IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 11 January 2024, accepted 18 January 2024, date of publication 24 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3357939

SURVEY

A Systematic Literature Review of Deep Learning Approaches for Sketch-Based Image Retrieval: Datasets, Metrics, and Future Directions

FAN YANG[1], NOR AZMAN ISMAIL[1], (Member, IEEE), YEE YONG PANG[1], VICTOR R. KEBANDE[2], (Member, IEEE), ARAFAT AL-DHAQM[3], AND TIENG WEI KOH[3]

[1]Faculty of Computing, Universiti Teknologi Malaysia (UTM), Skudai, Johor 81310, Malaysia
[2]Department of Computer Science (DIDA), Blekinge Institute of Technology, 37179 Karlskrona, Sweden
[3]Computer and Information Sciences Department, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Perak 32610, Malaysia

Corresponding authors: Fan Yang (fyang@graduate.utm.my) and Victor R. Kebande (victor.kebande@bth.se)

This work was supported by the Blekinge Institute of Technology, Sweden, through the Grant Funded Research.

ABSTRACT Sketch-based image retrieval (SBIR) utilizes sketches to search for images containing similar objects or scenes. Due to the proliferation of touch-screen devices, sketching has become more accessible and therefore has received increasing attention. Deep learning has emerged as a potential tool for SBIR, allowing models to automatically extract image features and learn from large amounts of data. To the best of our knowledge, there is currently no systematic literature review (SLR) of SBIR with deep learning. Therefore, the aim of this review is to incorporate related works into a systematic study, highlighting the main contributions of individual researchers over the years, with a focus on past, present and future trends. To achieve the purpose of this study, 90 studies from 2016 to June 2023 in 4 databases were collected and analyzed using the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) framework. The specific models, datasets, evaluation metrics, and applications of deep learning in SBIR are discussed in detail. This study found that Convolutional Neural Networks (CNN) and Generative Adversarial Networks (GAN) are the most widely used deep learning methods for SBIR. A commonly used dataset is Sketchy, especially in the latest Zero-shot sketch-based image retrieval (ZS-SBIR) task. The results show that Mean Average Precision (mAP) is the most commonly used metric for quantitative evaluation of SBIR. Finally, we provide some future directions and guidance for researchers based on the results of this review.

INDEX TERMS Sketch-based image retrieval, SBIR, SLR, PRISMA, deep learning.

I. INTRODUCTION

Text-based image retrieval (TBIR) [1], [2] and content-based image retrieval (CBIR) [3], [4] are dominant paradigms in the field of image retrieval. TBIR facilitates the identification of images relevant to a given natural language query by utilizing the query itself. CBIR employs inherent content features of an image for similarity matching. These features encompass attributes like color, texture, shape, structure, and other higher-level characteristics extracted from the image [5], [6]. However, in practical scenarios, users may encounter challenges when attempting to articulate the desired image solely through keywords. Moreover, locating a suitable natural image that precisely encapsulates the user's retrieval intent can be a complex endeavor [7]. An alternative approach involves users expressing their retrieval objectives through hand-drawn sketches. SBIR emerging as a distinct iteration of CBIR assumes significance within the field of computer vision. This research integrates methodologies from diverse areas, including image processing [8], [9], pattern recognition [10], [11], and human-computer interaction (HCI) [12], [13]. SBIR empowers users to search for similar images through sketches. Users can depict the object or scene of interest by sketching fundamental lines, outlines, or defining

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

TABLE 1. Summary of the SBIR challenges.

Challenges	Description
Feature Extraction	Variations in style, stroke quality and level of abstraction make sketch feature extraction challenging.
Cross-Domain Retrieval	Retrieving natural images across different domains through sketches increases the complexity of the problem due to representation and domain gaps, and differences between sketches and natural images affect retrieval accuracy.
Limited Data	Compared to natural images, sketch resources are lacking. The scarce annotated data further affects the generalization and robustness of the model.
Interaction Design	Design intuitive interfaces and interactions for effective sketch input and user experience.

key shapes. This novel mode of retrieval enriches the interactive and communicative dimensions of image retrieval.

However, unlike natural images with rich color and texture information, sketches are sparse, abstract, and limited by the unique characteristics of the creator, making SBIR challenging. We summarize the challenges faced by SBIR in TABLE 1, which are mainly divided into feature extraction, cross-domain retrieval, limited data, and interaction design.

To address the above issues, researchers have been working on SBIR for decades. Manual methods and cross-domain deep learning methods are classified as two main categories of solutions. Early research performed image retrieval by manually extracting edge contours from natural images, which was limited by the matching of the extracted edge maps of natural images with sketches with large changes and ambiguities [14], [15], [16]. In recent years, due to the rapid development of deep learning technology, it has achieved state-of-the-art performance in various computer vision tasks. Artificial neural networks eschew the need for handcrafted features, instead autonomously acquiring intricate and complex features. This also brings new possibilities to SBIR research.

The general strategy of SBIR is to retrieve similar target images from the data pool based on sketch queries. SBIR is challenging because the system's input is a sparse and abstract sketch, and the output is a natural image composed of dense pixels that belong to entirely different domains. The most common strategy for SBIR with deep learning is to train the joint embedding space and perform nearest neighbor retrieval to solve the cross-domain problem. This solution focuses on the design of the network architecture and loss function. Typical examples are the Siamese network and the Triplet network, as shown in Fig. 1. A Siamese CNN aims to learn a target space in which the distances of similar sketch-image pairs are reduced, and the distances of dissimilar sketch-image pairs are enlarged [17]. The Siamese network architecture inputs sketch S and edge map of the natural image I into the same two CNNs. Y is a binary label. When S and I belong to the same category, the value is 0.

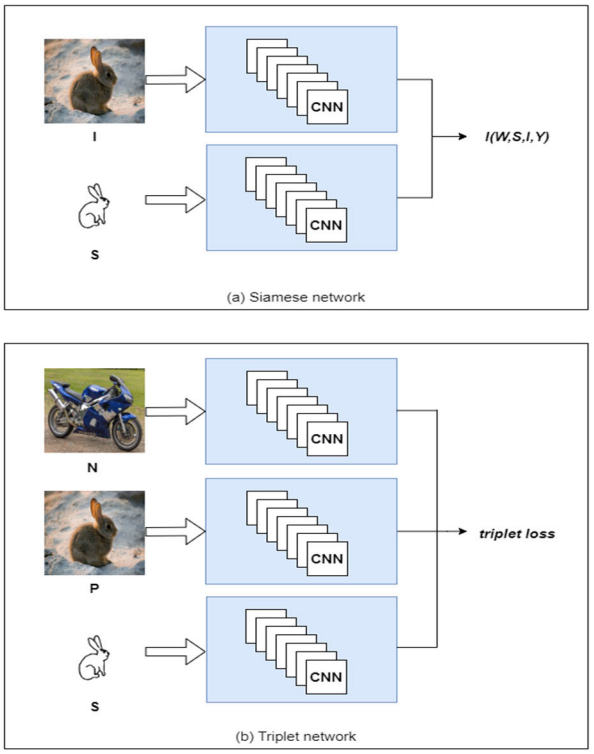

FIGURE 1. Illustrations of siamese network and triplet network.

Otherwise, the value is 1. W is the parameter to be learned. Therefore, the loss function can be written as

$$L(W) = \sum_{i=1}^{N} l(W, (S, I, Y)^i) \tag{1}$$

where $(S, I, Y)^i$ is the i-th training sample. The CNN with triplet architecture is an extension of the Siamese CNN and is trained with sketches, natural images of the same category as the sketches, natural images of different categories from the sketches, and then ranked using a distance metric comparison. The triplet network uses three branches with input data of sketch S, positive image P, and negative image N [18]. The triplet loss function guides the training. The triplet loss function is defined as:

$$L = \frac{1}{2N} \sum_{i=1}^{N} max\left[0, m + |S_i - P_i|^2 - |S_i - N_i|^2\right] \tag{2}$$

where m is the distance margin. Due to their excellent performance, triplet networks are the most commonly used model framework in SBIR. Another strategy is to use a generative method to convert a sketch into a pseudo-image, extract the features of the pseudo-image and the natural image, and compare the similarity. Conditional GAN generates pseudo-images by modeling the conditional distribution of the real image conditional on the sketch. This strategy transforms the cross-domain problem into a retrieval problem for the same domain.

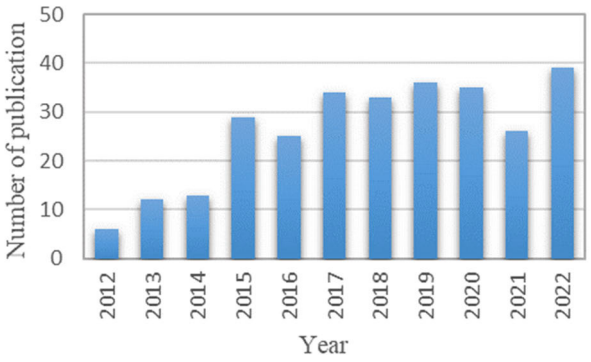

FIGURE 2. Distribution of web of science articles about SBIR from 2013 to 2022.

This research conducts a SLR on SBIR with deep learning. By analyzing previous works, this study has made it possible to emphasize existing research gaps in specific areas of deep learning methodology.

The main contributions of this study are as follows:

1) A comprehensive SLR incorporating 90 primary studies, structured following the PRISMA framework, providing a thorough and organized synthesis of existing research;

2) A detailed analysis of SBIR in deep learning, exploring the distinct dimensions: approaches, datasets, evaluation metrics, and potential applications;

3) An overview of a new perspective on the future research of SBIR.

The structure of this paper is organized as follows. Section II reviews the related work. Section III presents the research methodology adopted in conducting the systematic review. Section IV is an overview of the existing approaches. Section V contains the results and discussion, and Section VI offers the conclusion and limitation.

II. RELATED WORK

Indeed, by typing ''sketch-based image retrieval'' into Web of Science and setting the period from 2012 to 2022, the results obtained are visualized in terms of publication years on the horizontal axis and the number of records searched on the vertical axis, which shows that the level of publication activity in SBIR-related areas of research has been increasing, as shown in Fig. 2.

In this section, we discuss the related survey and review that examined SBIR. Indu and Kavitha [19] outlined the aim of SBIR methods, analyzed and reviewed diverse image retrieval approaches encompassing TBIR, CBIR, and SBIR. As a comprehensive review, the paper does not delve into the intricate details of individual works but strives to offer an exhaustive assessment of prevalent traditional image retrieval systems. It primarily focuses on methodologies, approaches, and the challenges associated with devising effective retrieval systems. Li and Li [7] provided a comprehensive review of SBIR by analyzing representative papers that have studied the SBIR problem. The paper endeavors to address two pivotal yet under-discussed inquiries prevalent in the literature: the primary objectives of SBIR and the overarching methodology guiding SBIR. It systematically arranges and scrutinizes the reviewed papers in chronological order, elucidating these critical aspects by offering comprehensive answers to the above questions. The paper also discusses the recent trend of Fine-grained sketch-based image retrieval (FG-SBIR) Zhang, et al. [20] conducted a comprehensive literature survey on recent developments in freehand sketch recognition research encompassing fundamental technologies and benchmark datasets. The survey delves into previous studies and recent progress within the field, categorizing recent research into two primary domains: SBIR and FG-SBIR. The paper provides specific recommendations and methodology evaluation criteria for readers desiring to pursue similar research and discusses promising directions for future research tasks in sketch recognition Ji [21] concentrated on the deep learning-oriented sketch retrieval approach, examining associated research endeavors covering deep feature extraction models, coarse-grained and fine-grained retrieval using deep learning, and category generalization. The paper summarizes the challenges encountered and forecasts potential future research directions Xu, et al. [22] provided a thorough review of the literature on deep learning approaches for free-hand sketch data and the applications that they enable. Through a rigorous taxonomy and experimental evaluation, the survey includes existing datasets, research subjects, and novel approaches. The survey also covers the inherent characteristics and specific limitations of free-hand drawing data, emphasizing the primary contrasts between sketch and pictures. The survey encourages future effort by discussing bottlenecks, outstanding challenges, and potential community research topics. Finally, the study provides TorchSketch, the first open-source deep learning framework built on PyTorch and available for future sketch research and applications. TABLE 2 lists surveys and reviews on SBIR in recent years. In this table, each paper was reviewed, taking into account the publication year, main topic, review types, and limitations.

Examining the above papers, several defects are found. These are all review and survey papers that provide a broad overview, can cover a wide range of research, and may not follow a systematic approach. The article selection process also needs to be clarified and may sometimes introduce personal biases or limitations due to the high level of subjectivity. No SLR on SBIR using deep learning was found. Our research investigates publications published between 2016 and June 2023. PRISMA procedure for SLR is employed to review 90 baseline articles. We have explored SBIR in five aspects: deep learning approaches, datasets, evaluation metrics, application, and future directions of SBIR in deep learning.

III. RESEARCH METHODOLOGY

This systematic review conforms to the PRISMA statement which provides the guidelines for the review in this

TABLE 2. Summary of the related works.

Ref	Year	Main topic	Type	Limitations
[19]	2016	methods or approaches to come up with an efficient retrieval system together with the limitations or challenges	Survey	non-deep learning approach; no discussion about evaluation; unclear selection process
[7]	2018	the objectives of SBIR; the general methodology of SBIR	Survey	no consideration of generalizability, e.g. zero-shot retrieval problem; unclear selection process
[20]	2019	offers an overview of the field's current state, outlines potential research directions, and provides valuable insights for researchers	Survey	method evaluation metrics are not discussed in detail; no consideration of zero-shot retrieval; unclear selection process
[21]	2020	focus on the deep learning-based sketch retrieval method and summarize and prospect the challenges and future research directions	Survey	method evaluation metrics are not discussed in detail; deep learning techniques are reviewed mainly in CNNs and not analyzed regarding the use of RNNs, GANs, etc.; unclear selection process
[22]	2022	presents a comprehensive literature survey of deep learning techniques for free-hand sketch data and the applications	Survey	unclear selection process

paper. PRISMA Statement proposes three distinct advantages: (1) the establishment of research questions (RQs) that allow systematic research, (2) the recognition of inclusion and exclusion criteria for the systematic review, and (3) the attempt to examine an extensive database of scientific literature in a definite time it offers [23].

A. RESEARCH QUESTIONS

The following RQs are developed to guide the systematic review, as shown in TABLE 3.

B. SEARCH STRATEGY

In order to conduct this study, four platform databases were used which are Web of Science, IEEE Xplore, ScienceDirect, and ACM Digital Library. All the database selected is well-known databases that index significant journals and conference papers in the fields of science and technology, as well as based on the analysis of preliminary investigation results. The search was conducted in June 2023 and was limited to studies around 8 years that were published between 2016 to June 2023. Only the English language was utilized, and TABLE 4 showcases the keywords employed for query execution. Keywords within each group are linked using the OR operator, while the groups themselves are connected using the AND operator, forming a comprehensive search query. The final row in TABLE 4 demonstrates the amalgamation of keywords from various groups to create a unified query executed across all four bibliographic databases. Since the ScienceDirect database limits the use of up to 8 logical operators in advanced searches, we deleted the three acronyms CNN, RNN, and GAN for the query, and the rest of the query content was consistent with other databases.

