Mathematical Biosciences and Engineering

MBE, 20(12): 21186–21210.
DOI: 10.3934/mbe.2023937
Received: 06 September 2023
Revised: 26 October 2023
Accepted: 30 October 2023
Published: 28 November 2023

http://www.aimspress.com/journal/MBE

Survey

A review of fine-grained sketch image retrieval based on deep learning

Qing Luo[1], Xiang Gao[1], Bo Jiang[1], Xueting Yan[1], Wanyuan Liu[1] and Junchao Ge[2,*]

[1] Yuxi Power Supply Bureau, Yunnan Power Grid Co., Ltd., Yuxi, China
[2] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, China

* **Correspondence:** Email: gjc0224@163.com.

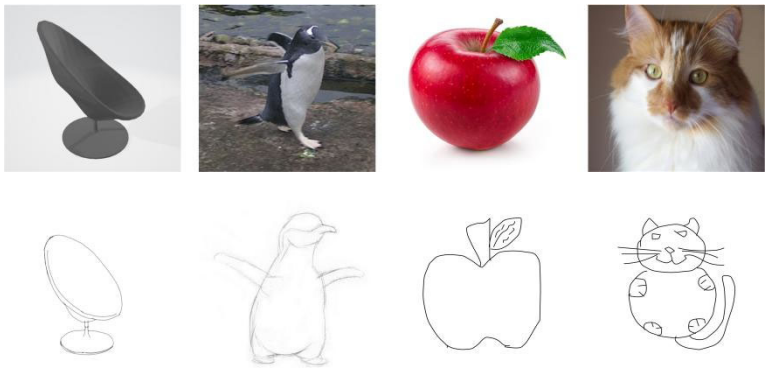
Abstract: Sketch image retrieval is an important branch of the image retrieval field, mainly relying on sketch images as queries for content search. The acquisition process of sketch images is relatively simple and in some scenarios, such as when it is impossible to obtain photos of real objects, it demonstrates its unique practical application value, attracting the attention of many researchers. Furthermore, traditional generalized sketch image retrieval has its limitations when it comes to practical applications; merely retrieving images from the same category may not adequately identify the specific target that the user desires. Consequently, fine-grained sketch image retrieval merits further exploration and study. This approach offers the potential for more precise and targeted image retrieval, making it a valuable area of investigation compared to traditional sketch image retrieval. Therefore, we comprehensively review the fine-grained sketch image retrieval technology based on deep learning and its applications and conduct an in-depth analysis and summary of research literature in recent years. We also provide a detailed introduction to three fine-grained sketch image retrieval datasets: Queen Mary University of London (QMUL) ShoeV2, ChairV2 and PKU Sketch Re-ID, and list common evaluation metrics in the sketch image retrieval field, while showcasing the best performance achieved for these datasets. Finally, we discuss the existing challenges, unresolved issues and potential research directions in this field, aiming to provide guidance and inspiration for future research.

Keywords: fine-grained sketch image retrieval; deep learning; image retrieval

1. Introduction

Image retrieval, as a core research direction in the field of computer vision, is dedicated to

retrieving images highly similar to a given input image from vast image databases. In practical applications, obtaining actual images of the target objects might not be feasible in specific scenarios. For instance, in online shopping, consumers might not be able to provide actual photographs of certain footwear; however, they can articulate the basic visual appearance and features of the shoes through hand-drawn sketches. With the widespread adoption of touchscreen devices such as smartphones and tablets, the acquisition of user-drawn sketches has become more convenient. Consequently, methods for retrieval based on sketch images have gradually garnered attention within the academic community [1,2]. Figure 1 [3,4] provides some hand-drawn sketch images depicting a variety of common objects.



(a)Chair (b)Penguin (c)Apple (d)Cat

Figure 1. Sketch images of common objects, where the images in (a) are from the ProSketch-3DChair [3] dataset and the images in (b), (c) and (d) are from the Sketchy dataset [4].

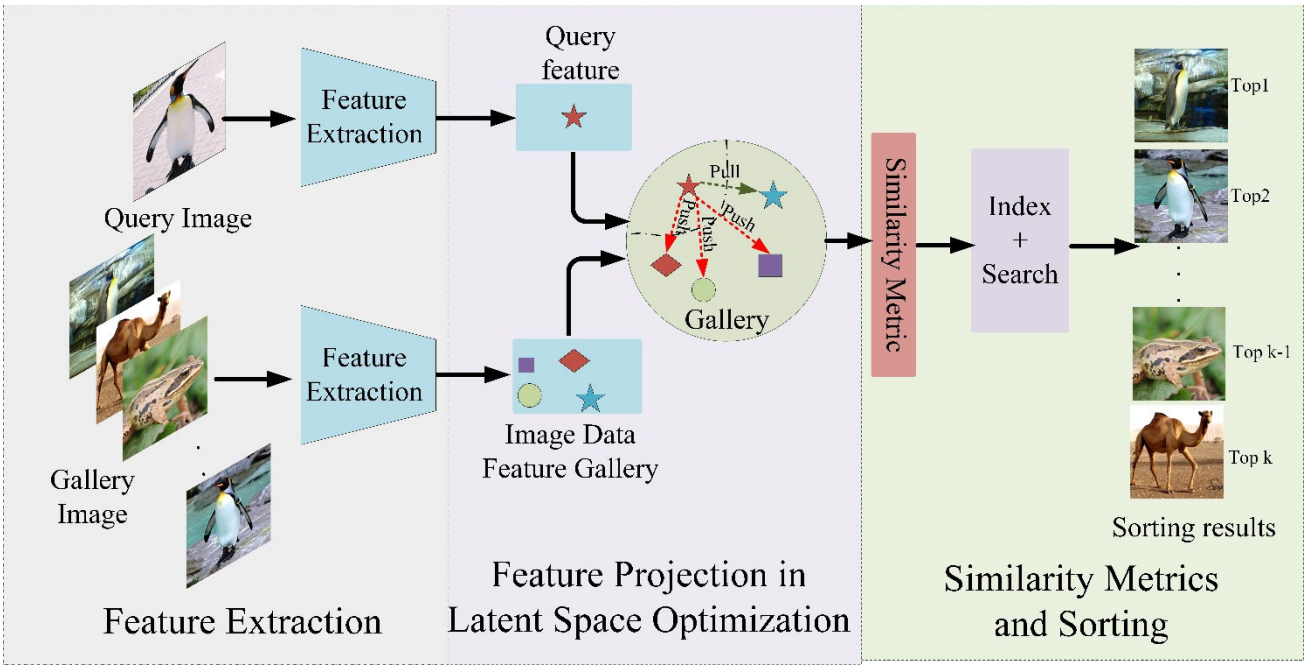
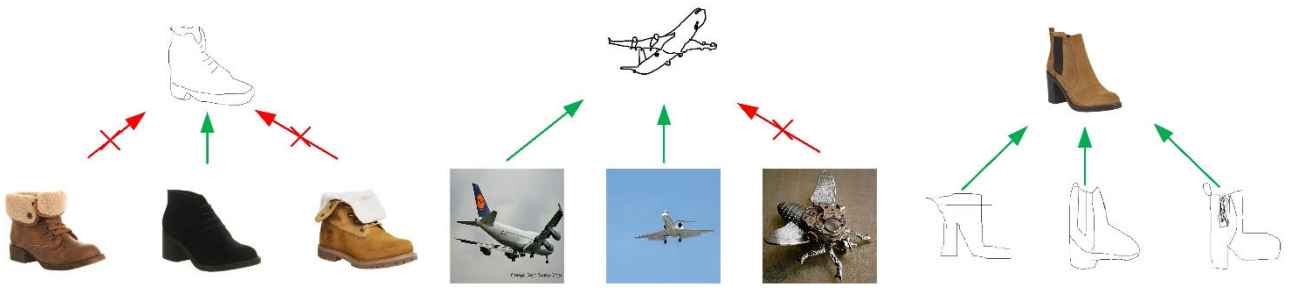


Figure 2. The workflow of intra-domain retrieval.

Based on the scope of retrieval, existing image retrieval methods can be categorized into two classes: Intra-domain retrieval and cross-domain retrieval. Intra-domain retrieval necessitates that both the query data and the data within the retrieval dataset belong to the same modality type. A common instantiation of this retrieval paradigm is exemplified by content-based image retrieval, as depicted in Figure 2. Given the inherent similarity in data distribution between query and retrieval image data, this retrieval paradigm typically employs manual features (such as Histogram of Oriented Gradients, HOG and other conventional feature extraction techniques) or shallow-depth deep learning networks to extract semantically indicative features. Subsequently, through the application of metric learning, the extracted features from both sets are subjected to similarity computation, facilitating the identification of the most analogous retrieval data. This, in turn, enables the arrangement of results in descending order of similarity, thereby achieving the objective of image retrieval.



(a) Fine-grained (b) Coarse-grained (c) Different results were drawn by different painters on the same shoe

Figure 3. Fine-grained sketch image retrieval and coarse-grained sketch image retrieval. Red means mismatch and green means correct match. The images in (a) and (c) are sourced from [4], the images in (b) are sourced from [5].

