Science and Information Technology.