After conducting the search query, 275 papers were obtained, as illustrated in Fig. 3. Subsequently, redundant studies present in multiple databases were identified, and only unique copies were preserved within EndNote for each primary sample. As part of the process to remove duplicate records, 38 studies were excluded.

TABLE 3. Research questions of SLR.

Index	Research Questions
RQ1	What are the existing deep learning approaches in SBIR?
RQ2	What are the major standard datasets for SBIR in the literature?
RQ3	What are the various metrics for evaluating the performance of SBIR?
RQ4	What are the potential applications for SBIR using deep learning?
RQ5	What are the future directions for SBIR using deep learning?

TABLE 4. Selected keywords in the different groups.

Group 1: SBIR related Keywords	"sketch-based image retrieval" OR "sketch-based visual search" OR "sketch-to-image retrieval" OR "sketch-based image matching" OR "image search by sketch"
Group 2: Deep Learning related Keywords	"deep learning" OR "neural networks" OR "CNN" OR "RNN" OR "GAN" OR "generative adversarial network"
Search Query	(Group 1) AND (Group 2)

C. SELECTION CRITERIA AND SELECTION PROCESS

The articles retrieved from the databases were screened in accordance with the PRISMA 2020 guidelines, as depicted in the flowchart and described in detail in Fig. 3. A total of 275 articles were obtained, distributed as follows: 55 from Web of Science, 90 from IEEE Xplore, 63 from Science Direct, and 67 from ACM Digital Library. Results were imported into an EndNote X9.1 library, where duplicate studies were subsequently eliminated. The resultant records were then transferred to Microsoft Excel for further processing.

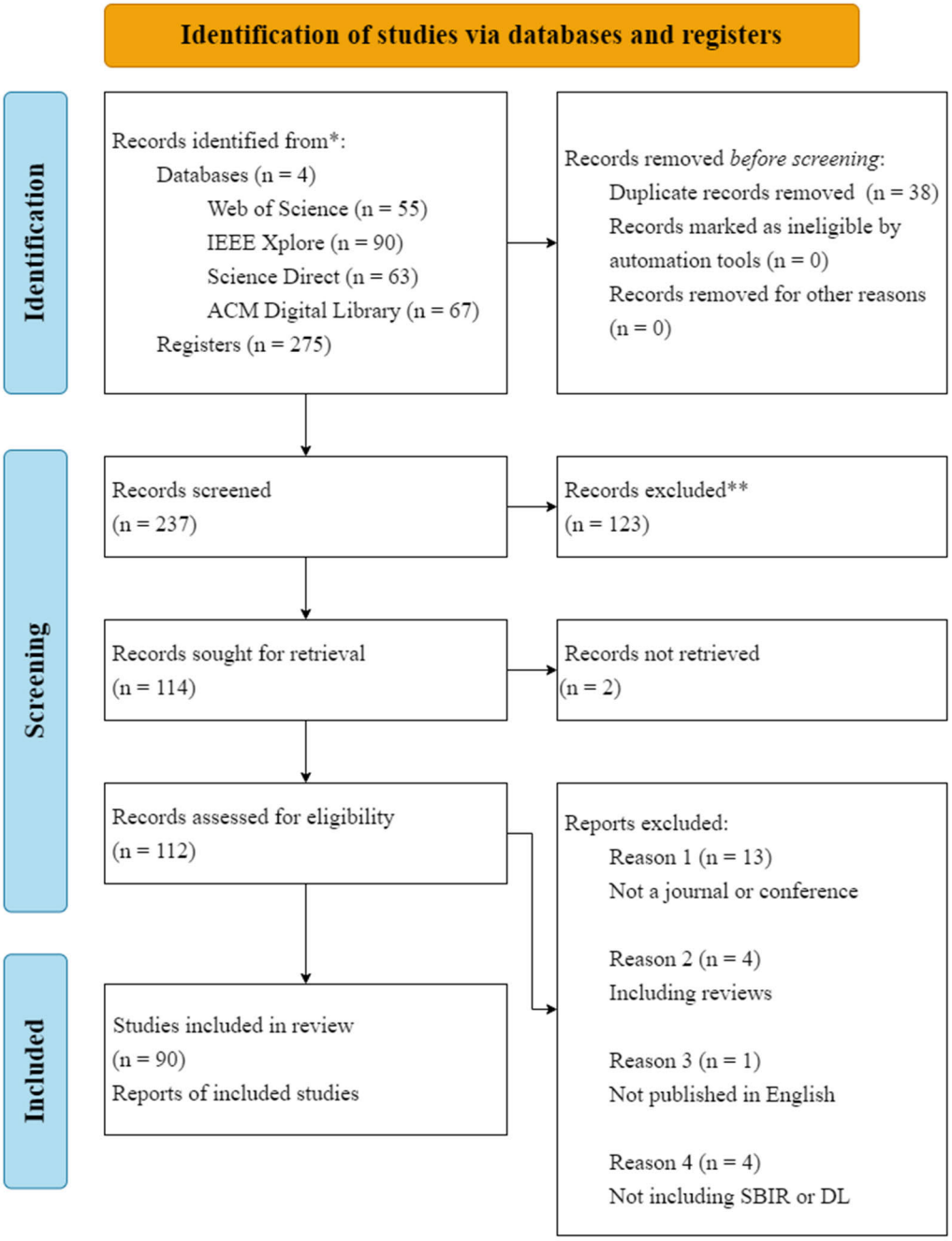

FIGURE 3. The PRISMA 2020 workflow depicting the identification, screening, eligibility, and inclusion process for the SLR.

1) SCREENING (STEP1: SCREENING BY TITLE & ABSTRACT)

After removing 38 duplicates, 237 articles remained. Subsequently, 123 articles were excluded based on title and abstract criteria, leaving 114 studies for full-text analysis.

2) SCREENING (STEP2: SCREENING BY EXCLUSIVE CRITERIA)

We use the following inclusion criteria:

- The article topic must include sketch-based image retrieval.
- The article must be published from 2016 to June 2023.
- The article must be published in a journal or a conference.
- The article must be written in English.

We use the following exclusion criteria:

- The articles that use deep learning techniques but do not address SBIR or study SBIR but are not related to deep learning methods are excluded.
- The articles written in languages other than English are excluded.

We could not obtain [24] and [25] and report them as not eligible, but the abstracts are highly relevant to SBIR. Out of

the remaining 112 papers, 22 were excluded for four specific reasons, while 90 papers were ultimately retained. Thirteen papers were excluded because they were not from journals or conferences (n = 13), and an additional four papers were excluded as they were reviews (n = 4). The third reason was ''the paper was not written in English'' (n = 1), and the fourth reason was ''the paper did not include SBIR or deep learning'' (n = 4).

D. QUALITY VERIFICATION

This section describes how to estimate bias risk. The SLR collected search results from each database in EndNote format. A single individual conducted the selection process utilizing EndNote and Excel tools. To ensure accuracy, all authors performed a double-check of the sorting results for quality assurance.

IV. OVERVIEW OF THE EXISTING APPROACHES

A. DEEP LEARNING

Deep learning, a subset of machine learning, has become the most popular learning technology for large-scale applications. With the rapid development of deep learning theory, there has been an explosive growth in image applications using neural networks. Recently, advanced deep learning techniques like CNN [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], Recurrent Neural Network (RNN) [30], [31], [32], and others have demonstrated remarkable success in image retrieval. The primary advantage lies in these methods' ability to directly learn features from both sketches and images, and end-to-end feature learning is significantly better than shallow features. Some deep learning approaches are given below.

1) CNN

Deep CNNs have shown robust performance across diverse computer vision tasks, surpassing traditional methods in numerous applications. Comprising convolutional, pooling, and fully connected (FC) layers, CNNs efficiently uncover significant data features [33]. Layers close to the input learn low-level general features, while higher layers in the network learn more complex features of the data. The convolutional layer performs feature extraction by combining convolution operations and activation functions. Filters and feature maps are included in the convolution operation. A filter is a collection of weights applied to an input, and a feature map is the output of that filter. Furthermore, the pooling procedure is employed to do down sampling since it aids in the detection of features. The output is then processed via a non-linear activation function as it creates non-linearity in the output. Following the convolutional layer, a FC layer is employed; by adding FC layers, the network can learn the mapping between features and targets. Sketch-a-Net [11] is the first CNN designed for sketches that surpasses the human recognition rate, providing new possibilities for SBIR.

2) RNN

RNN is a neural network model used to process sequence data [34]. It consists of an input layer that receives sequence data, such as a sequence of strokes of a sketch; a hidden layer that captures the temporal dependencies present in the sequence; and an output layer that produces an output based on the hidden state. The emergence of SketchRNN [35] makes the research on sketch representation no longer just static images, but into the vector format of dynamic strokes in time series.

3) GAN

The emergence of GAN can better achieve domain invariance of features, so it is widely used in cross-domain retrieval problems [36]. GAN consists of a generator and a discriminator. In an SBIR task, the generator receives an input sketch or sequence of sketches and attempts to generate an image that matches it. The discriminator is used to distinguish whether the image generated by the generator is a real image. It receives an input image and outputs a probability that represents the probability that the input image is an image in the real data set. The generator fools the discriminator by generating more realistic images, while the discriminator tries to distinguish between generated images and real images more accurately. The introduction of sketch-Gan [37] proves that the representation learned by GAN can be effectively used for sketch retrieval, and also proves that it is more stable in rotation, translation and scaling problems.

4) HASHING

The hash algorithm transforms high-dimensional image data into low-dimensional binary codes, ensuring that comparable images possess similar codes within the hash space. Widely applied for accelerating image and sketch similarity calculations, this algorithm enables rapid image retrieval. Notably, Deep Sketch Hashing (DSH) [38] represents the initial hashing approach employing an end-to-end deep architecture explicitly tailored for category-level SBIR. DSH adeptly captures cross-modal similarities and semantic correlations between distinct categories.

B. DATA COLLECTION

Since sketching is a dynamic process, the sketch storage format can be either a static raster image composed of pixels or a vector image composed of a sequence of strokes. The primary method of data collection is crowd-sourcing drawing [39], and the whole process is time-consuming and labor-intensive, especially for the dataset of the FG-SBIR task. There are also methods of web crawling or online drawing games [35].

During the process of creating a Sketchy database [26], the criteria defined in ''How do humans draw objects?'' [8] were followed during the category selection phase, adding the ''sketchability'' criterion. During the photo selection stage, parts of the content were eliminated based on requirements,

and volunteers then annotated each photo using a subjective ''sketchability'' score. To achieve fine-grained retrieval, the sketch collection process employs a strategy of creating sketches based on specific photos. Specifically, workers were prompted with a photo, then hidden, and then asked to draw a sketch corresponding to the image from memory. Each sketch is stored in SVG format with timing details.

When the Queen Mary University of London (QMUL) shoe and QMUL chair database [10] collects natural images, they are selected from UT-Zap50K and IKEA, Amazon, and Taobao shopping websites to cover different types and styles of images as much as possible. Sketches are collected by volunteers drawing sketches of the presented images on a blank canvas on a tablet. During data annotation, a subset of triples is selected, and a human annotator annotates attributes, generates candidate photos and triplet annotations for the sketch.

The user interface sketch dataset [12] was specifically crafted to facilitate the advancement of sketch-based data-driven design applications. This collection encompasses sketches akin in style and semantic nature to those generated by UI designers. Due to the absence of extensive public datasets featuring UI sketches alongside corresponding UI screenshots, the dataset's creation engaged four designers via Upwork. The dataset comprises sketches, each completed in an average time of 4.1 minutes, meticulously selected to represent 23 app categories found in the Google Play Store, demonstrating high levels of design quality.

C. EVALUATION

To determine the performance and effectiveness of the SBIR, it needs to be evaluated. Evaluation methods are usually categorized as quantitative and qualitative.

Quantitative evaluation: use metrics such as mAP, Recall, and Precision for sketch retrieval tasks to compare the state-of-the-art methods and quantify the effectiveness of this SBIR method.

Qualitative Evaluation: during SBIR research by visually comparing the retrieved images with real images or analyzing the discriminative and authenticity of the generated image features, highlighting the effectiveness of the proposed method in terms of better retrieval results. It can also be done by collecting subjective feedback from the users, including their satisfaction, usability, and preference for the retrieval results. Qualitative analysis helps to understand the strengths and limitations of the method.

Qualitative and quantitative evaluations usually complement each other, with quantitative evaluations providing objective performance measures and qualitative evaluations providing insights into user experience and application scenarios. To better evaluate sketch retrieval methods and applications from the perspective of professional users, some studies also employ expert evaluation. For example, by inviting some users with relevant industry experience to conduct actual evaluations and asking them to provide comments on the retrieval process and results [12].

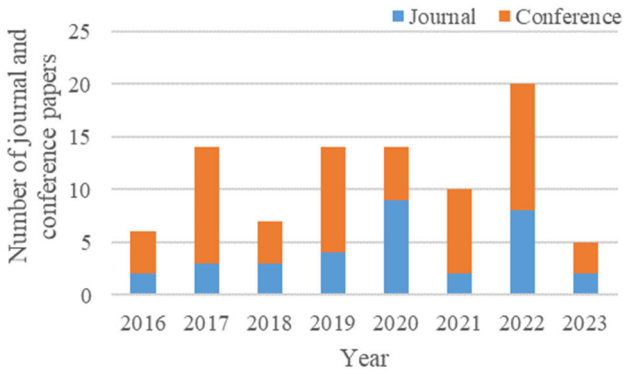

FIGURE 4. The accumulation of conference and journal papers from 2016 to June 2023.

V. RESULTS AND DISCUSSION

Ninety compliant articles were finally obtained according to the steps in Fig. 3. The articles selected for review were analyzed, summarized, and discussed. Categorized into conferences and journals based on the type of article, and Fig. 4 shows the number of conference and journal articles published between 2016 and June 2023. Research has been increasing using deep learning for SBIR tasks in recent years.