The realm of sketch image retrieval can be broadly categorized into two classes: coarse-grained retrieval at the category level and fine-grained retrieval at the instance level. This discussion particularly hones in on the intricacies of fine-grained sketch image retrieval. Noteworthy distinctions manifest exists between the tasks of coarse-grained and fine-grained retrieval. Within the realm of fine-grained sketch image retrieval, the central concern revolves around ensuring that the input sketch image and the retrieved results not only share the same category but also align with subtle nuances within that category. For instance, as illustrated in Figure 3(a), notwithstanding the common theme of depicting shoes across all images, the outcomes indicated by the red arrows do not correspond to the anticipated specific style or model, hence being deemed erroneous matches. Conversely, as illustrated in Figure 3(b), in the coarse-grained retrieval task, the requirement is merely for the input sketch and the retrieved image to belong to the same overarching category, thus qualifying as a correct match. Consequently, relative to coarse-grained retrieval, fine-grained sketch image retrieval imposes higher technical demands and presents greater challenges in terms of methodologies.

This exploration delves into the progress and application of deep learning techniques in the domain of fine-grained sketch image retrieval in recent years. Advanced algorithms within this field are meticulously categorized and summarized, offering insights into future research directions along with predictions and recommendations for the trajectory of the field.

Figure 4 serves as a visual representation, summarizing the intricate relationships and connections between the major approaches and topics discussed. This graphical overview lays the foundation for the subsequent sections, which are organized as follows: The subsequent sections are organized as follows: In Section 2, we provide an introduction to the background knowledge of fine-grained sketch image retrieval, emphasizing the key challenges encountered in its research. In Section 3, we enumerate and describe several prominent fine-grained sketch image retrieval datasets, delving into the fundamental issues currently confronted by research in this domain. Moving on to Section 4, we comprehensively expound upon recent representative approaches in fine-grained sketch image retrieval based on deep learning. We classify these approaches based on the types of datasets utilized and the research questions they address, while also showcasing their performance on well-established benchmark datasets. Finally, in Section 5, we present potential avenues for future developments in the field of fine-grained sketch image retrieval, outlining prospective research foci and directions to propel the field forward.

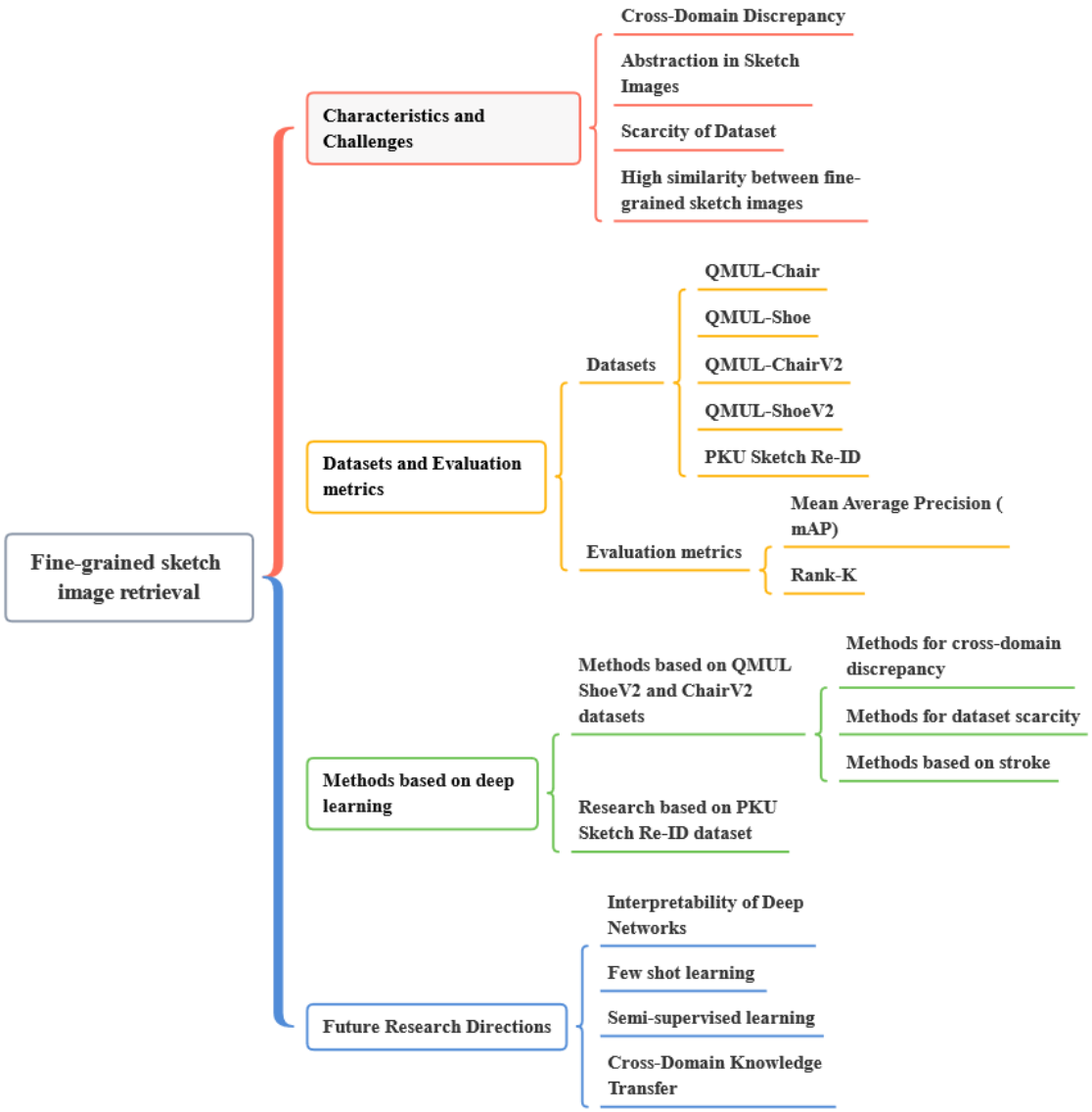


Figure 4. Overview of the structure of fine-grained sketch image retrieval based on deep learning.

2. Background

This section primarily delves into the background knowledge of fine-grained sketch image retrieval. First, it reviews the historical evolution and typical applications of sketch image retrieval. Subsequently, it analyzes the intrinsic features of sketch images, delving into the challenges and potential opportunities encountered within the domain of fine-grained sketch image retrieval.

2.1. Historical evolution and typical applications

Early research on sketch image retrieval [6–9] mainly focused on coarse-grained sketch image retrieval. However, in recent years, the rapid development of deep learning techniques has provided researchers with powerful tools that enable them to explore the field of fine-grained image retrieval more deeply and tap the potential of fine-grained image representations. The intersection of deep learning and fine-grained sketch retrieval methods opens up new opportunities for research and practical applications. This retrieval approach holds significant practical value, particularly in scenarios where the need arises to identify target objects with specific attributes from vast datasets, albeit lacking corresponding reference images. For instance, in e-commerce, fine-grained sketch image retrieval empowers users to search for products similar to their memories of a particular item through rudimentary hand-drawn sketches. This approach not only enhances search convenience but also broadens the spectrum of user retrieval methods. In criminal investigations, instances may arise where only eyewitness descriptions are available due to blind spots in surveillance devices or other reasons, leading to an absence of actual images of suspects. In such contexts, sketch-based person re-identification technology emerges as pivotal. This technique aims to match sketches drawn by professional artists based on eyewitness descriptions with personal images from surveillance videos, thereby furnishing robust technical support for investigative endeavors.

Early research on sketch image retrieval tended to convert images into edge maps and then directly compare them with sketch images [8]. However, this approach performed well only in coarse-grained sketch image retrieval tasks, facing challenges in fine-grained retrieval. This is primarily because fine-grained retrieval aims to distinguish highly similar objects within the same category. To address this, Li et al. [10] explored the use of deformable part models and graph-matching techniques for fine-grained sketch image retrieval in the presence of unaligned poses in 2014. With the advancement of deep learning techniques, Yu et al. [5] introduced the first deep learning-based model for fine-grained sketch image retrieval, where a deep triplet network was employed to learn a shared embedding space between photos and sketch images. To further enhance retrieval performance, researchers began to explore more advanced deep learning approaches. For instance, attention mechanisms with high-order retrieval loss were introduced [11], combining cross-modal image generation with joint discriminative learning [12], as well as leveraging text labels [13] and cross-modal hierarchical attention [14].

Person re-identification is a typical application area for fine-grained sketch image retrieval, which has made significant progress[15–19] in recent years. However, in some specific contexts, it is difficult to obtain photos of target pedestrians. This leads to a research framework for fine-grained sketch-based character re-identification, originally proposed by Pang et al. [20] in 2018. Their pioneering work included the creation of the first sketch-based character re-identification dataset and the use of adversarial learning techniques to narrow the gap between sketch images and photographs of people.

This framework provides new perspectives and possibilities for exploring applications of fine-grained sketch image retrieval. Based on this dataset, Gui et al. [21] adopted a triplet-based classification network as the backbone and incorporated an embedded gradient reversal layer to mitigate the modality discrepancy between sketch images and photographs. Although these approaches yielded commendable results, the challenges inherent in matching sketch images with personal photographs persist without definitive resolution. The principal factors contributing to this conundrum include: (1) The limited scale of sketch-based person re-identification datasets and the scarcity of relevant research therein; (2) the substantial dissimilarity between sketch-based representations and authentic photographic images of persons, with sketches containing mere contour information while photographic representations feature not only precise contours but also vivid color and intricate background details; (3) the predominant portrayal of frontal-view subjects in sketches, whereas real-world photographic images of persons stem from diverse camera angles and are consequently susceptible to the influences of camera perspectives, person postures and occlusions.