A. RQ1: WHAT ARE THE EXISTING DEEP LEARNING APPROACHES IN SBIR?

For the RQ1, TABLE 5 summarizes the deep learning methods used for SBIR. Based on the difficulty and complexity of the retrieval task, SBIR can be broadly categorized into coarse-grained retrieval, fine-grained retrieval and zero-shot retrieval. Coarse-grained SBIR, which can also be referred to as category-level SBIR, is given a sketch as a query, finds images similar to it based on the features of the sketch and returns a ranked list of images. If the photos in the ranked list have the same category label as the query, the retrieval result is correct. However, in FG-SBIR, also known as instance-level SBIR, the retrieval accuracy hinges on the correctness of the returned image result, requiring it to match the query sketch's specific instance pair. Zero-shot retrieval requires the system to perform image retrieval without a training sample. This means that the system should be able to understand an unseen sketch when seeing it and find images similar to it from a library of unseen images. It is useful for real-time scenarios in practical applications where it is difficult to train comprehensively for all possible sketch styles. TABLE 5 shows that CNN is the most commonly used deep learning method for SBIR, whether it is coarse-grained retrieval, fine-grained retrieval, or zero-shot retrieval. In addition to the use of CNNs, the study also extensively used them in conjunction with RNNs, GANs, attention mechanisms, hashing algorithms, and using multimodal fusion methods. The study found that CNNs are most widely used when it comes to coarse-grained retrieval. At fine-grained retrieval, CNNs remain the most popular, while the performance of CNNs combined with attention mechanisms is favored.

TABLE 5. Summary of deep learning methods used for SBIR.

SBIR	Model	Papers
Coarse-grained SBIR	CNN	[9], [12], [17], [18], [27], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52]
	RNN+CNN	[32], [53], [54]
	GAN+CNN	[55], [56], [57], [58], [59]
	CNN+Attention	[60]
	Hashing	[38], [61]
	Others	[29], [62], [63], [64]
Fine-grained SBIR	ANN	[65]
	CNN	[10], [12], [18], [26], [27], [28], [48], [51], [66], [67], [68], [69], [70], [71], [72], [73]
	RNN+CNN	[30], [31], [32], [74], [75], [76]
	GAN+CNN	[77], [78], [79]
	CNN+Attention	[80], [81], [82], [83], [84], [85], [86]
	CNN+Transformer	[87], [88]
	GNN+Transformer	[89]
	Transformer	[90], [91]
	Others	[92], [93], [94], [95]
Zero-shot SBIR	CNN	[96], [97], [98], [99]
	RNN+CNN+Hashing	[100]
	GAN+CNN	[101], [102], [103], [104]
	GAN+Auto-Encoder	[105]
	CNN+Attention	[106]
	CNN+Attention+GCN+Hashing	[107]
	Transformer	[108], [109], [110]
	Knowledge Distillation	[111], [112], [113]
	VAE	[114]
	Others	[115]

In zero-shot retrieval, the advantages of GAN are exploited, and many advanced deep learning models such as Transformer, Variational Autoencoder (VAE) are also introduced to solve the ZS-SBIR task. Most of the SBIR methods include data preprocessing, and many studies have used CNNs in conjunction with re-ranking of user feedback in others methods.

1) APPROACHES USED IN COARSE-GRAINED SBIR

TABLE 5 and Fig. 5 show that one of the most commonly used deep learning methods for SBIR in coarse-grained retrieval is the CNN. It is widely used for its special ability to find important features in sketch image data Qi et al. [17] first applied CNNs to category-level SBIR by proposing a CNN based on Siamese networks. Two CNNs connected by a loss function pull the output feature vectors of similar input sketch image pairs closer together and push dissimilar sketch image pairs further apart Xinggang et al. [41] used sketches and natural images to co-train CNNs, prior to which a specific image scaling method and a multi-angle voting scheme were designed for image data to be used together for SBIR Bui et al. [18] proposed a triplet ranked CNN for SBIR to learn embeddings between sketches and images with significantly improved performance. Experiments exploring different sharing levels between sketch and image edge graph branches show that partially shared weights have superior performance over fully shared and no sharing. The index storage is also reduced from 3200 bits to 56 bits using PCA and dimensional quantization, achieving a compact representation with accuracy. The excellent performance of triplet networks has played an important role in subsequent research, and many works have built on it to explore SBIR even further Seddati et al. [49] proposed quadruplet networks based on triplet networks, which make full use of global and local information to provide better embedding of data Yan et al. [27] used CNN models combined with classification loss and histogram loss, and their experimental results show that the joint loss learns cross-domain embedded features better than a single loss function. Inspired by triplet ranking [18], Bui et al. [44] learns joint embeddings using both contrast loss and triplet loss, compares different weight sharing, downscaling, and training data preprocessing strategies, and investigates cross-class generalization capabilities. The triplet architecture with GoogleNet branching structure with partially shared weights performs best Deng et al. [40] utilized the multi-layer attribute framework to derive profound semantic features from images, intending to bridge the domain gap between sketches and natural images. Their devised multi-layer deep neural network not only extracts multi-layer features from sketches but also captures binary edge maps from natural images. Merely employing these multi-layered visual representations of both sketches and natural images yields remarkable retrieval outcomes Ahmad et al. [9] conducted CNN fine-tuning employing an image augmentation dataset inclusive of natural images, edge maps, hand-drawn sketches, as well as decolorized and de-textured images. Their findings illustrate a notable enhancement in retrieval performance upon integrating color information into sketches. Moreover, the extent of this improvement is directly correlated with the quantity of added color information Song et al. [46] introduced an edge-guided cross-domain learning network aimed at minimizing the domain gap. Within a triple network, they incorporated edge map information to enhance the SBIR task. This was achieved through the introduction of an edge guidance module designed to fuse natural images with their corresponding edge maps. Additionally, they implemented a shape regression module, which delved into exploring the shape similarities existing between sketches and natural images Huang et al. [12] presented Swire, a sketch-based UI retrieval approach empowering designers to interact with extensive UI datasets using sketches. Swire incorporates a neural network structured into two identical sub-networks, reminiscent of the VGG-A deep convolutional neural network. These distinct networks, featuring diverse weights, aim to encode matching pairs of screenshots and sketches utilizing similar values. Zhang et al. [47] proposed a hybrid CNN structure consisting

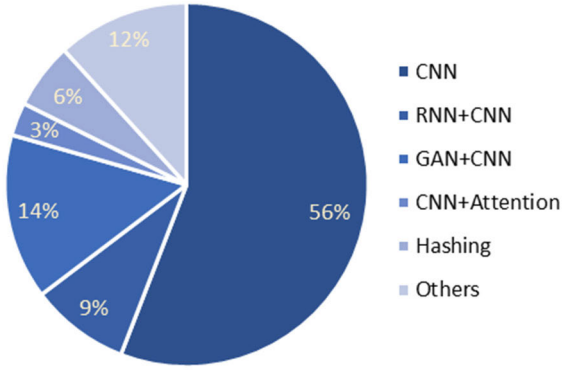

FIGURE 5. Distribution of coarse-grained SBIR approaches.

of A-Net and S-Net, with different branches dealing with appearance information and shape information respectively. Given that triplet networks are so popular in SBIR tasks, Seddati et al. [52] enhanced the SBIR pipeline performance by exploring multiple aspects, including embedding normalization, model sharing, margin selection, batch size, hard mining selection, and the number of hard triplets utilized during training. Furthermore, they introduced an innovative methodology for constructing SBIR solutions adaptable for deployment on low-power systems.

Inspired by the success of CNN architectures on a variety of computer vision tasks, researchers have begun to explore whether sketch image data can be similarly applied to these models. Many of these pre-trained models are publicly available, so some studies have conducted experiments using them as initial networks. Lei et al. [42] used the VGG19 network pre-trained on image data as the initialization network. Use generalized boundary extraction methods to generate sketch approximations. The sketch approximation data is then fed into the pretrained network for fine-tuning. The results show that using image data as auxiliary data can improve the ability of sketch feature extraction. The extracted sketch features are strongly correlated with the boundaries of the corresponding images Kumar et al. [45] employed the Siamese CNN in SBIR to acquire the feature space for both sketch input queries and images using transfer learning techniques. They utilized the pre-trained VGG19 model to extract feature descriptors from sketches and images Devis et al. [50] used the pre-trained VGG19 network and the cosine similarity function to find similarities between sketches and image databases, further demonstrating the effective use of transfer learning in SBIR.

There is a certain time order when drawing a sketch, so the data recorded in the form of stroke order in the SBIR problem is the key information for sketch retrieval. RNN models the time series by linking previous stroke information and the current sketch state. Long Short-Term Memory (LSTM) is a special type of RNN that can solve the long-term dependency problem He et al. [53] introduced the Deep Visual-Sequential Fusion model (DVSF), a sketch recognition framework combining CNN and RNN. They employed three-way CNNs to extract visual features, feeding them into a visual module comprising stacked residual FC layers. Subsequently, these features were passed into the sequential module, which comprised a specialized RNN known as Residual LSTM, culminating in the final prediction Collomosse et al. [32] proposed LiveSketch, which makes the query a dynamic iterative process through visual suggestions, reducing the ambiguous nature of the sketch itself. The research approach uses a triplet convolutional network architecture, combined with an RNN-based VAE, using sketch stroke vector queries to search target images, and real-time clustering to produce possible results. Back-propagation through possible results disrupts the stroke sequence input and guides continuous query improvement. The subsequent Gated Recurrent Unit (GRU) generally outperforms LSTM by learning a reduced number of parameters. Jia et al. [54] introduced a Sequential Double Recurrent Neural Network (SD-RNN) architecture for sketch recognition, leveraging this advantage. The sketched shapes and textures of the five images constructed from accumulated strokes are fed to two cascaded GRUs, which are classified based on the output.

Besides its remarkable advancements in image generation research, the adversarial learning principles of GAN offer innovative solutions across various image-related research domains Guo et al. [55] utilized the conditional generative adversarial network (cGAN) to produce synthetic images from input sketches. They employed the VGG encoder to process both synthetic and real images, culminating in the development of the interactive SBIR system known as MindReader Sharma et al. [56] implemented the task of querying floor plan image retrieval through hand-drawn sketches through recurrent GAN and CNN Sabry et al. [58] based on the features obtained by the unsupervised learning information maximization GAN (InfoGAN) model to meet the user's needs for retrieval of large-scale data sets Xu et al. [57] employed hand-drawn sketches to address remote sensing image search challenges. They devised a novel Sketch-Based Remote Sensing Image Retrieval (SBRSIR) model that learned a deep joint embedding space, integrating discriminative loss to foster domain-invariant representations through adversarial training. Additionally, they contributed sketch and remote sensing image datasets to SBRSIR and established a baseline for future researchers in the field Bai et al. [59] employed deep neural networks to extract features, utilizing the same network structure without sharing parameters. They augmented these networks with domain classifiers to create an adversarial network. The mutual adversarial training between domain classifiers and feature extractors facilitated the learning of a shared feature space suitable for both sketches and natural images. To streamline the network with fewer parameters while ensuring accurate feature extraction, the study utilized the lightweight MobileNet model for extracting features from different modality data.

In order to solve the problem of high computational effort and parameter redundancy, Lu et al. [60] proposed

Domain-Aware Squeeze-and-Excitation (DASE) network that can emphasize different channels based on domain knowledge. With the criterion that the maximum intra-class distance is less than the minimum inter-class distance, Multiplicative Euclidean Margin Softmax (MEMS) introduces the margin to optimize the feature space.

Hash technology maps high-dimensional image features into low-dimensional binary codes, so that similar images have similar codes in the hash space, achieving fast similarity retrieval. As the number of digital images continues to grow, resulting in some retrieval time and efficiency issues, hashing algorithms have attracted attention in the SBIR research field Liu et al. [38] introduced DSH, a deep hashing framework designed for rapid SBIR. Their research focused on encoding sketches and natural images using a semi-heterogeneous deep architecture featuring auxiliary sketch-tokens. This approach effectively mitigates geometric distortion between the two modalities. DSH demonstrates advantages in retrieval accuracy as well as time and storage complexity Wu and Xiao [61] proposed a deep hashing framework that uses a prototype hash code set to constrain feature representation, and maps data from two different domains of sketches and natural images to a common Hamming space to achieve good retrieval performance.

Some studies use deep learning techniques in addition to a combination of traditional methods of SBIR, multimodal information fusion, and re-ranking Huang et al. [62] used low-level visual features and convolutional kernel network (CKN) to capture local visual information, CNN captured semantic information of sketches and images, and finally obtained deep discriminative representation through multi-modal feature fusion to narrow the visual and semantic gap in SBIR Huang et al. [64] introduced a deep cross-modal correlation learning technique aimed at investigating the correlation between visual and semantic features within sketches and images, thereby mapping them into a shared space. They combined various methods to construct a real-time SBIR framework that involves mining multi-modal attributes. Typically, SBIR re-ranking methods employ relevance feedback for re-ranking. Relevant information is gathered from the initial retrieval outcomes, forming the foundation for subsequent re-ranking processes Huang et al. [63] introduced an SBIR re-ranking optimization technique utilizing sketch-like transformations and deep feature representation via deep visual semantic descriptors. They employed an efficient deep visual semantic descriptor capable of encoding both low-level and high-level features of sketches and sketch-like images derived from original natural images. Their approach involved constructing a clustering-based re-ranking optimization method to dynamically adjust image similarity within the ranking list. However, a drawback of this method lies in its limited consideration of the abundant semantic information contained within annotated photos Wang et al. [29] captured semantic information from both types of data by training CNNs separately for sketch classification and natural image classification. They employed a proposed category similarity measure to re-rank preliminary test results, aiming to improve system performance. The combination of SBIR training and reranking improves the results by further refining the initially learned embedding space.

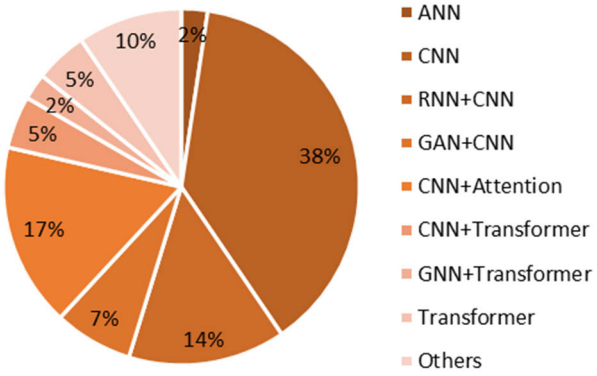

FIGURE 6. Distribution of FG-SBIR approaches.

2) APPROACHES USED IN FINE-GRAINED SBIR

Sketches are more likely to convey fine-grained information compared to text and tags. FG-SBIR is highly regarded for its value in business applications, but is more complex and therefore more challenging than the need for category-level retrieval. The distribution of various approaches is shown in Fig. 6.

Shi et al. [65] introduced an image retrieval method termed FAIR, employing Freak and ANN-based techniques. Image features were extracted and transformed into binary descriptors, mimicking human retina recognition. Artificial neural networks were utilized to handle extensive data for feature classification. While FAIR proved effective in accuracy, its time efficiency lacked objectivity.