2.2. Characteristics and challenges

The major challenges in the field of fine-grained sketch image retrieval include cross-domain differences, i.e., significant differences between sketch images and photographs, and intrinsic differences within the same domain, i.e., high levels of similarity within sketch images. In addition, the abstract nature of sketch images and the scarcity of datasets are also challenges in this domain. Together, these factors affect the performance and accuracy of fine-grained sketch image retrieval algorithms.

(1) Cross-domain discrepancy: A notable challenge in fine-grained sketch image retrieval resides in the significant cross-domain disparity between sketch images and photographs. Specifically, sketch images are characterized by sparse monochromatic lines, while photographs exhibit dense arrays of colored pixels, capturing the perspective projections of visual entities. Concurrently, sketch images encapsulate distinguishing features through subjectively abstract renderings of emblematic contours, whereas photographs meticulously represent images using a plethora of pixels. This cross-domain discrepancy problem hinders the performance of existing fine-grained sketch image retrieval.

(2) High similarity between fine-grained sketch images: The complexity of the challenge in the field of fine-grained image retrieval is not only manifested in the significant differences across domains but also in the intrinsic differences within the same domain. In particular, it should be emphasized that fine-grained sketch image datasets usually exhibit significant internal similarity, mainly in the sense that these images have similar textural and structural features within them, which makes them show a high degree of similarity in local areas. The presence of this internal similarity poses a complex challenge to the performance of fine-grained sketch image retrieval algorithms. This is because this similarity may trigger mismatches and confusion during local feature matching, especially when dealing with a large number of sketch image retrievals. Therefore, solving this problem requires algorithms to more accurately capture and distinguish small feature differences within fine-grained sketch images to improve the accuracy and reliability of fine-grained sketch image retrieval. This may also require further research and development of techniques and methods specifically for fine-grained images.

(3) Abstraction in sketch images: Similar to photographic images, grayscale images and sketch images all fall under the category of homogeneous images. Grayscale images and photographic images exhibit a certain resemblance, as both maintain the fundamental image structure to a considerable extent.

However, sketch images, being products of manual artistic rendering and imbued with a certain level of subjectivity, result in a heightened level of abstraction. As illustrated in Figure 3(c), a single shoe, when drawn by different artists, yields divergent outcomes. Considering the limited drawing capabilities of the majority, it becomes imperative that the creation of the dataset does not rely on professional illustrators for generating sketch images. This, in turn, impinges upon the precision of feature extraction for sketch images, thereby introducing complexities in fine-grained sketch images retrieval.

(4) Scarcity of dataset: Despite the relative ease of acquiring sketch images, the number of sketch image datasets available for fine-grained retrieval tasks that require a high degree of accuracy is very limited. This limits the choices and variety available to researchers in developing and evaluating algorithms, thus putting some constraints on the progress of the field. Therefore, more work is needed to expand and enrich sketch image datasets to advance the field of fine-grained sketch image retrieval. This may include collecting more sketch images of different styles, subjects and qualities to cater to a variety of real-world applications and to improve the generalisability and performance of algorithms.

3. Datasets and evaluation metrics

This section provides a comprehensive summary of the dataset comprising sketch images, as depicted in Table 1. Additionally, a discourse on the challenges encountered during the acquisition of pertinent data for sketch images is presented. Moreover, an exposition of the prevalent evaluation metrics employed for the task of sketch image retrieval is offered.

Table 1. Summary of fine-grained sketch image data sets.

Dataset	Number of photographs	Number of sketch images	Photograph-sketch images
QMUL-Chair	297	297	1 to 1
QMUL-Shoe	419	419	1 to 1
QMUL-ChairV2	400	1275	1 to 3
QMUL-ShoeV2	2000	6730	1 to 3
PKU Sketch Re-ID	400	200	2 to 1

3.1. QMUL ShoeV2 dataset

The QMUL-ShoeV2 [5] dataset stands as one of the prevailing benchmarks in fine-grained image retrieval. Comprising a total of 2000 pairs of meticulously crafted sketch images and corresponding photographs, each pair encompasses a single RGB photograph matched with three distinct sketch images. All images are standardized to dimensions of 256 × 256 pixels. In the dataset, a subset of 1800 pairs of data samples is earmarked for training, contributing to the iterative enhancement of the model. The remaining subset, consisting of 200 pairs, assumes the role of the testing dataset, exclusively employed during the evaluation phase.

3.2. QMUL ChairV2 dataset

The QMUL-ChairV2 [5] dataset is also a fine-grained image retrieval dataset. This dataset

encompasses a total of 1675 images, comprising a collection of 400 pairs of hand-drawn sketch images alongside their corresponding photographic images. All images are standardized to dimensions of 256 × 256 pixels. Within this dataset, a subset of 300 pairs of data samples is earmarked for training, contributing to the iterative enhancement of the model. The remaining subset, consisting of 100 pairs, assumes the role of the testing dataset, exclusively employed during the evaluation phase.

QMUL-ShoeV2 and QMUL-ChairV2 are advancements made by the QMUL laboratory upon the foundation laid by QMUL-Shoe and QMUL-Chair. In these developments, a deliberate emphasis has been placed on diversifying sketch styles. The initial paradigm, where each photograph corresponds to a single sketch image, has been augmented to encompass three sketch images. This augmentation serves the purpose of mitigating the uncertainties in retrieval that stem from the multifaceted nature of user stylistic preferences.

3.3. PKU Sketch Re-ID dataset

The PKU Sketch Re-ID dataset [20], established by the National Engineering Laboratory of Video Technology (NELVT) at Peking University, encompasses a collection of hand-drawn sketch images, conceived for re-identifying individuals. The dataset is composed of depictions of 200 distinct individuals, each portrayed through a single sketch image and two accompanying photographs. These photographs were captured under daylight conditions through the lenses of two orthogonal-view cameras. Prior to inclusion in the dataset, meticulous manual curation was undertaken, involving the extraction of subjects from original images or video frames, thereby ensuring the isolation of each individual within the frame. A noteworthy facet of this dataset is that all the sketch images, which encapsulate the artistic impressions of the subjects, were crafted by five distinct artists, each endowed with their unique stylistic approach to the portrayal.

Figure 5. Examples in PKU Sketch Re-ID datasets [20] (the same box represents a pedestrian ID, including a personal sketch image and two photographs.).

Figure 5 illustrates images from the PKU Sketch Re-ID dataset, revealing that sketch-based person re-identification (Sketch Re-ID) presents a notably challenging endeavor. The primary

difficulties encompass significant modality variations between person sketch images, instances of occluded person subjects, disparities in lighting conditions resulting from distinct camera sources and further compounded by the inherent diversity introduced by various sketch artists. These intricacies inherently differentiate sketch-based person re-identification from conventional image retrieval tasks. In the contemporary landscape of deep learning methodologies, apart from the conventional strategies of augmenting training data and refining network architectures, a concerted effort is directed towards devising algorithms tailored specifically for the challenges inherent in the Sketch Re-ID.

The issue of retrieving specific photographs based on user-provided query sketch images has been addressed through the framework of fine-grained sketch image retrieval. The intricacy of fine-grained sketch image retrieval lies in the acquisition of distinguishing details essential for discerning specific targets. The extraction of discriminative detail features necessitates the support of vast datasets. Nonetheless, the present challenge resides in the limitation of performance attributed to the scarcity of extensive fine-grained sketch-image datasets. This scarcity stems from the fact that the creation of fine-grained sketches demands the involvement of artists, consequently entailing substantial time and cost expenditures in the collection of sketch images. Hence, how to obtain large-scale fine-grained sketch image retrieval datasets is a major challenge.

3.4. Evaluation metrics

In fine-grained image retrieval tasks, the evaluation process often relies upon two key metrics, namely Rank-K [22] and mean average precision (mAP) [23]. The Rank-K metric measures the proportion of correctly matched true labels relative to the first K retrieval results. Specifically, for a single sketch image Query, the library samples are ranked based on similarity from smallest to largest, with K representing the ordinal number that appears in the ranked list. Hence, the Rank-K metric is instrumental in assessing the algorithm's capability to identify accurate labels within the top K outcomes.

$$Rank-K = \frac{1}{M}\sum_{i=1}^{M}\varepsilon \tag{3.1}$$

where ε is the indicator function, which ε equal to 1 only the current top k-ranked list samples contain results consistent with the identity of the query image, otherwise ε is 0.

Precision [24] and recall [25] are ways to measure two different dimensions of model performance. A threshold is needed to obtain four values in a multi-classification task: True positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). Image retrieval can also be viewed as a multi-categorization task, aiming at correctly categorizing the query and the target. Then, precision and recall can be expressed as:

$$precision = \frac{TP}{TP+FP} \qquad recall = \frac{TP}{TP+FN} \tag{3.2}$$

The mAP metric encompasses a more comprehensive evaluation strategy by aggregating the average precision (AP) scores from multi-class tasks and subsequently calculating their average. Thus, AP is calculated as

$$AP_i = \sum_k precision_i^k (recall_i^k - recall_i^{k-1}) \tag{3.3}$$

where $precision_i^k$ and $recall_i^k$ represent precision and recall respectively concerning query q_i in the Rank-k ranked gallery items. The mAP is the mean of all queries and can be expressed as:

$$mAP = \frac{1}{N} \sum_i^N AP_i \tag{3.4}$$

where N is the number of identities of the query sample. This approach facilitates a more holistic evaluation across all queries, capturing the algorithm's performance at different recall levels. Furthermore, the mAP metric excels in its ability to depict the algorithm's efficacy across queries of varying difficulty levels.

4. Fine-grained sketch image retrieval methods based on deep learning

The research field of fine-grained sketch image retrieval is mainly centered around three core datasets due to the scarcity of datasets, namely the QMUL ShoeV2, ChairV2 datasets and the PKU Sketch Re-ID dataset. Each dataset has its unique characteristics and research focus. The QMUL ShoeV2 and ChairV2 datasets have been widely used internationally since a long time ago, mainly for solving the problem of fine-grained sketch image retrieval in commercial applications. Different from this, the PKU Sketch Re-ID dataset focuses more on the research of fine-grained sketch person re-identification. In the field of fine-grained sketch image retrieval, there are problems such as cross-domain differences and dataset scarcity. To explore these challenges and corresponding solutions more clearly, we divide our research into three key directions. First, we focus on how to address cross-domain differences to cope with the significant cross-domain differences between sketch images and photographs. Second, we investigate the problem of dataset scarcity, especially when the limited availability of datasets becomes a central issue when highly accurate fine-grained retrieval tasks are required. Finally, we explore stroke-based research, aiming to conduct research on fine-grained sketch image retrieval based on strokes using techniques such as reinforcement learning.

4.1. Methods based on QMUL ShoeV2 and ChairV2 datasets

4.1.1. Methods for cross-domain discrepancy

As the field of deep learning continues to evolve, the conventional approach involving the training of two branches with edge maps and validation losses has been superseded by the utilization of triplet-branch network incorporating ranking losses, as depicted in the illustrative diagram in Figure 6. Presently, research about fine-grained sketch image retrieval has embraced more intricate network architectures and sophisticated loss formulations to enhance performance and accuracy in the retrieval process.

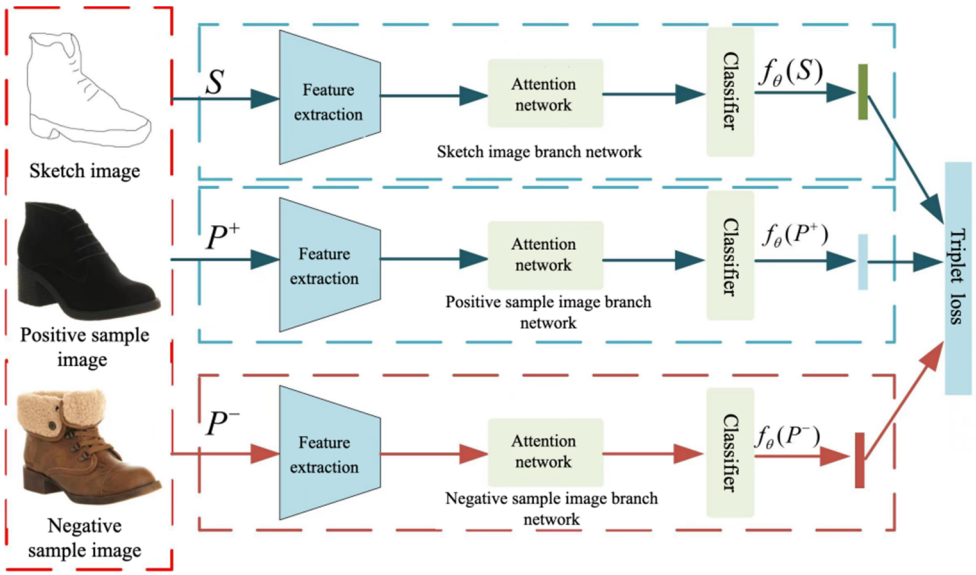
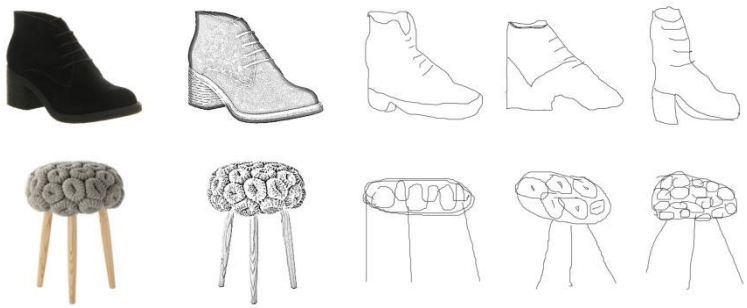


Figure 6. Triplet-branch network. The images in the figure are from [5].

Yu et al. [5] pioneered the task of fine-grained image retrieval based on hand-drawn sketches. They established two fine-grained sketch image retrieval datasets, namely QMUL Chair and QMUL Shoe, and elaborated extensively on their data collection methodology. Furthermore, they introduced a triplet branch network for instance-level fine-grained sketch image retrieval, termed the deep triplet ranking model, and conducted empirical investigations on how deep learning models could attain enhanced performance from augmented datasets. Building upon the aforementioned model, Song et al. [11] have incorporated attention modules into each deep branch of the neural network, enabling the model to focus more on salient regions during the feature learning process. This approach facilitates the fusion of coarse and fine-grained semantic information through feature integration techniques. Additionally, they introduced a high-order learnable energy loss function, establishing correlations between two modal features. This enhancement enhances the robustness of the model to misaligned features across different modalities. Furthermore, the researchers expanded the scope of their study by augmenting the QMUL Chair and Shoe datasets, resulting in the creation of the more diverse QMUL-ShoeV2 and ChairV2 datasets. These enhancements yield three corresponding sketch images for each photographic image. Pang et al. [26] proposed a new mixed-modal puzzle-solving scheme as an effective pre-training strategy. However, due to the serious misalignment problem between the sketch and the photo, resulting in local mismatch, the effect is not ideal in practical applications. Radenovi et al. [27] employed a technique of automated recovery from motion to acquire standard image edge maps for training their network, resulting in enhanced performance. However, a significant proportion of sketch images are produced and gathered by non-professional artists, rendering them inherently abstract, as illustrated in Figure 7. Substantial disparities persist between sketch images and edge maps. For the task of sketch image retrieval, the utilization of edge maps can at times yield counterproductive outcomes. This is primarily attributed to the fact that sketch images lack the well-defined boundaries characteristic of edge maps. Collomosse et al. [9] introduced a novel metric for measuring visual similarity in image retrieval, wherein the metric integrates both structural and aesthetic (i.e., stylistic) constraints. This metric exhibits a significant advancement in style recognition compared to preceding networks. Lin et al. [28] proposed a framework termed deep variational metric learning (DVML),

which explicitly models intra-class variance and addresses intra-class invariance. By leveraging the learned distribution of intra-class variance, it becomes feasible to generate discriminative samples concurrently, thereby enhancing robustness. Xu et al. [29] postulated that local features bear superior discriminative capabilities compared to global features. Consequently, they introduced the local alignment network (LANet), which addresses the challenge of fine-grained sketch image retrieval by aligning intermediate-level local features directly. DLA-Net [29] calculates the distance between the sketch at the same location and all the local features in the photo by this idea, but it increases the computational overhead. Sun et al. [30] proposed that DLI-Net eliminates the background features and only utilizes the foreground features for the matching, which can reduce the computational cost to some extent. However, this approach ignores an important point: The local spatial misalignment between sketches and photographs greatly limits the matching accuracy. Zhang et al. [31] proposed a new extended window local alignment weighted network (EWLAW-Net) based on this approach by using the extended window mechanism, which aligns the extracted local features with the same semantics between photographs and sketches. This innovative approach showcases the significance of emphasizing local features in tackling the intricacies of such retrieval tasks. In the work presented by Ling et al. [32], a multi-level region matching approach termed Multi-Level Region Matching (MLRM) is introduced for the retrieval of fine-grained sketch images. The methodology encompasses two essential components: The discriminative region extraction (DRE) module and the region and level attention (RLA) module.



(a) photographs (b) edge maps (c)sketch images

Figure 7. The comparison of photographs, edge maps and sketch images. The images in the figure are from [5].

The primary objective of investigating the issue of cross-domain disparity lies in the development of an efficient and accurate retrieval model. The endeavor to enable models to comprehend the essence of sketch images in a manner akin to human cognition underscores the significance of formulating a universal and proficient framework. Presently, the Transformer architecture has demonstrated commendable adaptability to vast datasets in fine-grained sketch retrieval tasks. This propensity can be attributed to its capacity to glean features through attention mechanisms, thereby capturing intricate interrelationships among elements. This affords the model a heightened universality, rendering it less reliant solely on the intrinsic characteristics of the data. Moreover, the Transformer is attuned not only to local details but also to holistic representations, facilitating an information flow from local to global scales. Conversely, convolutional neural networks (CNNs), predicated on convolutional operations,

are oriented towards preserving translational invariance in features. The pronounced divergence observed among instances of sketch images contributes to the relative inconspicuousness of CNNs in addressing the task of sketch image retrieval.