Addressing the problems of cross-domain retrieval, highly abstract sketches, and scarcity of training data in FG-SBIR, Yu et al. [10] solved these problems for the first time. They introduced a database containing two categories of shoes and chairs along with rich annotation information, developed a deep triplet model for fine-grained retrieval, investigated novel data augmentation strategies, and contributed to the realization of SBIR for commercial applications Portenier et al. [67] proposed an interactive image retrieval system reordered by sketching and semantic clustering, using the power of CNNs to refine the query results, and ultimately designing a user study to demonstrate the effectiveness of the approach. Considering the limitations of digital manga retrieval, [69] constructed an interactive manga retrieval system based on sketches by extracting features of sketches and manga images using two differently trained CNNs Xia et al. [28] utilize the same and homogeneous three-branch Triplet network and implement a ranking method that accounts for both shape and color matching to accomplish fine-grained retrieval. The use of edge maps in instance-level SBIR systems poses challenges,

requiring pre-training of significant edge map data and sensitivity to edge map quality. To overcome these limitations, Lin et al. [71] propose an end-to-end iSBIR system called TC-Net, which incorporates a triplet Siamese network and an auxiliary classification loss that eliminates the need for edge maps Qi et al. [72] proposed a personalized SBIR approach that incorporates a deep full convolutional neural network as a general model and a personalized model using transfer learning to achieve fine-grained image semantic features. The personalized model is trained based on user-selected images and user history feedback, allowing the system to learn the user's intent and provide more optimal search results. Zhang et al. [51] introduced a groundbreaking weakly-supervised deep architecture named the landmarks-aware network. This network comprises two primary modules: the representative landmarks discovering module and the category-aware representation learning module. Through extensive SBIR task experiments, the paper validates the effectiveness of the proposed method. In Black et al. [66] research they match the relative positions of multiple objects in addition to matching the main object contained in the image. The visual features are encoded by training a CNN that aggregates these features into spatial descriptors, which in turn encode spatial relationships and appearance. Focusing on exploring the use of local features to solve the problem of fine-grained retrieval, Xu et al. [73] introduced the Local Aligned Network (LANet) to address the FG-SBIR problem by identifying matching pairs of sketches and photographs with shared fine-grained details. Additionally, to resolve spatial misalignment issues between sketches and photographs, they proposed the Dynamic Local Aligned Network (DLA-Net) and achieved superior performance compared to humans in QMUL FG-SBIR, QMUL Handbag, and Sketchy datasets.

Muhammad et al. [30] developed a model based on RNN to determine stroke retention or deletion, trained a sketch abstraction model by reinforcement learning (RL), fed images and generated edge maps into the abstraction model but trained FG-SBIR only for a given photo, and used three-branch CNN learning for comparing joint embeddings of the photos and sketches. Experiments demonstrate that FG-SBIR models can now be trained using only photos. Noting the important impact of sketch integrity and sketch-drawing quality on SBIR systems, Choi et al. [31] proposed the first stroke-guided SBIR system. The system's primary stroke guidance network comprises a Siamese CNN and an LSTM network. Employing a stroke loss function enables comparison between the next stroke feature vectors of the CNN and LSTM networks. Through deep binary hash retrieval, the system effectively navigates extensive databases and promptly provides target guidance sketches for stroke guidance stages in real-time Wang, et al. [74] designed a framework that employs deep reinforcement attentional regression to support on-the-fly image-based retrieval. A hybrid loss function was developed which will be used to train RL agents for dynamic ranking rewards and for supervised learning for updating RNNs. Based on this technique, Mohian and Csallner [75] present PSDoodle, the first system that allows on-screen interactive searching through interactive sketches. it achieves similar search accuracy to SWIRE, with a much reduced search time.

Yang et al. [77] incorporated the CNN model into the edge branch of DA-SBIR. They integrated dark region information into FG-SBIR and employed the SGAN module to efficiently distinguish between the edge structure and dark region information. This approach aimed to enhance retrieval accuracy without compromising convenience Zhou et al. [78] effectively narrowed the gap in the sketch image domain by using GAN for stroke coordination, CNN for feature extraction and fine-tuning, and aligning the sketch and natural image to the same stroke style domain Zhang et al. [79] proposed deformable triplet CNNs combining depth features and attribute features using the generation of pseudo-sketches instead of the traditional method in the preprocessing step.

According to visual habits, people do not directly process the global scene, but preferentially capture local salient parts. Attention models have been widely studied in various vision problems [116], [117], [118], [119]. The most prevalent attention model utilized is soft attention, which, due to its differentiable nature, can be seamlessly learned alongside the rest of the network in an end-to-end fashion. Typically, soft attention models are designed to learn attention masks assigning varying weights to distinct regions of the image. Hard attention models only indicate one region at a time and are usually learned using reinforcement learning because they are non-differentiable Song et al. [83] add an attention model to each branch based on multi-branch CNN, focusing on specific local discriminative information. Using a shortcut connection architecture to feed details into the attention module, a higher-order learnable energy function (HOLEF) is introduced to model correlation, which in turn enhances robustness Shaojun et al. [81] mainly solve the Sketch Re-ID problem, and in order to validate the generalization ability of the model, experiments are also performed on SBIR related datasets, where the framework employs CNN combined with a spatial attention module using a gradient reverse layer to reduce the domain gap Yu et al. [80] designed the attention module for FG-SBIR based on the Siamese network, and extensive experiments on QMUL FG-SBIR database proved the effectiveness of the model Chaudhuri et al. [84] solved the problem of cross-modal retrieval of photo-sketched remote sensing (RS) data via CNN combined with cross-attention networks Dawei et al. [82] proposed a multi-granularity association-learning method for FG-SBIR to bring the joint embedding space representation of incomplete and complete sketches closer for dynamic retrieval.

Self-attention, a prominent attention mechanism, gained recognition through the Transformer proposal [120], which has found applications not only in natural language processing [121], [122] but also in computer vision. In image-related data, self-attention contributes to improving neural network feature extraction capabilities. Chen et al. [86] highlighted

the importance of channel context and spatial sequence information by strengthening both channel attention and spatial attention modules. They integrated a Transformer into their model to enhance its ability to comprehend spatial sequence data and introduced Mutual Loss to bolster intra-class discrimination.

Ribeiro et al. [89] introduced the Scene Designer, a model utilizing hand-drawn sketches to both search for and generate the appearance and relative positions of multiple objects. This approach involves learning embeddings through a hybrid combination of graph neural network (GNN) and Transformer architecture. These embeddings capture the entirety of the scene layout from either complete or partial sketch compositions. Future exploration could focus on deploying this interactive model within creative practice.

Chen et al. [90] proposed an asymmetrical disentanglement scheme and a dynamic updatable auxiliary sketch (A-sketch) modality for photo disentanglement to solve the asymmetry between sketch and image modalities. Problem. The proposed visual converter-based method achieves better performance than CNN.

In their proposition of the Multi-Level Region Matching model (MLRM) [87], two integral components contribute to its architecture. The initial segment involves the Discriminative Region Extraction module (DRE), responsible for extracting multi-level CNN-based features. The subsequent segment, the Region and Level Attention module (RLA) utilizes a transformer encoder to assign distinct weights to the transformed multi-level region features. The method's effectiveness is validated across a combined dataset comprising five different datasets.

There have also been some studies at fine-grained retrieval focusing on the impact of multimodal information and re-ranking models on SBIR retrieval Yanfei et al. [92] proposed Deep Cascaded Cross-modal Ranking Model (DCCRM) constructs representations of sketches, images and annotations to learn deep associations between multimodal information. Image visual features are extracted using pre-trained and fine-tuned GoogLeNet, image semantic features are described using Skip-thought, multimodal information is co-learned in Deep Multimodal Embedding Model (DMEM), and Deep Triplet Ranking Model (DTRM) is utilized to improve retrieval efficiency. Bhattacharjee et al. [93] proposed a graph-based re-ranking method using Sketch-a-Net and CNN to obtain shape and appearance features while considering the query as a sub-graph selection problem Pang et al. [94] employed an unsupervised embedding network and a dynamically parameterized feature extractor equipped with triplet loss for FG-SBIR. The unsupervised embedding network employs an encoder-decoder framework to map sketches into visual trait descriptors. Meanwhile, the dynamically parameterized feature extractor adapts the feature extraction and retrieval processes based on the generated descriptor. Deep CNN feature extractors are used in the FG-SBIR model to find the photo that minimizes the distance to the query sketch, and the model is trained in a supervised way on the training sketch categories Sabry et al. [95] tackled the challenges in Facial Sketched-Real Image Retrieval (FSRIR) and bridged the research gap in content-based similarity matching and retrieval. They expanded the Chinese University Face Sketch (CUFS) dataset, introduced the Extended Sketched-Real Image Retrieval (ESRIR) dataset, and introduced three novel systems for sketched-real image retrieval. These systems are based on convolutional autoencoder, InfoGAN, and Vision Transformer (ViT) unsupervised models tailored for large dataset.

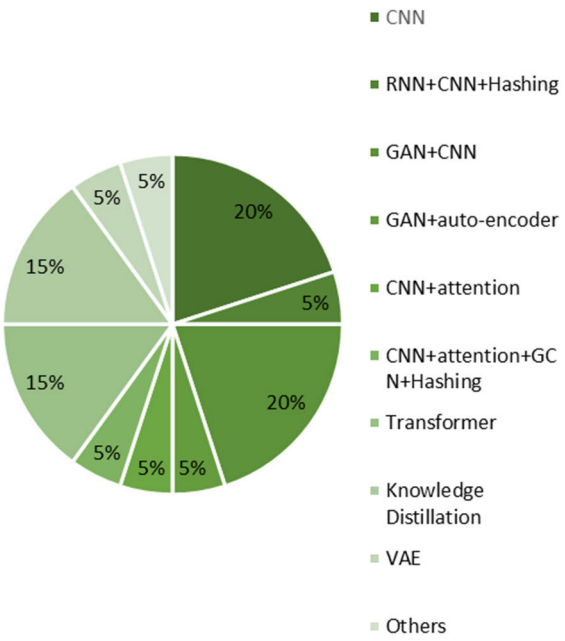

FIGURE 7. Distribution of ZS-SBIR approaches.

3) APPROACHES USED IN ZERO-SHOT SBIR

Zero-shot learning in computer vision refers to recognizing objects for which no examples have been seen during the training phase. ZS-SBIR addresses the limitation of traditional SBIR methods that rely on a large amount of sample data with the same category annotations. It can retrieve relevant images from a gallery without any example or training data. In recent years, many methods for ZS-SBIR have been proposed to address the modality gap and semantic gap problems. The distribution of various ZS-SBIR approaches is shown in Fig. 7.

In the CNN model-based approach, Tursun et al. [96] proposed a framework that includes three kinds of losses. The first one is domain-balanced quadruplet loss, which solves the imbalance problem in the underlying triplet loss. The second is the semantic classification loss which is used to learn semantic features. The third is the semantic knowledge preservation loss which preserves the knowledge learnt from the pre-trained model used. Experiments have shown that the introduction of each loss has led to new enhancements. Based on intermediate and advanced CNN feature embeddings,

Chaudhuri et al. [97] proposed BDA-SketRet that performs two-layer domain adaptation to progressively align spatial and semantic features of visual data pairs.

Xu et al. [100] introduced a two-branch CNN-RNN network architecture, capturing both the static and temporal patterns of sketch strokes. Their experiments investigate the acquisition of sketch-oriented semantic representations in two distinct settings: hash retrieval and zero-shot recognition. These evaluations extend to millions of abstract sketches generated from real-time online interactions.

Modelling the joint distribution between sketch and image using GAN can significantly reduce the domain gap between sketches and images [101]. Continuing previous SBIR research, [102] [103] simplifies the SBIR problem to an image-to-image retrieval problem by synthesizing image samples from sketch features. From only focusing on one-side mapping of multimodal features to class embeddings to emphasizing the correlation between them, Xu, et al. [104] combined multimodal feature synthesis, knowledge transfer and common subspace learning to successfully transfer knowledge to unseen classes Dutta and Akata [105] focuses on the any-shot SBIR task, which combines zero-shot and few-shot learning and aims to accurately predict the class of a query and retrieve relevant images. The paper introduces an innovative approach called SEM-PCYC, a semantically aligned paired cycle-consistent generative adversarial network. This model effectively maps sketches and images into a shared semantic space, enhancing their alignment and ensuring greater consistency between the two. The proposed model achieves state-of-the-art performance on three popular datasets for any-shot SBIR and introduces the few-shot setting for SBIR, which enables the model to handle queries with very few examples.

In addition to using GANs for the ZS-SBIR task, Graph Convolutional Networks can also be utilized to pair sketches and images in their shared semantic space. The ZS-SBIR framework proposed in [107] introduced a cross-modal attentional learning strategy based on cross-modal ternary loss, while using GCN to propagate semantic information into the shared space and Hashing to shorten retrieval time.

CNNs, limited by their local convolutional operations, face constraints in modeling the global structural information of objects. Yet, this global structural information holds significance in the ZS-SBIR task, as it serves as a critical indicator for correlating images and sketches. Transformer is capable of modeling global structural information, and it has emerged as an effective alternative to the CNN framework. Tian et al. [108] model the global context using a global self-attention mechanism and propose a Transformer-based approach called Three-Way Vision Transformer (TVT). Fusion and distillation tokens are also added so that they complement each other. Inter-modal and intra-modal alignment is performed without loss of uniformity to learn multimodal hyperspheres to bridge the modal gap between sketch and image modalities. Utilizing ViT's capacity for global structure modeling in handling generic cross-domain retrieval tasks, Tian et al. [109] introduced the Structure-Aware Semantic-Aligned Network (SASA). This novel approach aims to retain global structural information and accomplish cross-domain alignment of multi-source data using an efficient domain-biased sampling strategy Song et al. [110] proposed a framework called Local Feature Contrastive Network (LFCN) that utilized transformers to obtain similarity scores of local features in cross-modal image retrieval scenarios in order to narrow the domain gap. It is also demonstrated that the potential risk of overfitting can be reduced by transferring limited knowledge to unseen categories and through assisted learning.

Existing approaches propose various ways to address the modal gap, e.g., using category language information, generative architectures. However, these approaches still do not adequately consider distinguishability and generalizability. There are also works that improve the generalization capability in the direction of knowledge distillation Tian et al. [111] proposed Relationship-Preserving Knowledge Distillation (RPKD) for ZS-SBIR. It preserves instance-level inter-class relations without semantic similarity, contrast relation preservation and local relation preservation for teacher networks that mimic rich knowledge Wang et al. [112] propose Prototype-based Selective Knowledge Distillation (PSKD). For teacher modelling, PSKD utilizes variants of the Transformer module to capture contextual information from images and sketches with a multi-tailoring strategy; and category prototyping to mitigate domain gaps. Teacher-student optimization through a combination of these designs Tian et al. [113] propose a new method called Adaptive Balanced Discriminability and Generalizability (ABDG). ABDG utilizes a two-stage knowledge distillation scheme, a task-specific teacher model, and an entropy-based weighting strategy to balance the learning of discriminability and generalizability for each instance. The aim is to achieve state-of-the-art performance in ZS-SBIR by considering both discriminative and generalization properties.