4.1.2. Methods for dataset scarcity

Pang et al. [33] introduced a novel unsupervised learning approach aimed at modeling the intrinsic manifold of prototype visual sketch image features. This manifold serves as a basic structure for parameterizing the representation of sketch images and photographs. Subsequently, by embedding new sketch images into this manifold and appropriately updating the representation and retrieval functions, adaptation of the model to new categories is achieved. This advance has significantly advanced research in zero-shot sketch retrieval. Sain et al. [34] proposed an innovative style-agnostic model for sketch image retrieval. Departing from the prevailing methods, their approach introduces a cross-modal variational autoencoder (VAE) to explicitly decompose each sketch image into two distinct components: A semantic content segment shared with the corresponding photograph, and a stylistic segment specific to the sketch artist. This decomposition is achieved by incorporating adaptive components corresponding to the two different styles during the training process of the cross-modal VAE. Ling et al. [35] proposed an unsupervised stroke disentangling algorithm that shows remarkable performance in stroke extraction and sketch image enhancement. Furthermore, two weaknesses of the triplet ranking model are identified and a dual-anchor loss is introduced to mitigate the cosine distance between sketch and photo image pairs. Bhunia et al. [36] introduced an innovative semi-supervised cross-modal retrieval framework that provides a novel solution to data scarcity problems by exploiting a large corpus of unlabelled photographs. In addition, they incorporated discriminator-guided mechanisms to control image synthesis and introduced a regularisation component based on distillation loss. This regularisation component enhances the framework's ability to tolerate noisy training samples. By treating generation and retrieval as two related but distinct problems, they establish a symbiotic relationship between the two. Bhunia et al. [37] have introduced a novel framework based on model-agnostic meta-learning (MAML), which aims to improve the adaptability of fine-grained sketch image retrieval models across different categories/styles. This approach entails using meta-learning to facilitate the rapid adaptation of the model to different user-specific drawing styles with minimal sample input from the user.

One of the barriers to progress in fine-grained sketch image retrieval is data scarcity. Existing hand-drawn sketch image datasets do not comprehensively cover all categories encountered in everyday life, thereby perpetuating the occurrence of zero-shot scenarios. In such cases, users search for objects during the retrieval process that do not fall into the categories present in the dataset used for training. Zero-shot sketch retrieval has gradually gained traction in current research; however, its retrieval accuracy remains relatively constrained, exhibiting a notable disparity from practical applications.

4.1.3. Methods based on stroke

As the field of reinforcement learning continues to evolve, researchers have commenced investigating the interrelationships among strokes and employing reinforcement learning methodologies to delve into fine-grained sketch image retrieval. This endeavor conceptualizes the

process of fine-grained sketch image retrieval as an ongoing sequence of successive decisions. Ha et al. [38] conducted an in-depth exploration involving recurrent neural networks (RNNs) for sketch generation, thereby proposing a comprehensive framework for both conditional and unconditional sketch image synthesis. The study further elucidated a novel methodology geared towards cultivating the robust training of RNNs, specifically tailored for the generation of coherent sketch images in vector format. This pioneering approach established a foundational bedrock for subsequent research endeavors in stroke analysis pertinent to sketch image retrieval. Muhammad et al. [39] proposed the first stroke-level abstraction model for sketch images. This model involves a delicate balance between the recognizability of sketch images and the number of strokes employed in rendering them. Specifically, the model employs reinforcement learning through stroke removal strategies to train a generative framework for abstract sketch image synthesis. The framework learns to predict the strokes that can be safely eliminated without compromising recognizability. Bhunia et al. [40] have introduced a dynamic approach to design wherein retrieval commences as users initiate the drawing process. They have further devised a cross-modal retrieval framework based on reinforcement learning, aimed at refining the ranking of authentic sketch images directly across the entire sketch image dataset. Additionally, a novel reward scheme has been introduced to mitigate issues associated with strokes in unrelated sketch images. Sain et al. [14] designed a novel network that is capable of cultivating sketch-specific hierarchies and exploiting them to match sketches with photos at corresponding hierarchical levels. In a recent contribution by Wang et al. [41], a novel framework has been introduced, which leverages a uniquely designed deep reinforcement learning model to undertake a dual-level exploration aimed at addressing the challenges posed by partially sketched images during training, as well as the selection of attention regions. By directing the model's attention towards pivotal areas within the original sketched images, it demonstrates robustness against superfluous stroke noise, thereby substantially enhancing retrieval precision. Dai et al. [42] posit that during the sketching phase of generating outline images from photographs, significant correlations exist among these incomplete sketch representations. To glean a more efficacious shared joint embedding space between photographs and their corresponding incomplete sketch depictions, they introduce a multi-scale associative learning framework. This framework subsequently refines the embedding space for all partial sketch illustrations. Furthermore, to mitigate the impact of noisy strokes, Bhunia et al. [43] devised a stroke subset selector aimed at identifying strokes with noise, retaining only those strokes that positively contribute to the successful retrieval process. Leveraging a reinforcement learning framework, they quantified the significance of each stroke within a given subset by formulating its contribution to retrieval effectiveness.

The exploration of fine-grained sketch image retrieval based on strokes holds significant potential for commercial applications. However, the translation of this potential into practical problem-solving remains a pivotal challenge. This challenge primarily stems from the difficulty in establishing a universal reward scheme across disparate datasets due to the domain gap issue. The formulation of a suitable reward function, which guides agents in reinforcement learning to acquire desired behaviors, proves to be a formidable task. There exists no definitive optimal approach for constructing a reward function for reinforcement learning, and in certain instances, attempts at formulating such functions can inadvertently lead models astray from intended objectives. Furthermore, as reinforcement learning is tailored to specific data distributions, it fails to accommodate the distinct variability in data distributions among different users for real-world fine-grained sketch image retrieval tasks. The efficient realization of a real-time online system for hand-drawn fine-grained sketch image retrieval,

capable of effectively retrieving images aligned with user requirements from the vast expanse of internet imagery, emerges as a promising avenue for future research endeavors in the domain of sketch image retrieval.

Table 2 Comparison of methods based on QMUL ShoeV2 and ChairV2 datasets.

	QMUL ChairV2			QMUL ShoeV2		
Methods	Rank-1(%)	Rank-5(%)	Rank-10(%)	Rank-1(%)	Rank-5(%)	Rank-10(%)
Triplet SN[5]	47.4	——	84.3	28.7	——	71.6
HOLEF SN[13]	50.7	——	86.3	31.2	——	74.6
SN-RL[9]	51.2	——	86.9	30.8	——	74.2
Triplet Attn[11]	53.4	——	87.6	31.7	——	75.8
CC-Gen [33]	54.2	——	88.2	33.8	——	77.9
Triplet RL[40]	51.2	76.34	89.6	34.1	——	78.8
CMHM[14]	62.5	——	90.7	36.3	——	80.7
Edgemap[27]	53.9	——	87.7	33.8	——	80.9
Edge2sketch[39]	54.3	——	87.5	34.2	——	81.2
DVML[28]	52.8	——	85.2	32.1	——	76.2
SSL[36]	53.3	——	87.5	33.4	——	80.7
StyleMeUp[34]	62.9	——	91.1	36.5	——	81.8
DARP-SBIR[41]	——	35.65	——	——	60.02	——
MGAL[42]	——	81.73	92.56	——	65.31	78.22
SWNT[43]	64.8	79.1	——	43.7	74.9	——
DLA-Net[29]	——	——	——	50.15	——	——
DLI-Net[30]	77.81	——	——	50.0	——	——
EWLAW-Net[31]	81.48	——	——	53.0	——	——
CSR[35]	——	——	——	52.1	——	87.9
MAML[37]	——	——	——	38.3	76.6	——
MLRM[32]	74.3	——	98.2	50.4	——	87.9
SketchTrans[46]	81.7	——	97.4	38.7	——	80.9
CCSC[47]	74.3	——	97.4	33.5	——	80.2

As evident from Table 2, the current research methodologies have exhibited promising outcomes on the relatively modest dataset QMUL ChairV2. Nevertheless, when applied to the more expansive dataset, QMUL ShoeV2, there remains significant room for improvement. The above three research methods have their advantages and disadvantages. The research method based on cross-domain differences mainly focuses on the processing of complete images, which enables fast image retrieval and has a low research threshold. However, in practical research, it is extremely expensive to collect the fine-grained sketch image data to integrate it, and the trained dataset may not perform well in unknown fields, so it becomes crucial to solve the dataset scarcity problem. In addition, in commercial applications, users prefer the immediate retrieval process to reduce the waiting time. Such research methods are called stroke-based methods, and they mainly rely on training strategies for reinforcement learning. However, due to domain gap problems, it is difficult to establish universal reward schemes on different datasets, so it also has some limitations. Currently, fine-grained sketch image retrieval algorithms based on deep learning have surpassed humans in accuracy and speed, indicating significant

progress in this field. We firmly believe that with the continuous development of deep learning technology, more excellent algorithms will emerge to provide more convenience and innovation for the practical application of fine-grained sketch image retrieval.

4.2. Research based on PKU Sketch Re-ID dataset

In the realm of person re-identification, the utilization of professional sketch images to query databases for novel personal information holds paramount significance, which means using sketch photos to find the best matching pedestrian from a large number of pedestrian photos, which is an extremely challenging task of fine-grained sketch image retrieval. Sketch-based person re-identification constitutes a pivotal technology within security systems, facilitating the identification of persons from diverse camera sources. This technique efficiently addresses scenarios in which comprehensive facial sketch information cannot be obtained, thus serving as a complementary approach. A framework for sketch-based person re-identification is illustrated in Figure 8. Pang et al. [20] initially introduced the PKU Sketch Re-ID dataset, providing a foundational dataset for research in sketch-based person re-identification. Moreover, they proposed a cross-domain adversarial framework to filter out domain-sensitive information, aiming to learn domain-invariant features, thereby narrowing the modality gap and enhancing model retrieval performance. However, this approach incurred the loss of certain modality-specific information conducive to pedestrian identity discrimination. It failed to jointly optimize the representation capacity of sketch and pedestrian photograph features, and the feature extraction network was incapable of achieving effective semantic alignment across multiple modalities.

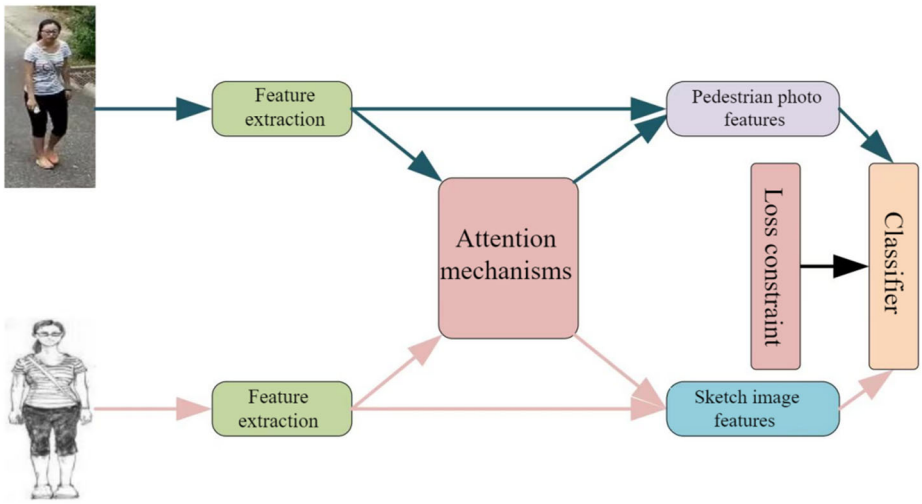


Figure 8. A sketch-based person re-identification framework. The images in the figure are from [20].

Yang et al. [44] attempted to enhance the generalizability of the sketch-to-photo model through domain adaptation techniques, introducing a specialized framework designed for instance-level heterogeneous image retrieval tasks. They overcame the limitations of conventional fine-tuning strategies and traditional domain adaptation methods. Gui et al. [21] delved into the multi-level feature representation of sketch and photo images, employing a triplet classification network as the

foundational architecture. By incorporating spatial attention modules, they amalgamated high-level and intermediate CNN-generated features to represent the input images. Additionally, they employed gradient reversal layers to address domain discrepancies, leading to performance improvements. However, the challenge of modality discrepancy remains intricate and unresolved. Gong et al. [45] proposed a strategy for bias elimination termed Random Color Dropout (RCD). This strategy postulates the existence of color bias between query images and database images. It aims to mitigate the influence of color bias by discarding a portion of color information from the training data. This, in turn, serves to balance the weight between color-specific features and color-agnostic features within the neural network. However, it is noteworthy that for photographic images, the presence of diverse color information enables a more nuanced focus on finer-grained details. In a different vein, Chen et al. [46] introduced a novel asymmetric disentanglement and dynamically synthesized learning approach within the Transformer framework. This approach operates with the intent to explore a shared embedding space across modalities. Notably, they introduced a dynamic updatable auxiliary sketch (A-Sketch) modality generated from the photographic modality. This auxiliary modality guides the process of asymmetric disentanglement within a singular framework. In the context of a multi-modal joint learning framework, the incorporation of this auxiliary modality imparts heightened diversity to the training samples and diminishes inter-modal disparities. Drawing upon the principles of non-exclusive transplantation, Zhang et al. [47] proposed an innovative method within the framework of a dual-path Transformer. This approach introduces a cross-compatible embedding technique, enabling cross-modal exchange at a local token level, thereby facilitating the extraction of modal-compatible features and mitigating the disparities between them. Furthermore, they introduce a scheme for constructing semantically coherent features, which serves to enhance feature recognition and greatly bolster feature robustness, ultimately attaining the most state-of-the-art performance to date. Zhu et al. [48] introduced a novel cross-domain attention mechanism that employs distinct strategies to partition feature maps within two separate branches, subsequently computing the relationships between distinct segments of sketched images and person photographs. Additionally, a cross-domain center loss was devised, surpassing the constraints of conventional center loss that necessitates domain consistency within datasets. This innovation effectively mitigates the gap between the two domains, promoting proximity among features about the same identity. Rachmadi et al. [49] leveraged three distinct target dropout regularizations, encompassing per-block dropout, horizontal per-block dropout and vertical horizontal per-block dropout. Consequently, an augmentation in the performance of deep neural network classifiers was achieved. Yuan et al. [50] devised an unbiased feature extractor aimed at mitigating the bias stemming from modality-specific information, thus enhancing the capacity of the extracted features in bridging inter-domain disparities. Furthermore, a multi-stream classifier was introduced to ensure the comprehensive attainment of intra-class consistency by the feature extractor.

The Comparison of methods based on the PKU Sketch Re-ID dataset is shown in Table 3. Notwithstanding, due to the relatively limited scale of the utilized PKU Sketch Re-ID dataset, achieving favorable recognition outcomes in authentic scenarios might prove challenging. The prospect of training and deploying models on diminutive datasets for real-world applications remains an objective for prospective advancement.

Table 3 Comparison of methods based on PKU Sketch Re-ID dataset.

Methods	Rank-1(%)	Rank-10(%)	mAP(%)
Dense-HOG+LBP+rankSVM [20]	5.1	28.3	——
Triplet SN [10]	9.0	42.6	——
GN Siamese [4]	28.9	62.4	——
AFLNet [20]	34.0	72.5	——
RCD [45]	42.5	87.5	——
LMDF [21]	49.0	80.2	——
UFE [50]	57.1	89.8	——
CDAC [48]	60.8	88.8	——
FT-SwinTrans-VHDPL [49]	73.2	99.6	72.5
SketchTrans [46]	84.6	98.2	——
IHDA [44]	85.6	98.0	——
CCSC [47]	86.0	100.0	83.7

5. Future research directions

In this section, we will provide a synopsis from the perspective of prospective applications and foundational research value, outlining several potential research directions deemed promising for the future.

In recent years, notable breakthroughs have been achieved in the domain of scene-level fine-grained sketch image retrieval, notably by works such as SketchyCOCO [51] and SceneSketcher [52]. These contributions have been instrumental in furnishing extensive datasets of scene-level fine-grained sketch images, thereby laying the foundation for the exploration of novel research avenues. Particularly, investigations into novel topics have been undertaken, including image generation predicated on scene-level fine-grained sketch image retrieval [53], as well as data retrieval anchored in fine-grained scene contexts [54]. These research trajectories hold significant practical import, as they proffer the technical feasibility of effectuating scene retrieval employing fine-grained sketch images, thereby engendering pronounced appeal among end-users.

The advent of novel data acquisition devices has substantially facilitated the collection of 3D sketch images, thereby lending substantial support to various compelling research avenues in 3D sketch image analysis. Notably, this technological advancement has paved the way for synergistic integration with immersive technologies like virtual reality (VR) [55] and augmented reality (AR) [56]. Delving into the domain of 3D sketch image investigations holds the potential to extend the horizons of human-computer interaction rooted in 2D planar touchscreens towards the expanse of 3D spatial environments, consequently affording a heightened level of immersive experience. While extant repositories of sketch image datasets and applications primarily concentrate on depictions of objects through sketch imagery, it is noteworthy that in practical scenarios, users might evince a keen interest in the machine's comprehension of a broader spectrum of sketch image concepts. Such concepts encompass but are not limited to diagrams, curves, histograms [57], maps [58], engineering sketch images [53] and prototypes of User Interfaces [54].

The advancement of deep learning in sketch image analysis is predominantly propeled by the accumulation of progressively expansive sketch image datasets. However, the exigency for manual curation in generating sketch images engenders a scarcity of such datasets in comparison to their

photographic counterparts. Consequently, the efficacy of methodologies addressing pivotal tasks encompassing sketch image analysis, ranging from recognition to retrieval, bears paramount significance. Optimal avenues for achieving this efficacy manifest through judicious approaches such as few-shot learning, self-supervised learning, or intermodal knowledge transfer originating from the domain of photographic images.

As for the future research direction, it is considered from the following aspects, and some possible solutions are given:

(1) Establishing a high-quality standard fine-grained sketch dataset more adapted to realistic environments: Currently, fine-grained sketch image datasets are relatively limited in size, and there is a significant quantitative gap compared to those datasets with millions of sketches of facial data. In addition, existing datasets lack diverse scene coverage, and data can usually be collected only under limited environmental conditions and over relatively short periods. In addition, there are significant differences between the fine-grained sketch data drawn by different artists, which significantly degrade the model performance when applied to different datasets and underperform in real-world applications. To comprehensively evaluate the robustness of the algorithm under different conditions, there is an urgent need to establish large-scale, high-quality and diverse fine-grained sketch datasets.

(2) Interpretability of deep learning for fine-grained sketch image retrieval: Even though deep learning methods have demonstrated excellent performance in fine-grained image retrieval tasks, the pursuit of higher accuracy is accompanied by a relatively limited understanding of the key factors affecting fine-grained sketch image retrieval. This leads to a lack of interpretability of the model decision-making process, making it difficult to understand in practical applications. Therefore, researchers need to delve into the inner workings [59] of deep learning models to improve the interpretability and understandability of fine-grained sketch image retrieval.

(3) Few shot learning for fine-grained sketch image retrieval: The human brain can reason from very little knowledge to unknown knowledge domains through instances, and the most advanced fine-grained sketch image retrieval algorithms nowadays are not able to have this reasoning ability. For fine-grained sketch image retrieval, large-scale instance-level dataset collection is an unsolvable problem, and because fine-grained sketch images need to be drawn by professional painters and labeled by professionals to collect, it is both time-consuming and labor-intensive to collect the data. Thus, fine-grained sketch image retrieval using few shot learning [60] will be one of the mainstream ways of future research.