The answer to RQ1 shows that the use of deep learning methods in SBIR has continued to increase over the years. For SBIR tasks, no matter what network structure is used, since sketches and natural images belong to two different domains, at least two network branches are usually needed to process different types of data in order to better bridge the domain gap, which is the reason why many methods are based on Siamese, Triplet, and Quadruplet-ranking methods. The current new trend is to use multimodal data for retrieval in the study of SBIR. However, the number of articles on SBIR using multimodal data and deep learning methods is still relatively small. Although scientific research methods and reasonable retrieval strategies are used whenever possible, it is still possible to miss some valuable research.

B. RQ2: WHAT ARE THE MAJOR STANDARD DATASETS FOR SBIR IN THE LITERATURE?

In order to meet the needs of large-scale deep network training in the SBIR study, RQ2 answered the datasets used in

the study and further analyzed the details of the data using in various types of SBIR tasks. TABLE 6 summarizes the name, scale, number of categories, and public availability of the sketch dataset.

As can be seen from TABLE 6 and Fig. 8, Sketchy and Flickr 15k are widely used by researchers because they contain a wide variety of sketches and corresponding natural images, and they are open to the public and serve as benchmarks for many studies. The CNN model processes the sketch as a static raster image, and Sketchy is often used for model training and testing. The QuickDraw dataset contains considerable sketch data, which is crowdsourced by different users and can be publicly accessed. The data is a sequence of strokes consisting of coordinates and timestamps. Research methods using RNN models often use the QuickDraw data set because they focus on the stroke sequence of sketches. The QMUL dataset is widely used for its high-quality images, including QMUL Shoe, QMUL-ShoeV2, QMUL Chair, QMUL ChairV2 and QMUL Handbag datasets, etc. They come with detailed annotations, including attribute labels, which are valuable for training and evaluating fine-grained retrieval models. Therefore, methods used for FG-SBIR, especially those combined with attention mechanisms, usually use the QMUL series of datasets. The latest research direction, ZS-SBIR, often conducts experiments on the benchmark Sketchy extension, TU Berlin extension, and QuickDraw extension dataset to verify the effectiveness of the method.

C. RQ3: WHAT ARE THE VARIOUS METRICS FOR EVALUATING THE PERFORMANCE OF SBIR?

SBIR involves retrieving images from sketches drawn by users that match their intent. In this case, in order to more comprehensively evaluate the performance and effect of the algorithm, most studies adopt a combination of quantitative and qualitative evaluation. Quantitative evaluation measures system performance through digital indicators and quantitative methods. As depicted in TABLE 7, many studies utilize common quantitative evaluation metrics such as mAP, Recall@K and Precision@K. In the case of query sketches, any image among the initial K retrieved images sharing the same category as the query sketch qualifies as a matching result [92]. Recall@K indicates the percentage of relevant images retrieved among all relevant images within the top K positions and is used to evaluate the probability of detecting a positive sample correctly. It is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

where TP is the true positive sample, and FN is the false negative sample, which is the number of true positive samples assigned with false labels. Precision@K represents the percentage of relevant images among those retrieved within the top K positions, measures how accurately the system finds positive examples, and is defined as:

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

where TP stands for True Positive, and FP stands for False Positive. mAP is a comprehensive metric that considers both precision and recall aspects. This is more in line with practical needs. Users have limited energy to view search results, and the system can put the most relevant images at the front. mAP can be calculated in the following way:

$$mAP = \frac{\sum_{q=1}^{S} Avg\,(Pr\,(q))}{S} \tag{5}$$

where $Pr(q)$ denotes the retrieval accuracy of the query sketch q, $Avg()$ denotes the average, and S denotes the number of query sketches.

Some CNNs for category-level retrieval also prefer to quantify the Kendal score by comparing the ranked lists obtained by the proposed new method with the rankings of the users identified as the ''ground-truth''. The Kendal rank correlation coefficient for SBIR is defined as (6):

$$\tau = \frac{n_c - n_d}{[(N - U)\,(N - V)]^{\frac{1}{2}}} \tag{6}$$

where n_c denotes the number of consistent, n_d denotes the number of inconsistent, N denotes the number of possible pairs in the set of different elements, U denotes the list of the number of tied pairs in the baseline ranking, and V denotes the number of tied pairs in the ranked list of the proposed method.

When SBIR is combined with Hashing, the memory loads and retrieval time of the system are often evaluated. Because hashing methods map high-dimensional feature vectors into compact binary codes, searches for similar items in large databases can be made faster, running efficiently without performance degradation or crashes due to excessive memory consumption. Therefore, evaluating retrieval time and memory load is critical to evaluating the efficiency gains achieved through hashing. Quantitative evaluation helps provide specific performance data to quantify the performance of the methods involved in TABLE 5 on different datasets.

Qualitative evaluation is based on subjective judgement and intuition to assess system performance. Such evaluation methods may include visual displays [73], [90], [109], [110], [111], [113], user studies [75], [76], etc. We observed that most of the studies conducted qualitative evaluation by displaying top k's visualized retrieval results. Qualitative assessment can reveal issues that cannot be captured by quantitative evaluation, such as how users feel when using a sketch image retrieval system, how easy it is to interact, and whether the system meets their expectations Xinggang et al. [41] conducted qualitative assessments by visualizing the model's filters and the diverse extracted features. The GAN-based method [55] demonstrated generated images that were semantically consistent with the input sketch and consistent with the content of the target

TABLE 6. Summary of the sketch datasets.

Datasets Name	Scale	Category	Public	Notes	Paper
Flickr15k [123]	330 sketches and 15,000 images	33	Public		[17], [18], [40], [42], [44], [47], [48], [49], [59], [62], [63], [64], [67], [72], [93], [102], [124]
Flickr25k [8]	25,000 images and 20,000 sketches	250	Public		[48], [50]
Flickr30K-Sketchy intersection dataset [92]	7,643 images and 4,127 sketches	8	Public		[92]
FlickrLarge (a subset of Flickr3M) [93]	1,600 object images and 200K distractors	80	Public		[93]
Flickr Logos 27 [125]	4,207 logo images	27	Public		[95]
eBDtheque [126]	100 pages	25	Public		[93]
MSCOCO-Sketchy intersection dataset [92]	56,999 images and 16,908 sketches	32	Private		[92]
HUST SI [41]	31,824 images	250	Public	Matching TU-Berlin sketch categories	[27], [41]
PASCAL Car and Bicycle [127]	50 sketches	2	Public	using query images to retrieve the PASCAL VOC 2007 dataset	[41]
PASCAL parts [128]	10,103 images	20	Public		[68]
M. Eitz-SBIR Benchmark [15]	31 sketches and 1,240 images	-	Public		[27], [41], [65]
MECD [63]	30 sketches and 900 images	30	Public	a large Chinese museum collection image set	[62], [63], [64]
Sketchy [26]	75471sketches, 12500 photos	125	Public		[26], [29], [38], [43], [44], [45], [47], [48], [49], [50], [51], [52], [55], [67], [71], [73], [74], [80], [86], [87], [92], [94], [99], [101], [104], [107], [108], [109], [110], [111], [112], [113], [114], [115]
Sketchy extension [26]	provides another 60,502 images	125	Public		[60], [61], [96], [97], [98], [103], [105], [106]
SketchyCOCO [129]	14K pairs	14	Public		[89]
TU-Berlin [8]	20,000 sketches	250	Public		[9], [18], [27], [44], [49], [50], [51], [53], [54], [58], [68], [88], [98], [99], [101], [104], [107], [108], [109], [110], [111], [112], [113], [114], [115]
TU-Berlin Extension [8]	80 sketches for each category and 204,489 images	250	Public		[29], [38], [46], [60], [61], [96], [97], [102], [103], [105], [106]
QuickDraw [35]	50M+ sketches	345	Public		[31], [32], [39], [75], [76], [100], [108], [111], [112], [113]
QuickDraw Extended [106]	203,885 images and 330,111 sketches	110	Public		[95], [96], [97], [105], [106]
QuickDrawCOCO-92c [89]	115800	92	Public		[89]
QMUL Shoe [10]	419 sketches, 419 photos	1	Public		[10], [70], [71], [77], [79], [81], [83], [87]
QMUL-ShoeV2 [80]	2,000 image-sketch pairs	1	Public		[30], [71], [74], [78], [82], [87], [90], [91], [94]
QMUL Chair [10]	297 sketches, 297 photos	1	Public		[10], [71], [79], [81], [83], [87]
QMUL-ChairV2 [10]	400 image-sketch pairs	1	Public		[30], [74], [78], [82], [87], [90], [91]
QMUL Handbag [83]	568 sketches, 568 photos	1	Public		[73], [80], [81], [83]
QMUL FG-SBIR [80]	3,116 images and 8721 sketches	2	Public		[73], [80]
self-built clothing dataset [79]	2,000 sketch-image pairs	1	Public		[79]
self-built UI dataset [12]	3,802 sketches	1	Public		[12]
Color images dataset [9]	35,000 images	250	Private	corresponding to TU-Berlin sketch classes	[9]
Hairstyle Photo-Sketch Dataset (HPSD) [130]	3,600 photos and sketches, and 2400 photo-sketch pairs.	40	Public		[71]
Saavedra-SBIR [131]	53 sketches and 1326 images	50	Public		[44]
S-ROBIN [132]	510 floor plan sketches	3	Private		[56]

TABLE 6. *(Continued.)* **Summary of the sketch datasets.**

CSBIR dataset [28]	419 shoe color sketch-image pairs	1	Private		[28]
[29] Dataset	270 sketches and 73314 images	31	Private		[29]
self-built Manga dataset [69]	903 images and 300 sketches	6	Private		[69]
OI-TrainVal [133]	1.3M training and 26K validation images	141	Public		[66]
OI-Test-LQ [66]	125k test images	141	Public		[66]
Stock4.5M [66]	100 sketches and 4.5 million images	3	Public		[66]
SJTU-Cloth [77]	440 images	4	Public		[77]
RSketch [57]	4,000 remote sensing images and 900 sketches	20	Public		[57], [115]
eCommerce [134]	50,701 images; 5,665 images and 666 sketches	141	Public		[48]
ImageNet-Sketch [58]	50,000 images	1000	Public		[58]
Earth on Canvas [135]	Each modality comprises 100 images per class	14	Private		[84]
ComicLib [39]	181,354 sketches and 2,107,648 objects	17	Private		[39]
ESRIR [95]	53,000 sketches and 53,000 images	606	Public		[95]
CUHK Face Sketch FERET (CUFSF） [136]	1,800 face images and sketches	1	Public		[95]
Labeled Faces in the Wild home (LFW) [137]	21,174 facial images	1	Public		[95]
256_Object Categories [138]	30,607 images	256	Public		[95]
PKU-Sketch [70]	200 sketches and 400 images	200	Public		[90]
DoodleUINet [139]	11k doodles	16	Public		[75], [76]
Sketch Re-ID [70]	200 sketches and 400 images	1	Public		[91]
SketchyScene [140]	11,000 sketches	44	Public		[88]

image during qualitative evaluation Tian et al. [111] show the results of the visualization of the RPKD method on Sketchy. Qualitative analyses revealed that sketch query and retrieval results are more similar in shape and structure than category information. The reason for this problem may be due to the highly abstract nature of sketches. Also, they visualized the distribution of unseen data from random samples of graph images and sketches using the t-SNE tool. From these qualitative evaluations, it is easy to recognize that high-quality queries are crucial for retrieving the correct candidates, and that redundant noise and over-abstracted low-quality queries tend to cause erroneous retrieval results. For the ZS-SBIR task, when the gallery contains all candidate images, the retrieval results for queries targeting the unseen class tend to have results that are visually highly similar to the seen class.

Although the focus of SBIR research is mainly related to methods, data, and user interfaces, hardware platforms also play an essential role in the efficiency and effectiveness of SBIR systems. Deep learning SBIR algorithms are computationally demanding. Utilizing Graphics Processing Units (GPUs) can significantly accelerate computation compared to traditional Central Processing Units (CPUs) due to their parallel processing capabilities [96], [115]. SBIR systems typically require significant memory resources, especially with large image databases. Memory limitations will affect the database size that can be indexed efficiently or the complexity of the models that can be used. Experiments have demonstrated that deep hashing architectures in SBIR enable retrieval that outperforms general deep learning methods in terms of time complexity and storage complexity, saving nearly four orders of magnitude in memory load and retrieval time [38], [61] and achieving real-time performance even for retrieval from large, multi-million scale databases like QuickDraw [100]. Hardware limitations of mobile and tablet devices may affect the development and performance of SBIR applications designed for these platforms [75], [76]. Optimizing algorithms for mobile devices is critical; otherwise, providing real-time performance in SBIR systems may be difficult, which can affect the user experience, especially in interactive applications that must respond quickly to sketching queries. SBIR systems often rely on stylus or touch-based input devices for sketching. Compatibility with various input devices and their integration with hardware platforms is critical for a good user experience. Addressing these issues involves a combination of hardware platforms, algorithm optimization, and software development practices to utilize available resources efficiently.

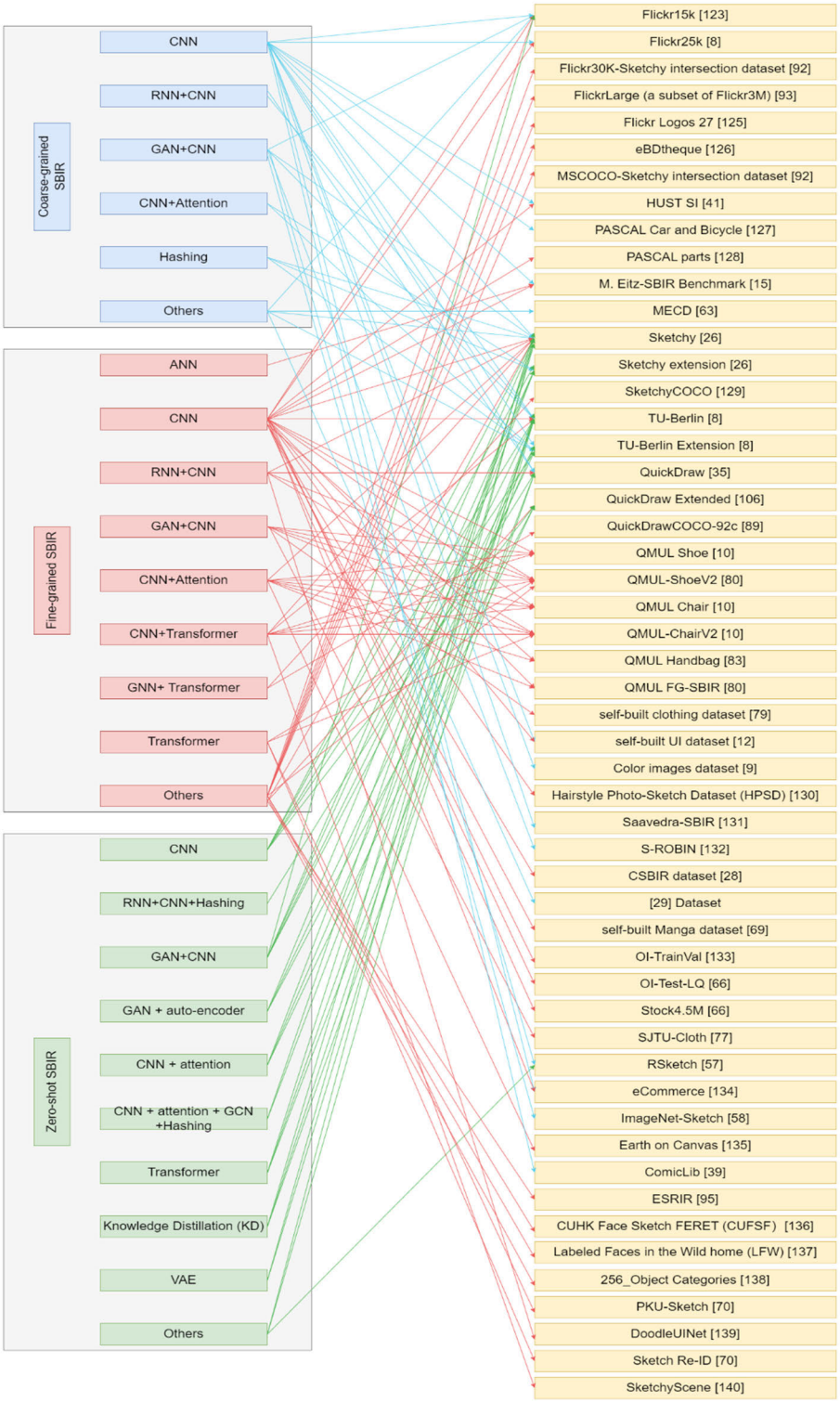

FIGURE 8. Method-data association diagram.

D. RQ4: WHAT ARE THE POTENTIAL APPLICATIONS FOR SBIR USING DEEP LEARNING?

SBIR using deep learning has shown great potential as another image retrieval solution in addition to TBIR and CBIR. By reviewing the research on SBIR in recent years, Question 4 will outline some future potential applications.

Applications in e-commerce. Shoppers search for products that meet their wishes through hand-drawn sketches.

TABLE 7. Evaluation metrics covered in the papers reviewed.

Evaluation method	Evaluation metrics	Papers
Qualitative Evaluation		[9], [12], [17], [18], [28], [29], [30], [31], [32], [38], [40], [41], [42], [44], [47], [48], [49], [50], [51], [55], [56], [60], [63], [66], [67], [68], [69], [70], [71], [72], [73], [74], [75], [76], [77], [79], [81], [82], [83], [84], [87], [88], [89], [90], [93], [94], [95], [96], [97], [99], [100], [101], [102], [103], [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114]
Quantitative Evaluation	AP (Average Precision)	[29], [62], [63], [79]
	AUC (area under the curve)	[31]
	AUIR (area under inverse rank)	[74]
	classification precision (cla-precision)	[79]
	HD2 (precision of Hamming distance with radius 2)	[38]
	IOU (Intersection Over Union between split sketches and their ground truths)	[77]
	Kendal score	[27], [44], [49]
	m@A (ranking percentile)	[82]
	m@B (1/rank)	[82]
	mAP	[17], [18], [32], [38], [40], [41], [42], [44], [45], [46], [47], [48], [49], [51], [56], [59], [60], [61], [64], [66], [67], [69], [72], [84], [91], [93], [96], [97], [98], [99], [100], [101], [102], [103], [104], [105], [107], [108], [109], [110], [111], [112], [113], [114], [115]
	memory loads	[38], [58], [61]
	MRP (Mean Rank Precision)	[62], [63]
	NDCG(Normalized Discounted Cumulative Gain)	[66]
	precision@k	[29], [32], [38], [55], [58], [61], [66], [84], [89], [92], [96], [97], [98], [99], [101], [103], [104], [105], [106], [107], [108], [109], [110], [111], [112], [113], [114], [115]
	precision-recall curve	[18], [38], [40], [44], [56], [67], [72]
	precision/recall scores	[93], [95]
	recall @ K	[26], [43], [49], [52], [58], [86], [89], [92]
	retrieval accuracy	[9], [10], [12], [28], [30], [39], [50], [51], [65], [70], [71], [73], [74], [75], [76], [77], [78], [79], [80], [81], [82], [83], [87], [90], [94]
	retrieval time	[18], [31], [38], [58], [61], [65], [75], [87]
	% correctly ranked triplets	[10]

For example, when you are walking on the street and like someone else's clothes but are too embarrassed to take pictures to get pictures, you can use the SBIR method. This makes the online shopping experience more interactive and personalized. FG-SBIR lays the foundation for the promotion of this application [28].

Application in children's education. By tracing shadow lines that guide the image across the canvas, children's hand sketches retrieve rich visual content, allowing them to explore a wide range of visual references to enhance their learning [31].

Applications in medical diagnosis. Doctors and other professionals can use hand-drawn sketches to retrieve relevant medical images to aid medical diagnosis. Promote medical research and education by searching for case images similar to this content based on hand-drawn sketches [141].

Applications in law enforcement. Law enforcement agencies can use SBIR to help identify suspects or missing persons based on sketches when investigating offenses. This technology may assist in criminal investigations and forensic reconstructions [58].

Applications in art and design. Artists, designers, and architects can quickly search for visual references using SBIR systems. This can help with ideation, concept development, and the creation of new designs or artwork [76].

E. RQ5: WHAT ARE THE FUTURE DIRECTIONS FOR SBIR USING DEEP LEARNING?

Sketch as a modality has its limitations, such as its inability to reflect the texture and color of natural images and the distortion of sketches caused by different drawing abilities. In view of this, future SBIR research directions, especially FG-SBIR, should integrate color or texture information to improve performance [28]. From an HCI perspective, creating interactive interfaces that allow users to refine or modify retrieval results using sketch input can enhance the user experience [12]. Additionally, integrating sketch information with other modalities such as attributes, text descriptions, or audio helps generate more diverse and contextual results [31], [80], [105]. The popularity of smart terminals and HCI technologies drives the idea that our research should be more orientated towards generalized SBIR tasks and that exploring scene-level SBIR is more in line with real-world needs, in which sketches and images can contain multiple instances of objects and, in particular, new requirements for exploring retrievals that contain non-rigid objects are being put forward [142].

Currently, supervised CNN is mostly used to solve the SBIR problem, which relies on labeled data sets. In the future, weakly supervised or unsupervised learning methods can be explored more to reduce the reliance on large labeled datasets, making SBIR more accessible and adaptable to various fields. At the same time, we encourage the use of QuickDraw's large-scale fine-grained data sets as a research benchmark. The content of its sketches is closer to the uneven levels of

ordinary people and more consistent with actual situations. The combination of FG-SBIR and the attention mechanism effectively improves the retrieval results, but the combination of the attention mechanism lacks effective interpretability. Future focus on the interpretability of methods could play an important role in critical applications such as forensics and medical imaging. Most studies include data preprocessing, and it is recommended to jointly train preprocessing and models into a unified deep framework [79]. ZS-SBIR can be extended to handle more complex cross-modal retrieval tasks, such as retrieval involving multiple modalities. In the future, we can evaluate the SBIR-based approach in progressively more challenging and realistic environments, exploring other modalities to improve retrieval performance and enable more natural and intuitive interaction with the system. Another direction is to develop a Data-Free Learning (DFL) approach for SBIR, which utilizes pre-trained unimodal classification models to learn the cross-modal metric space for retrieval without accessing the training data. Accessing paired photo sketch datasets in current research is challenging, so this data-free learning approach is practical for data-scarce tasks like SBIR [143].

VI. CONCLUSION AND LIMITATION

This paper presents a systematic literature review of deep learning methods in the field of SBIR. This SLR was conducted to highlight the existing research gaps in the specific area of deep learning methodology and to provide useful information on the factors influencing potential future research. This review analyzed 90 relevant papers collected based on the PRISMA framework (2016 to June 2023) in five aspects: the SBIR approach, the dataset, evaluation metrics, the applications, and future directions. Overall, deep learning methods for SBIR provide better performance and accuracy than traditional manual feature methods. The deep learning methods are all equally capable of retrieving the target image based on the input sketches used in the model. The most effective deep learning methods for SBIR are CNN-based methods. The hierarchical structure of CNN helps the model to automatically extract and learn the abstract features of the sketch. In addition, RNN and Transformer are able to model the sequence information of sketches and capture the dependencies between sequences. We observed that Sketchy, the most commonly used sketch dataset, contains multimodal data including sketches, associated textual descriptions, and natural images associated with them, and it is valuable in studying cross-modal retrieval. It is also seen that the factors affecting SBIR are based on the relevance of the model, the data, and the way it is evaluated in relation to other factors. It is also shown that the research direction for SBIR using deep learning methods lies in how to improve the working model and utilize multimodal information in order to improve accuracy and provide more possibilities for commercial applications of image retrieval. This SLR will be useful for researchers interested in SBIR methods, data, evaluation approaches, and future trends.

Regarding the limitations of this paper, our study only covers the relatively recent period from 2016 to June 2023 and does not cover the full range of SBIR studies involving deep learning. Furthermore, in this review paper, four valid scientific databases were used to search for papers and other databases could be used to supplement the search in the future. Only journals and conference papers published in English were used in this paper. No other publications, such as books and non-English papers, were used. Five RQs were mentioned and answered in this paper, but many questions are worth pondering under this research topic, such as comparing different loss functions. SBIR is relevant to the research on recognition, detection, classification, and segmentation of sketches, and future research can compare and migrate with algorithms related to these researches.

REFERENCES

[1] M. Alkhawlani, M. Elmogy, and H. E. Bakry, "Text-based, content-based, and semantic-based image retrievals: A survey," *Int. J. Comput. Inf. Technol.*, vol. 4, no. 1, pp. 58–66, 2015.

[2] U. Singhania and B. Tripathy, "Text-based image retrieval using deep learning," in *Encyclopedia of Information Science and Technology*, 5th ed. USA: ISI Global, 2021, pp. 87–97.

[3] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000, doi: 10.1109/34.895972.

[4] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, and J. Zhu, "Deep learning for content-based image retrieval: A comprehensive study," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 157–166.

[5] L. Piras and G. Giacinto, "Information fusion in content based image retrieval: A comprehensive overview," *Inf. Fusion*, vol. 37, pp. 50–60, Sep. 2017.

[6] S. R. Dubey, "A decade survey of content based image retrieval using deep learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 5, pp. 2687–2704, May 2022.

[7] Y. Li and W. Li, "A survey of sketch-based image retrieval," *Mach. Vis. Appl.*, vol. 29, no. 7, pp. 1083–1100, Oct. 2018, doi: 10.1007/s00138-018-0953-8.

[8] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–10, Aug. 2012.

[9] J. Ahmad, K. Muhammad, and S. W. Baik, "Data augmentation-assisted deep learning of hand-drawn partially colored sketches for visual search," *PLoS ONE*, vol. 12, no. 8, Aug. 2017, Art. no. e0183838, doi: 10.1371/journal.pone.0183838.

[10] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C. C. Loy, "Sketch me that shoe," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 799–807, doi: 10.1109/CVPR.2016.93.

[11] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Sketch-a-net: A deep neural network that beats humans," *Int. J. Comput. Vis.*, vol. 122, no. 3, pp. 411–425, May 2017, doi: 10.1007/s11263-016-0932-3.

[12] F. Huang, J. F. Canny, and J. Nichols, "Swire: Sketch-based user interface retrieval," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–10, doi: 10.1145/3290605.3300334.

[13] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang, "MindFinder: Interactive sketch-based image search on millions of images," in *Proc. 18th ACM Int. Conf. Multimedia*, Oct. 2010, pp. 1605–1608.

[14] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004, doi: 10.1023/B:Visi.0000029664.99615.94.

[15] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[16] R. Hu, M. Barnard, and J. Collomosse, "Gradient field descriptor for sketch based retrieval and localization," in *Proc. IEEE Int. Conf. Image Process.*, 2010, pp. 1025–1028.

[17] Y. Qi, Y.-Z. Song, H. Zhang, and J. Liu, "Sketch-based image retrieval via Siamese convolutional neural network," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 2460–2464.

[18] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse, "Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network," *Comput. Vis. Image Understand.*, vol. 164, pp. 27–37, Nov. 2017, doi: 10.1016/j.cviu.2017.06.007.

[19] M. Indu and K. V. Kavitha, "Survey on sketch based image retrieval methods," in *Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT)*, Mar. 2016, pp. 1–4.

[20] X. Zhang, X. Li, Y. Liu, and F. Feng, "A survey on freehand sketch recognition and retrieval," *Image Vis. Comput.*, vol. 89, pp. 67–87, Sep. 2019, doi: 10.1016/j.imavis.2019.06.010.

[21] J. Ziheng, "State-of-the-art survey of deep learning based sketch retrieval," in *Proc. Int. Conf. Artif. Intell. Comput. Eng. (ICAICE)*, Oct. 2020, pp. 6–14.

[22] P. Xu, T. M. Hospedales, Q. Yin, Y.-Z. Song, T. Xiang, and L. Wang, "Deep learning for free-hand sketch: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 285–312, Jan. 2023, doi: 10.1109/TPAMI.2022.3148853.

[23] R. Sarkis-Onofre, F. Catalá-López, E. Aromataris, and C. Lockwood, "How to properly use the PRISMA statement," *Systematic Rev.*, vol. 10, no. 1, pp. 1–3, Apr. 2021, doi: 10.1186/s13643-021-01671-z.

[24] X. Wang, J. Tang, and S. Tan, "Successively focused details: Attention-based recurrent multiscale network for sketch-based image retrieval," *J. Electron. Imag.*, vol. 31, no. 6, Dec. 2022, Art. no. 06304, doi: 10.1117/1.jei.31.6.063048.

[25] O. Seddati, S. Dupont, and S. Mahmoudi, "Triplet networks feature masking for sketch-based image retrieval," in *Proc. Int. Conf. Image Anal. Recognit.*, 2017, pp. 296–303.

[26] P. Sangkloy, N. Burnell, C. Ham, and J. Hays, "The sketchy database: Learning to retrieve badly drawn bunnies," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–12, Jul. 2016, doi: 10.1145/2897824.2925954.

[27] Y. L. Yan, X. G. Wang, X. Yang, X. Bai, and W. Y. Liu, "Joint classification loss and histogram loss for sketch-based image retrieval," in *Proc. Int. Conf. Image Graph.*, 2017, pp. 238–249.

[28] Y. Xia, S. B. Wang, Y. R. Li, L. H. You, X. S. Yang, and J. J. Zhang, "Fine-grained color sketch-based image retrieval," in *Proc. Adv. Comput. Graph.*, 2019, pp. 424–430.