(4) Semi-supervised learning for fine-grained sketch image retrieval: In semi-supervised learning, we can use a small number of samples that have been labeled and a large number of unlabeled samples for training [34], thus reducing the burden of fine-grained sketch images on the collection. This type of semi-supervised method can enhance the performance of the model and improve the accuracy of fine-grained sketch image retrieval by utilizing the features of the unlabeled data [61], the use of semi-supervised or fine-grained small number of samples sketch image retrieval research is a more realistic research significance.

(5) Cross-domain knowledge transfer: Given the substantial differences between sketch images and photographic counterparts, the exploration of cross-domain knowledge transfer, drawing insights from photographic image analysis, holds promise in advancing the capabilities of fine-grained sketch image retrieval algorithms. This approach can facilitate the transfer of expertise from the domain of photographic images to improve the performance and robustness of fine-grained sketch image retrieval methods.

It is anticipated that these research directions hold the potential to lay the foundational framework

for more efficient and enlightening methodologies. This, in turn, is expected to enhance scholarly interest in the field of fine-grained sketch image retrieval, thereby catalyzing its development.

6. Conclusions

In the past seven years, fine-grained sketch image retrieval based on deep learning has made remarkable progress, showing great research value and potential. This paper comprehensively describes the relevant concepts, problems, evaluation metrics, methods and datasets, optimal performance, and future research directions in recent years regarding deep learning-based fine-grained sketch image retrieval, aiming to provide readers with a comprehensive understanding of the current state of the art of research in this field. It is worth emphasising that all images or datasets mentioned in the paper are for academic purposes only.

The field of fine-grained sketch image retrieval based on deep learning is continuing to show vigorous vitality, and we are optimistic about its prospects. The future development of this research direction will hopefully further expand our understanding of fine-grained sketch image retrieval and provide new insights for research in related fields. As we continue our in-depth research, we encourage researchers in academia and industry to work together to promote more in-depth and reliable results in this field.

Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Author Contributions

This work was started under the suggestion of Junchao Ge. Qing Luo wrote the manuscript with support from Junchao Ge. Xiang Gao; Bo Jiang, Xueting Yan and Wanyuan Liu participated in collecting and organizing the papers involved in the survey. All authors reviewed the manuscript.

Acknowledgments (All sources of funding of the study must be disclosed)

This work was supported by the Science and Technology Project of China Southern Power Grid Co., Ltd. (No. YNKJXM20222219).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. P. Xu, T. M. Hospedales, Q. Yin, Y. Z. Song, T. Xiang, L. Wang, Deep learning for free-hand sketch: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, **45** (2022), 285–312. https://doi.org/10.1109/TPAMI.2022.3148853
2. A. K. Bhunia, P. N. Chowdhury, Y. Yang, T. M. Hospedales, T. Xiang, Y. Z. Song, Vectorization and rasterization: Self-supervised learning for sketch and handwriting, in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, (2021), 5668–5677. https://doi.org/10.1109/CVPR46437.2021.00562
3. A. Qi, Y. Gryaditskaya, J. Song, Y. Yang, Y. Qi, T. M. Hospedales, et al., Toward fine-grained sketch-based 3D shape retrieval, *IEEE Trans. Image Process.*, **30** (2021), 8595–8606. https://doi.org/10.1109/TIP.2021.3118975
4. P. Sangkloy, N. Burnell, C. Ham, J. Hays, The sketchy database: Learning to retrieve badly drawn bunnies, *ACM Trans. Graphics*, **35** (2016), 1–12. https://doi.org/10.1145/2897824.2925954
5. Q. Yu, F. Liu, Y. Z. Song, T. Xiang, T. M. Hospedales, C. C. Loy, Sketch me that shoe, in *Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 799–807. https://doi.org/10.1109/CVPR.2016.93
6. Y. Cao, C. Wang, L. Zhang, L. Zhang, Edgel index for large-scale sketch-based image search, in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 761–768. https://doi.org/10.1109/CVPR.2011.5995460
7. Y. Cao, H. Wang, C. Wang, Z. Li , L. Zhang, L. Zhang, Mindfinder: Interactive sketch-based image search on millions of images, in *Proceedings of The 18th ACM International Conference on Multimedia (MM'10)*, (2010), 1605–1608. https://doi.org/10.1145/1873951.1874299
8. M. Eitz, K. Hildebrand, T. Boubekeur, M. Alexa, Sketch-based image retrieval: Benchmark and bag-of-features descriptors, *IEEE Trans. Visualization Comput. Graphics*, **17** (2011), 1624–1636. https://doi.org/10.1109/TVCG.2010.266
9. J. Collomosse, T. Bui, M. Wilber, C. Fang, H. Jin, Sketching with style: Visual search with sketches and aesthetic context, in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 2679–2687. https://doi.org/10.1109/ICCV.2017.290
10. Y. Li, T. M. Hospedales, Y. Song, S. Gong, Fine-grained sketch-based image retrieval by matching deformable part models, in *The British Machine Vision Conference(BMVC)*, (2014).
11. J. Song, Q. Yu, Y. Z. Song, T. Xiang, T. M. Hospedales, Deep spatial-semantic attention for fine-grained sketch-based image retrieval, in *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 5551–5560. https://doi.org/10.1109/ICCV.2017.592
12. J. Zhang, F. Shen, L. Liu, F. Zhu, M. Yu, L. Shao, et al., Generative domain-migration hashing for sketch-to-image retrieval, in *Proceedings of the 2018 European Conference on Computer Vision (ECCV)*, (2018), 297–314. https://doi.org/10.1007/978-3-030-01216-8_19
13. J. Song, Y. Z. Song, T. Xiang, T. M. Hospedales, Fine-Grained image retrieval: The text/sketch input dilemma, in *The British Machine Vision Conference(BMVC)*, (2017).
14. A. Sain, A. K. Bhunia, Y. Yang, T. Xiang, Y. Song, Cross-modal hierarchical modelling for fine-grained sketch based image retrieval, preprint, arXiv: 2007.15103 .
15. S. L. Yan, Y. F. Zhang, M. H. Xie, D. C. Zhang, Z. T. Yu, Cross-domain person re-identification with pose-invariant feature decomposition and hypergraph structure alignment, *Neurocomputing*, **467** (2022), 229–241. https://doi.org/10.1016/j.neucom.2021.09.054

16. H. Li, M. Liu, Z. Hu, F. Nie, Z. Yu, Intermediary-guided bidirectional spatial-temporal aggregation network for video-based visible-infrared person re-identification, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 4962–4972. https://doi.org/10.1109/TCSVT.2023.3246091
17. H. Li, K. Xu, J. Li, Z. Yu, Dual-stream reciprocal disentanglement learning for domain adaptation person re-identification, *Knowl. Based Syst.*, **251** (2022), 109315. https://doi.org/10.1016/j.knosys.2022.109315
18. S. Wang, R. Liu, H. Li, G. Qi, Z. Yu, Occluded person re-identification via defending against attacks from obstacles, *IEEE Trans. Inf. Forensics Secur.*, **18** (2022), 147–161. https://doi.org/10.1109/TIFS.2022.3218449
19. H. Li, N. Dong, Z. Yu, D. Tao, G. Qi, Triple adversarial learning and multi-view imaginative reasoning for unsupervised domain adaptation person re-identification, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2021), 2814–2830. https://doi.org/10.1109/TCSVT.2021.3099943
20. L. Pang, Y. Wang, Y. Z. Song, T. J. Huang, Y. H. Tian, Cross-domain adversarial feature learning for sketch re-identification, in *Proceedings of the 26th ACM international conference on Multimedia (MM'18)*, (2018), 609–617. https://doi.org/10.1145/3240508.3240606
21. S. Gui, Y. Zhu, X. Qin, X. Ling, Learning multi-level domain invariant features for sket ch re-identification, *Neurocomputing*, **403** (2020), 294–303. https://doi.org/10.1016/j.neuco m.2020.04.060
22. D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in *Proceedings of the IEEE international workshop on performance evaluation for tracking and surveillance (PETS)*, (2007), 1–7.
23. L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, (2015), 1116–1124. https://doi.org/10.1109/ICCV.2015.133
24. Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, J. R. Smith, Learning locally-adaptive decision functions for person verification, in *Proceedings of the 2013 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2013), 3610–3617. https://doi.org/10.1109/CVPR.2013.463
25. R. Kushwaha, N. Nain, PUG-FB: Person-verification using geometric and Haralick featuresof footprint biometric, *Multimedia Tools Appl.*, **79 (**2020), 2671–2701. https://doi.org/10.1007/s11042-019-08149-0
26. K. Pang, Y. Yang, T. M. Hospedales, T. Xiang, Y. Z. Song, Solving mixed-modal jigsaw puzzle forfine-grained sketch-based image retrieval, in *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020) 10347–10355. https://doi.org/10.1109/CVPR42600.2020.01036
27. F. Radenovic, G. Tolias, O. Chum, Deep shape matching, in *Proceedings of the 2018 European Conference on Computer Vision (ECCV)*, (2018), 751–767. https://doi.org/10.1007/978-3-030-01228-1_46
28. X. Lin, Y. Duan, Q. Dong, J. Lu, J. Zhou, Deep variational metric learning, in *Proceedings of the 2018 European Conference on Computer Vision (ECCV)*, (2018), 689–704. https://doi.org/10.1007/978-3-030-01267-0_42