[29] L. Wang, X. Qian, Y. Zhang, J. Shen, and X. Cao, "Enhancing sketch-based image retrieval by CNN semantic re-ranking," *IEEE Trans. Cybern.*, vol. 50, no. 7, pp. 3330–3342, Jul. 2020, doi: 10.1109/TCYB.2019.2894498.

[30] U. R. Muhammad, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Learning deep sketch abstraction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8014–8023, doi: 10.1109/CVPR.2018.00836.

[31] J. Choi, H. Cho, J. Song, and S. M. Yoon, "SketchHelper: Real-time stroke guidance for freehand sketch retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2083–2092, Aug. 2019, doi: 10.1109/TMM.2019.2892301.

[32] J. Collomosse, T. Bui, and H. Jin, "LiveSketch: Query perturbations for guided sketch-based visual search," presented at the *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2874–2882.

[33] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022, doi: 10.1109/TNNLS.2021.3084827.

[34] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D: Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306, doi: 10.1016/j.physd.2019.132306.

[35] D. Ha and D. Eck, "A neural representation of sketch drawings," 2017, *arXiv:1704.03477*.

[36] I. Goodfellow, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[37] A. Creswell and A. A. Bharath, "Adversarial training for sketch retrieval," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 798–809.

[38] L. Liu, F. Shen, Y. Shen, X. Liu, and L. Shao, "Deep sketch hashing: Fast free-hand sketch-based image retrieval," presented at the *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2298–2307.

[39] X. Wei, H. Zhao, Z. Gao, Y. Zhang, J. Zhou, Z. Chen, and Q. Lu, "ComicLib: A new large-scale comic dataset for sketch understanding," in *Proc. Int. Conf. Digit. Image Computing: Techn. Appl. (DICTA)*, Nov. 2022, pp. 1–8, doi: 10.1109/DICTA56598.2022.10034579.

[40] D. Yu, Y. Liu, Y. Pang, Z. Li, and H. Li, "A multi-layer deep fusion convolutional neural network for sketch based image retrieval," *Neurocomputing*, vol. 296, pp. 23–32, Jun. 2018, doi: 10.1016/j.neucom.2018.03.031.

[41] X. Wang, X. Duan, and X. Bai, "Deep sketch feature for cross-domain image retrieval," *Neurocomputing*, vol. 207, pp. 387–397, Sep. 2016, doi: 10.1016/j.neucom.2016.04.046.

[42] J. Lei, K. Zheng, H. Zhang, X. Cao, N. Ling, and Y. Hou, "Sketch based image retrieval via image-aided cross domain learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 3685–3689.

[43] O. Seddati, S. Dupont, and S. Mahmoudi, "DeepSketch 3: Analyzing deep neural networks features for better sketch recognition and sketch-based image retrieval," *Multimedia Tools Appl.*, vol. 76, no. 21, pp. 22333–22359, Nov. 2017, doi: 10.1007/s11042-017-4799-2.

[44] T. Bui, L. Ribeiro, M. Ponti, and J. Collomosse, "Sketching out the details: Sketch-based image retrieval using convolutional neural networks with multi-stage regression," *Comput. Graph.*, vol. 71, pp. 77–87, Apr. 2018, doi: 10.1016/j.cag.2017.12.006.

[45] V. A. Kumar, K. S. Rajesh, and M. Wilscy, "Cross domain descriptor for sketch based image retrieval using Siamese network," in *Proc. 5th Int. Conf. Image Inf. Process. (ICIIP)*, Nov. 2019, pp. 591–596.

[46] Y. Song, J. Lei, B. Peng, K. Zheng, B. Yang, and Y. Jia, "Edge-guided cross-domain learning with shape regression for sketch-based image retrieval," *IEEE Access*, vol. 7, pp. 32393–32399, 2019, doi: 10.1109/ACCESS.2019.2903534.

[47] X. Zhang, Y. Huang, Q. Zou, Y. Pei, R. Zhang, and S. Wang, "A hybrid convolutional neural network for sketch recognition," *Pattern Recognit. Lett.*, vol. 130, pp. 73–82, Feb. 2020, doi: 10.1016/j.patrec.2019.01.006.

[48] J. M. Saavedra, J. Morales, and N. Murrugarra-Llerena, "SBIR-BYOL: A self-supervised sketch-based image retrieval model," *Neural Comput. Appl.*, vol. 35, no. 7, pp. 5395–5408, Mar. 2023, doi: 10.1007/s00521-022-07978-9.

[49] O. Seddati, S. Dupont, and S. Mahmoudi, "Quadruplet networks for sketch-based image retrieval," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2017, pp. 189–196, doi: 10.1145/3078971.3078985.

[50] N. Devis, N. J. Pattara, S. Shoni, S. Mathew, and V. A. Kumar, "Sketch based image retrieval using transfer learning," in *Proc. 3rd Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, Jun. 2019, pp. 642–646, doi: 10.1109/ICECA.2019.8822021.

[51] H. Zhang, P. She, Y. Liu, J. Gan, X. Cao, and H. Foroosh, "Learning structural representations via dynamic object landmarks discovery for sketch recognition and retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4486–4499, Sep. 2019, doi: 10.1109/TIP.2019.2910398.

[52] O. Seddati, S. P. Dupont, S. D. Mahmoudi, and T. Dutoit, "Towards human performance on sketch-based image retrieval," in *Proc. Cbmi*, 2022, pp. 77–83, doi: 10.1145/3549555.3549582.

[53] J.-Y. He, X. Wu, Y.-G. Jiang, B. Zhao, and Q. Peng, "Sketch recognition with deep visual-sequential fusion model," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 448–456, doi: 10.1145/3123266.3123321.

[54] Q. Jia, X. Fan, M. Yu, Y. Liu, D. Wang, and L. J. Latecki, "Coupling deep textural and shape features for sketch recognition," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 421–429, doi: 10.1145/3394171.3413810.

[55] L. Guo, J. Liu, Y. Wang, Z. Luo, W. Wen, and H. Lu, "Sketch-based image retrieval using generative adversarial networks," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1–12.

[56] D. Sharma, N. Gupta, C. Chattopadhyay, and S. Mehta, "REXplore: A sketch based interactive explorer for real estates using building floor plan images," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2018, pp. 61–64.

[57] F. Xu, W. Yang, T. Jiang, S. Lin, H. Luo, and G.-S. Xia, "Mental retrieval of remote sensing images via adversarial sketch-image feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7801–7814, Nov. 2020, doi: 10.1109/TGRS.2020.2984316.

[58] E. S. Sabry, S. Elagooz, and F. E. Abd El-Samie, "Sketch-based retrieval approach using artificial intelligence algorithms for deep vision feature extraction," *Axioms*, vol. 11, no. 12, p. 663, Dec. 2022, doi: 10.3390/axioms11120663.

[59] Z. Bai, H. Hou, and N. Kong, "Sketch based image retrieval with adversarial network," in *Proc. ICVIP*, 2020, pp. 148–152, doi: 10.1145/3376067.3376070.

[60] P. Lu, G. Huang, H. Lin, W. Yang, G. Guo, and Y. Fu, "Domain-aware SE network for sketch-based image retrieval with multiplicative Euclidean margin Softmax," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 3418–3426, doi: 10.1145/3474085.3475499.

[61] X. H. Wu and S. J. Xiao, "Sketch-based image retrieval via compact binary codes learning," in *Proc. Neural Inf. Process., 25th Int. Conf.*, 2018, pp. 294–306.

[62] F. Huang, Y. Cheng, C. Jin, Y. J. Zhang, and T. Zhang, "Enhancing sketch-based image retrieval via deep discriminative representation," in *Proc. ECAI*, 2016, p. 1626.

[63] F. Huang, C. Jin, Y. Zhang, K. Weng, T. Zhang, and W. Fan, "Sketch-based image retrieval with deep visual semantic descriptor," *Pattern Recognit.*, vol. 76, pp. 537–548, Apr. 2018, doi: 10.1016/j.patcog.2017.11.032.

[64] F. Huang, C. Jin, Y. Zhang, and T. Zhang, "Towards sketch-based image retrieval with deep cross-modal correlation learning," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 907–912, doi: 10.1109/ICME.2017.8019432.

[65] X. Shi, H. Chen, and X. Zhao, "A novel sketch-based image retrieval method inspired by retina," in *Proc. 2nd Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2019, pp. 389–393, doi: 10.1109/ICAIBD.2019.8837001.

[66] A. Black, T. Bui, L. Mai, and H. J. Collomosse, "Compositional sketch search," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 2668–2672.

[67] T. Portenier, Q. Hu, P. Favaro, and M. Zwicker, "SmartSketcher: Sketch-based image retrieval with dynamic semantic re-ranking," in *Proc. Symp. Sketch-Based Interfaces Model.*, Jul. 2017, pp. 1–12, doi: 10.1145/3092907.3092910.

[68] R. K. Sarvadevabhatla, I. Dwivedi, A. Biswas, and S. Manocha, "SketchParse: Towards rich descriptions for poorly drawn sketches using multi-task hierarchical deep networks," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 10–18, doi: 10.1145/3123266.3123270.

[69] R. Narita, K. Tsubota, T. Yamasaki, and K. Aizawa, "Sketch-based Manga retrieval using deep features," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 3, Nov. 2017, pp. 49–53, doi: 10.1109/ICDAR.2017.291.

[70] L. Pang, Y. W. Wang, Y. Z. Song, T. J. Huang, and Y. H. Tian, "Cross-domain adversarial feature learning for sketch re-identification," in *Proc. 26th ACM Int. Conf. Multimedia*, 2018, pp. 609–617, doi: 10.1145/3240508.3240606.

[71] H. Lin, Y. Fu, P. Lu, S. Gong, X. Xue, and Y.-G. Jiang, "TC-net for iSBIR: Triplet classification network for instance-level sketch based image retrieval," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1676–1684, doi: 10.1145/3343031.3350900.

[72] Q. Qi, Q. Huo, J. Wang, H. Sun, Y. Cao, and J. Liao, "Personalized sketch-based image retrieval by convolutional neural network and deep transfer learning," *IEEE Access*, vol. 7, pp. 16537–16549, 2019, doi: 10.1109/ACCESS.2019.2894351.

[73] J. Xu, H. Sun, Q. Qi, J. Wang, C. Ge, L. Zhang, and J. Liao, "DLA-net for FG-SBIR: Dynamic local aligned network for fine-grained sketch-based image retrieval," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5609–5618, doi: 10.1145/3474085.3475705.

[74] D. Wang, H. Sapkota, X. Liu, and Q. Yu, "Deep reinforced attention regression for partial sketch based image retrieval," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Dec. 2021, pp. 669–678, doi: 10.1109/ICDM51629.2021.00078.

[75] S. Mohian and C. Csallner, "PSDoodle: Fast app screen search via partial screen doodle," in *Proc. IEEE/ACM 9th Int. Conf. Mobile Softw. Eng. Syst. (MobileSoft)*, May 2022, pp. 89–99, doi: 10.1145/3524613.3527816.

[76] S. Mohian and C. Csallner, "PSDoodle: Searching for app screens via interactive sketching," in *Proc. IEEE/ACM 9th Int. Conf. Mobile Softw. Eng. Syst. (MobileSoft)*, May 2022, pp. 84–88, doi: 10.1145/3524613.3527807.

[77] Z. Yang, X. Zhu, J. Qian, and P. Liu, "Dark-aware network for fine-grained sketch-based image retrieval," *IEEE Signal Process. Lett.*, vol. 28, pp. 264–268, 2021, doi: 10.1109/LSP.2020.3043972.

[78] G. Zhou, Z. Ji, X. Chen, and B. Wang, "StrokeNet: Harmonizing stoke domains between sketches and natural images for sketch-based image retrieval," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 3370–3375.

[79] X. Zhang, M. Shen, X. Li, and F. Feng, "A deformable CNN-based triplet model for fine-grained sketch-based image retrieval," *Pattern Recognit.*, vol. 125, May 2022, Art. no. 108508, doi: 10.1016/j.patcog.2021.108508.

[80] Q. Yu, J. Song, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Fine-grained instance-level sketch-based image retrieval," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 484–500, Feb. 2021, doi: 10.1007/s11263-020-01382-3.

[81] S. Gui, Y. Zhu, X. Qin, and X. Ling, "Learning multi-level domain invariant features for sketch re-identification," *Neurocomputing*, vol. 403, pp. 294–303, Aug. 2020, doi: 10.1016/j.neucom.2020.04.060.

[82] D. Dai, X. Tang, Y. Liu, S. Xia, and G. Wang, "Multi-granularity association learning for on-the-fly fine-grained sketch-based image retrieval," *Knowl.-Based Syst.*, vol. 253, Oct. 2022, Art. no. 109447, doi: 10.1016/j.knosys.2022.109447.

[83] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Deep spatial-semantic attention for fine-grained sketch-based image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5552–5561, doi: 10.1109/ICCV.2017.592.

[84] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "Attention-driven cross-modal remote sensing image retrieval," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. IGARSS*, Jul. 2021, pp. 4783–4786, doi: 10.1109/IGARSS47720.2021.9554838.

[85] Y. Li and X. Liu, "Sketch based thangka image retrieval," in *Proc. IEEE 5th Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Mar. 2021, pp. 2066–2070, doi: 10.1109/IAEAC50856.2021.9390657.

[86] Y. Chen, Z. Zhang, Y. Wang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "AE-net: Fine-grained sketch-based image retrieval via attention-enhanced network," *Pattern Recognit.*, vol. 122, Feb. 2022, Art. no. 108291, doi: 10.1016/j.patcog.2021.108291.

[87] Z. Ling, Z. Xing, J. Li, and L. Niu, "Multi-level region matching for fine-grained sketch-based image retrieval," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 462–470, doi: 10.1145/3503161.3548147.

[88] A. Agarwal, A. Srivastava, I. Nair, S. S. Mishra, V. Dorna, S. R. Nangi, and B. V. Srinivasan, "SketchBuddy: Context-aware sketch enrichment and enhancement," in *Proc. 14th Conf. ACM Multimedia Syst.*, Jun. 2023, pp. 217–228, doi: 10.1145/3587819.3590980.

[89] L. S. Ferraz Ribeiro, T. Bui, J. Collomosse, and M. Ponti, "Scene designer: A unified model for scene search and synthesis from sketch," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2424–2433, doi: 10.1109/ICCVW54120.2021.00275.

[90] C. Chen, M. Ye, M. Qi, and B. Du, "Sketch transformer: Asymmetrical disentanglement learning from dynamic synthesis," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 4012–4020, doi: 10.1145/3503161.3547993.

[91] Y. Zhang, Y. Wang, H. Li, and S. Li, "Cross-compatible embedding and semantic consistent feature construction for sketch re-identification," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3347–3355, doi: 10.1145/3503161.3548224.

[92] Y. Wang, F. Huang, Y. Zhang, R. Feng, T. Zhang, and W. Fan, "Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107148, doi: 10.1016/j.patcog.2019.107148.

[93] S. D. Bhattacharjee, J. Yuan, W. Hong, and X. Ruan, "Query adaptive instance search using object sketches," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1306–1315, doi: 10.1145/2964284.2964317.

[94] K. Pang, K. Li, Y. Yang, H. Zhang, T. M. Hospedales, T. Xiang, and Y.-Z. Song, "Generalising fine-grained sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 677–686, doi: 10.1109/CVPR.2019.00077.

[95] E. S. Sabry, S. S. Elagooz, F. E. Abd El-Samie, W. El-Shafai, and N. A. El-Bahnasawy, "Image retrieval using convolutional autoencoder, InfoGAN, and vision transformer unsupervised models," *IEEE Access*, vol. 11, pp. 20445–20477, 2023, doi: 10.1109/ACCESS.2023.3241858.

[96] O. Tursun, S. Denman, S. Sridharan, E. Goan, and C. Fookes, "An efficient framework for zero-shot sketch-based image retrieval," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108528, doi: 10.1016/j.patcog.2022.108528.

[97] U. Chaudhuri, R. Chavan, B. Banerjee, A. Dutta, and Z. Akata, "BDA-SketRet: Bi-level domain adaptation for zero-shot SBIR," *Neurocomputing*, vol. 514, pp. 245–255, Dec. 2022, doi: 10.1016/j.neucom.2022.09.104.

[98] H. Yu, M. Huang, and J. J. Zhang, "Domain adaptation problem in sketch based image retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 3, pp. 1–17, Aug. 2023, doi: 10.1145/3565368.

[99] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "A simplified framework for zero-shot cross-modal sketch data retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 699–706.

[100] P. Xu, Y. Huang, T. Yuan, T. Xiang, T. M. Hospedales, Y.-Z. Song, and L. Wang, "On learning semantic representations for large-scale abstract sketches," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 9, pp. 3366–3379, Sep. 2021, doi: 10.1109/TCSVT.2020.3041586.

[101] A. Pandey, A. Mishra, V. K. Verma, and A. Mittal, "Adversarial joint-distribution learning for novel class sketch-based image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1391–1400.

[102] C. Bai, J. Chen, Q. Ma, P. Hao, and S. Chen, "Cross-domain representation learning by domain-migration generative adversarial network for sketch based image retrieval," *J. Vis. Commun. Image Represent.*, vol. 71, Aug. 2020, Art. no. 102835, doi: 10.1016/j.jvcir.2020.102835.

[103] A. Pandey, A. Mishra, V. K. Verma, A. Mittal, and H. A. Murthy, "Stacked adversarial network for zero-shot sketch based image retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 2529–2538, doi: 10.1109/WACV45572.2020.9093402.

[104] X. Xu, K. Lin, H. Lu, L. Gao, and H. T. Shen, "Correlated features synthesis and alignment for zero-shot cross-modal retrieval," in *Proc. 43rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2020, pp. 1419–1428, doi: 10.1145/3397271.3401149.

[105] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for any-shot sketch-based image retrieval," *Int. J. Comput. Vis.*, vol. 128, nos. 10–11, pp. 2684–2703, Nov. 2020, doi: 10.1007/s11263-020-01350-x.

[106] S. Dey, P. Riba, A. Dutta, J. L. Lladós, and Y.-Z. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2174–2183, doi: 10.1109/CVPR.2019.00228.

[107] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "CrossATNet—A novel cross-attention based framework for sketch-based image retrieval," *Image Vis. Comput.*, vol. 104, Dec. 2020, Art. no. 104003, doi: 10.1016/j.imavis.2020.104003.

[108] J. L. Tian, X. Xu, F. Shen, Y. Yang, and H. T. Shen, "TVT: Three-way vision transformer through multi-modal hypersphere learning for zero-shot sketch-based image retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 2370–2378.

[109] J. L. Tian, X. Xu, K. Wang, Z. Cao, X. L. Cai, and H. T. Shen, "Structure-aware semantic-aligned network for universal cross-domain retrieval," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2022, pp. 278–289, doi: 10.1145/3477495.3532061.

[110] Y. Song, Y. Yu, H. Tang, J. Guo, and Y. Wang, "Cross-modal relation and sketch prototype learning for zero-shot sketch-based image retrieval," in *Proc. 6th Int. Conf. Comput. Sci. Artif. Intell.*, Dec. 2022, pp. 121–128, doi: 10.1145/3577530.3577550.

[111] J. Tian, X. Xu, Z. Wang, F. Shen, and X. Liu, "Relationship-preserving knowledge distillation for zero-shot sketch based image retrieval," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 5473–5481, doi: 10.1145/3474085.3475676.

[112] K. Wang, Y. Wang, X. Xu, X. Liu, W. Ou, and H. Lu, "Prototype-based selective knowledge distillation for zero-shot sketch based image retrieval," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 601–609, doi: 10.1145/3503161.3548382.

[113] J. Tian, X. Xu, Z. Cao, G. Zhang, F. Shen, and Y. Yang, "Zero-shot sketch-based image retrieval with adaptive balanced discriminability and generalizability," in *Proc. ACM Int. Conf. Multimedia Retr.*, Jun. 2023, pp. 407–415, doi: 10.1145/3591106.3592287.

[114] J. Tian, K. Wang, X. Xu, Z. Cao, F. Shen, and H. T. Shen, "Multimodal disentanglement variational AutoEncoders for zero-shot cross-modal retrieval," in *Proc. 45th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2022, pp. 960–969, doi: 10.1145/3477495.3532028.

[115] S. Jiao, X. Han, F. Xiong, X. Yang, H. Han, L. He, and L. Kuang, "Deep cross-modal discriminant adversarial learning for zero-shot sketch-based image retrieval," *Neural Comput. Appl.*, vol. 34, no. 16, pp. 13469–13483, Aug. 2022, doi: 10.1007/s00521-022-07169-6.

[116] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[117] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2156–2164.

[118] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Apr. 2015, pp. 842–850.

[119] P. Sermanet, A. Frome, and E. Real, "Attention for fine-grained categorization," 2014, *arXiv:1412.7054*.

[120] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, and Ł. Kaiser, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[121] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3285–3294.

[122] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2020, pp. 10076–10085.

[123] R. Hu and J. Collomosse, "A performance evaluation of gradient field HOG descriptor for sketch based image retrieval," *Comput. Vis. Image Understand.*, vol. 117, no. 7, pp. 790–806, Jul. 2013, doi: 10.1016/j.cviu.2013.02.005.

[124] D. Xu, X. Alameda-Pineda, J. K. Song, E. Ricci, and N. Sebe, "Academic coupled dictionary learning for sketch-based image retrieval," in *Proc. ACM Multimedia Conf.*, 2016, pp. 1326–1335, doi: 10.1145/2964284.2964329.

[125] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis, "Scalable triangulation-based logo recognition," in *Proc. 1st ACM Int. Conf. Multimedia Retr.*, Apr. 2011, pp. 1–7.

[126] C. Guérin, C. Rigaud, A. Mercier, F. Ammar-Boudjelal, K. Bertet, A. Bouju, J.-C. Burie, G. Louis, J.-M. Ogier, and A. Revel, "EBDtheque: A representative database of comics," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1145–1149.

[127] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. A. Efros, "Data-driven visual similarity for cross-domain image matching," *ACM Trans. Graph.*, vol. 30, no. 6, pp. 1–10, Dec. 2011, doi: 10.1145/2070781.2024188.

[128] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. Yuille, "Detect what you can: Detecting and representing objects using holistic models and body parts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1979–1986.

[129] C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, and C. Zou, "SketchyCOCO: Image generation from freehand scene sketches," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5173–5182.

[130] W. Yin, Y. Fu, Y. Ma, Y.-G. Jiang, T. Xiang, and X. Xue, "Learning to generate and edit hairstyles," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1627–1635.

[131] J. M. Saavedra and B. Bustos, "An improved histogram of edge local orientations for sketch-based image retrieval," in *Proc. DAGM German Conf. Pattern Recognit. (DAGM)*. Darmstadt, Germany: Springer, Sep. 2010, pp. 432–441.

[132] D. Sharma, N. Gupta, C. Chattopadhyay, and S. Mehta, "DANIEL: A deep architecture for automatic analysis and retrieval of building floor plans," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 420–425.

[133] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *Int. J. Comput. Vis.*, vol. 128, no. 7, pp. 1956–1981, Jul. 2020, doi: 10.1007/s11263-020-01316-z.

[134] P. Torres and J. M. Saavedra, "Compact and effective representations for sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2115–2123.

[135] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, "A zero-shot sketch-based intermodal object retrieval scheme for remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3056392.

[136] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Proc. CVPR*, Jun. 2011, pp. 513–520.
[137] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognition*, 2008, pp. 11–15.
[138] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," CaltechData, USA, Tech. Rep. 20087, 2007, doi: 10.22002/D1.20087.
[139] S. Mohian and C. Csallner, "DoodleUINet: Repository for DoodleUINet drawings dataset and scripts," Zenodo, Switzerland, Tech. Rep. 5144472, 2021, doi: 10.5281/zenodo.5144472.
[140] C. Zou et al., "SketchyScene: Richly-annotated scene sketches," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 421–436.
[141] K. Kobayashi, L. Gu, R. Hataya, T. Mizuno, M. Miyake, H. Watanabe, M. Takahashi, Y. Takamizawa, Y. Yoshida, S. Nakamura, N. Kouno, A. Bolatkan, Y. Kurose, T. Harada, and R. Hamamoto, "Sketch-based medical image retrieval," 2023, *arXiv:2303.03633*.
[142] C. Ge, J. Wang, Q. Qi, H. Sun, T. Xu, and J. Liao, "Scene-level sketch-based image retrieval with minimal pairwise supervision," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, no. 1, 2023, pp. 650–657.
[143] A. Chaudhuri, A. K. Bhunia, Y.-Z. Song, and A. Dutta, "Data-free sketch-based image retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12084–12093.

FAN YANG was born in 1995. She received the B.S. degree in software engineering and the M.S. degree in agricultural engineering from Shanxi Agricultural University, China, in 2018 and 2021, respectively. She is currently pursuing the Ph.D. degree in computer science with Universiti Teknologi Malaysia (UTM), Johor, Malaysia. Her research interests include image retrieval, deep learning, and human–computer interaction.

NOR AZMAN ISMAIL (Member, IEEE) was born in Penang, Malaysia. He received the B.S. degree in computer science and education (mathematics) from Universiti Teknologi Malaysia (UTM), in 1995, the M.S. degree in information technology from Universiti Kebangsaan Malaysia (UKM), in 2000, and the Ph.D. degree in human–computer interaction from Loughborough University, U.K., in 2007.

Throughout his distinguished career, he has held various leadership positions, including the Deputy Dean of Research and Innovation, the Associate Chair of Research and Academic Staff, the Deputy Director (the UTM Digital Media Director), and the Head of the Virtual Visualization Vision (ViCubeLab) Research Group. Currently, he is an Associate Professor with the Faculty of Computing, UTM, Johor Bahru. His research interests include multimodal interaction, UI/UX experimentation, image retrieval, web mining, human-centric AI, and computer vision.

Dr. Ismail has been an active member of the Computing Research Community for over 25 years, contributing significantly to academia. He is an esteemed member of professional societies, including Association for Computing Machinery (ACM) SIGCHI Chapter. His contributions extend to serving on IEEE committees and publications, showcasing his commitment to advancing the field.

YEE YONG PANG was born in Johor, Malaysia, in 1986. He received the B.Eng. and Ph.D. degrees in computer science from Universiti Teknologi Malaysia (UTM), Johor, in 2009 and 2016, respectively. From 2016 to 2021, he was a Lecturer with the Southern University College and TAR University College. Since 2021, he has been with Universiti Teknologi Malaysia, where he is currently involved in teaching and research. His research interests include human–computer interaction (HCI), computer vision, image processing, machine learning, and multimedia. He has contributed to numerous publications in these areas.

VICTOR R. KEBANDE (Member, IEEE) received the Ph.D. degree in computer science (information and computer security architectures and digital forensics) from the University of Pretoria, Hatfield, South Africa. He was a Researcher with the Information and Computer Security Architectures (ICSA) Group and the DIgiFORS Research Group, University of Pretoria. He was a Postdoctoral Researcher with the Internet of Things and People (IOTAP) Center, Department of Computer Science, Malmö University, Malmö, Sweden. He was also a Postdoctoral Researcher in cyber and information security in information systems research subject with the Department of Computer Science, Electrical and Space Engineering, Luleå University of Technology, Luleå, Sweden. He is currently an Assistant Professor in IT security with the Department of Computer Science (DIDA), Blekinge Institute of Technology (BTH), Karlskrona, Sweden. His research interests include cyber, information security, digital forensics in the IoT, the IoT security, digital forensics-incident response, cyber-physical system protection, critical infrastructure protection, cloud computing security, computer systems, distributed system security, threat hunting and modeling and cyber-security risk assessment, blockchain technologies, and privacy-preserving techniques. He also serves as an Editorial Board Member for *Forensic Science International: Reports* journal.

ARAFAT AL-DHAQM received the B.Sc. degree in computer science from the University of Technology, Iraq, and the M.Sc. degree in information security and the Ph.D. degree in computer science from the University Technology Malaysia (UTM). He is currently a Senior Lecturer and a Cybersecurity Researcher with the Computer Information and Science Department, Universiti Teknologi PETRONAS, Bandar Seri Iskandar, Photorefractive Keratectomy, Malaysia. He has a solid foundation in information security, digital forensics, information security governance, and risk management. Furthermore, he was trained by Cybersecurity Malaysia (CSM) as a Certified Digital Forensic Investigator and a Certified Information Security Awareness Manager (CISM). He is a member of the High-Performance Cloud Computing Centre, Universiti Teknologi PETRONAS.

TIENG WEI KOH received the Bachelor of Computer Science (Hons.), Master of Science, and Ph.D. degrees in software engineering from Universiti Putra Malaysia (UPM), in 2004, 2007, and 2012, respectively. He is currently an Associate Professor with the Department of Computer and Information Sciences, Universiti Teknologi PETRONAS. He dedicated two decades to the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, holding the same academic position, since 2019. During this time, he was the Head of the Software Engineering Research Group (SERG) for a four-year term starting, in 2014. Subsequently, he assumed responsibilities as the Program Coordinator of the Master of Software Engineering Program and as a Research Associate with the Malaysian Research Institute on Ageing (MyAgeing). His involvement extended to diverse research projects funded by government agencies, including the Ministry of Higher Education Malaysia (MoHE) and the Ministry of Women, Family, and Community Development (KPWKM). He maintains robust collaborations with SMEs, particularly in the logistics, automotive, and food and beverage sectors. His research interests include empirical software engineering, software metrics, and robotic process automation.