29. J. Xu, H. Sun, Q. Qi, J. Wang, C. Ge, L. Zhang, et al., DlA-Net for FG-SBIR: Dynamic local aligned network for fine-grained sketch-based image retrieval, in *Proceedings of the 29th ACM international conference on Multimedia (MM'21)*, (2021), 5609–5618. https://doi.org/10.1145/3474085.3475705
30. H. Sun, J. Xu, J. Wang, Q. Qi, C. Ge, J. Liao, Dli-net: Dual local interaction network for fine-grained sketch-based image retrieval, *IEEE Trans. Circuits Syst. Video Technol.*, **32** (2022), 7177–7189.https://doi.org/10.1109/TCSVT.2022.3171972
31. Z. Zhang, Z. Xie, Z. Chen, Y. Han, X. Luo, X. Xu, Expansion window local alignment weighted network for fine-grained sketch-based image retrieval, *Pattern Recognit.*, **144** (2023), 109892. https://doi.org/10.1016/j.patcog.2023.109892
32. Z. Ling, Z. Xing, J. Li, L. Niu, Multi-level region matching for fine-grained sketch-based image retrieval, in *Proceedings of the 30th ACM international conference on Multimedia (MM'22)*, (2022), 462–470. https://doi.org/10.1145/3503161.3548147
33. K. Pang, K. Li, Y. Yang, H. Zhang, T. M. Hospedales, T. Xiang, et al., Generalising fine-grained sketch-based image retrieval, in *Proceedings of the 2019 IEEE/CVF Conference onComputer Vision and Pattern Recognition (CVPR)*, (2019), 677–686. https://doi.org/10.1109/CVPR.2019.00077
34. A. Sain, A. K. Bhunia, Y. Yang, T. Xiang, Y. Z. Song, Stylemeup: Towards style-agnostic sketch-based image retrieval, in *Proceedings of the 2021 IEEE/CVF Conference on ComputerVision and Pattern Recognition (CVPR)*, (2021), 8504–8513. https://doi.org/10.1109/CVPR46437.2021.00840
35. Z. Ling, Z. Xing, J. Zhou, X. Zhou, Conditional stroke recovery for fine-grained sketch-based image retrieval, in *Proceedings of the 2022 European Conference on Computer Vision (ECCV)*, (2022), 722–738. https://doi.org/10.1007/978-3-031-19809-0_41
36. A. K. Bhunia, P. N. Chowdhury, A. Sain, Y. Yang, T. Xiang, Y. Z. Song, More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval, in *Proceedings of the 2021 IEEE/CVF Conference on ComputerVision and Pattern Recognition (CVPR)*, (2021), 4247–4256. https://doi.org/10.1109/CVPR46437.2021.00423
37. A. K. Bhunia, A. Sain, P. H. Shah, A. Gupta, P. N. Chowdhury, T. Xiang, et al., Adaptive fine-grained sketch-based image retrieval, in *Proceedings of the 2022 European Conference on Computer Vision (ECCV)*, (2022), 163–181. https://doi.org/10.1007/978-3-031-19836-6_10
38. D. Ha, D. Eck, A neural representation of sketch drawings, preprint, arXiv: 1704.03477.
39. U. R. Muhammad, Y. Yang, Y. Z. Song, T. Xiang, T. M. Hospedales, Learning deep sketch abstraction, in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 8014–8023. https://doi.org/10.1109/CVPR.2018.00836
40. A. K. Bhunia, Y. Yang, T. M. Hospedales, T. Xiang, Y. Z. Song, Sketch less for more: On-the-fly fine-grained sketch-based image retrieval, in *Proceedings of the 2020 IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 9779–9788. https://doi.org/10.1109/CVPR.2018.00836
41. D. Wang, H. Sapkota, X. Liu , Q. Yu, Deep reinforced attention regression for partial sketch based image retrieval, in *Proceedings of the 2021 IEEE International Conference on Data Mining (ICDM)*, (2021), 669–678. https://doi.org/10.1109/CVPR.2018.00836

42. D. Dai, X. Tang, Y. Liu, S. Xia, G. Wang, Multi-granularity association learning for on-the-fly fine-grained sketch-based image retrieval, *Knowl. Based Syst.*, **253** (2022), 109447. https://doi.org/10.1016/j.knosys.2022.109447
43. A. K. Bhunia, S. Koley, A. F. U. R. Khilji, A. Sain, P. N. Chowdhury, T. Xiang, Sketching without worrying: Noise-tolerant sketch-based image retrieval, in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 999–1008. https://doi.org/10.1109/CVPR52688.2022.00107
44. F. Yang, Y. Wu, Z. Wang, X. Li, S. Sakti, S. Nakamura, Instance-level heterogeneous domain adaptation for limited-labeled sketch-to-photo retrieval, *IEEE Trans. Mult.*, **23** (2020), 2347–2360. https://doi.org/10.1109/TMM.2020.3009476
45. Y. Gong, L. Huang, L. Chen, Eliminate deviation with deviation for data augmentation and a general multi-modal data learning method, preprint, arXiv: 2101.08533.
46. C. Chen, M. Ye, M. Qi, B. Du, Sketch transformer: Asymmetrical disentanglement learning from dynamic synthesis, in *Proceedings of the 30th ACM international conference on Multimedia (MM'22)*, (2022), 4012–4020. https://doi.org/10.1145/3503161.3547993
47. Y. Zhang, Y. Wang, H. Li, S. Li, Cross-compatible embedding and semantic consistent feature construction for sketch re-identification, in *Proceedings of the 30th ACM International Conference on Multimedia (MM'22)*, (2022), 3347–3355. https://doi.org/10.1145/3503161.3548224
48. F. Zhu, Y. Zhu, X. Jiang, J. Ye, Cross-domain attention and center loss for sketch re-identification, *IEEE Trans. Inf. Forensics Securit.*, **17** (2022), 3421–3432. https://doi.org/10.1109/TIFS.2022.3208811
49. R. F. Rachmadi, S. M. S. Nugroho, I. K. E. Purnama, Revisiting dropout regularization for cross-modality person re-identification, *IEEE Access*, **10** (2022), 102195–102209. https://doi.org/10.1109/ACCESS.2022.3208562
50. B. Yuan, B. Chen, Z. Tan, X. Shao, B. K. Bao, Unbiased feature enhancement framework for cross-modality person re-identification, *Multimedia Syst.*, **28** (2022), 749–759. https://doi.org/10.1007/s00530-021-00872-9
51. C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, C. Zou, Sketchycoco: Image generation from freehand scene sketches, in *Proceedings of the 2020 IEEE/CVF Conference on ComputerVision and Pattern Recognition (CVPR)*, (2020), 5174–5183. https://doi.org/10.1109/CVPR42600.2020.00522
52. F. Liu, C. Zou, X. Deng, R. Zuo, Y. K. Lai, C. Ma, et al., Scenesketcher: Fine-grained image retrieval with scene sketches, in *Proceedings of the 2020 European Conference on Computer Vision (ECCV)*, (2020),718–734. https://doi.org/10.1007/978-3-030-58529-7_42
53. K. D. D. Willis, P. K. Jayaraman, J. G. Lambourne, H. Chu, Y. Pu, Engineering sketch generation for computer-aided design, in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2021), 2105–2114. https://doi.org/10.1109/CVPRW53098.2021.00239
54. V. Jain, P. Agrawal, S. Banga, R. Kapoor, S. Gulyani, Sketch2Code: Transformation of sketches to UI in real-time using deep neural network, preprint, arXiv: 1910.08930.

55. D. Giunchi, S. James, D. Degraen, A. Steed, Mixing realities for sketch retrieval in virtual reality, in *Proceedings of the 17th International Conference on Virtual-Reality Continuum and its Applications in Industry(VRCAI'19)*, (2019). https://doi.org/10.1145/3359997.3365751
56. B. Jackson, D. F. Keefe, Lift-off: Using reference imagery and freehand sketching to create 3D models in VR, *IEEE Trans. Visualization Comput. Graphics*, **22** (2016), 1442–1451. https://doi.org/10.1109/TVCG.2016.2518099
57. J. C. Roberts, C. Headleand, P. D. Ritsos, Sketching designs using the five design-sheet methodology, *IEEE Trans. Visualization Comput. Graphics*, **22** (2015), 419–428. https://doi.org/10.1109/TVCG.2015.2467271
58. F. Boniardi, A. Valada, W. Burgard , G. D. Tipaldi, Autonomous indoor robot navigation using a sketch interface for drawing maps and routes, in *Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA)*, (2016), 2896–2901. https://doi.org/10.1109/ICRA.2016.7487453
59. F. Lin, M. Li, D. Li, T. Hospedales, Y. Z. Song, Y. Qi, Zero-shot everything sketch-based image retrieval, and in explainable style, in *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, (2023), 23349–23358. https://doi.org/10.1109/CVPR52729.2023.02236
60. X. S. Wei, Y. Z. Song, O. M. Aodha, J. Wu, Y. Peng, J. Tang, et al., Fine-grained image analysis with deep learning: A survey, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021) ,8927–8948. https://doi.org/10.1109/TPAMI.2021.3126648
61. A. Sain, A. K. Bhunia, S. Koley, P. N. Chowdhury, S. Chattopadhyay, T. Xiang, et al., Exploiting unlabelled photos for stronger fine-grained SBIR, in *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR)*, (2023), 6873–6883. https://doi.org/10.1109/CVPR52729.2023.00664

AIMS Press