Bi-LSTM Sequence Modeling for on-the-fly Fine-Grained Sketch-based Image Retrieval

Yingge Liu, Dawei Dai*, Xiaoyu Tang, Shuyin Xia, Guoyin Wang

Abstract—Fine grained sketch-based image retrieval (FG-SBIR) addresses the problem of retrieving a particular photo for a given query sketch. However, its widespread applicability is limited by the fact that it is difficult to draw a complete sketch, and the drawing process often takes more time than the text/tag method. On-the-fly FG-SBIR was proposed to address this problem, in which image retrieval is performed after each stroke. The aim is to retrieve the target photo using the least number of strokes. Each photo corresponds to a sketch drawing episode, in which a significant correlation exists between these incomplete sketches. This correlation will allow a more efficient learning embedding space for incomplete sketches, which is considered in this study. First, a triplet network, as used in the classical FG-SBIR framework, was designed to learn the joint embedding space shared between the photo and its corresponding complete sketch. Second, assuming strong time correlation, each sketch drawing episode is considered a sequence, and each incomplete sketch in the drawing episode is extracted as a feature vector. A learnable Bi-LSTM module and triplet loss function map the feature space of incomplete sketches obtained from the base model for efficient representation. In the experiments, we proposed more realistic challenges, and our method achieved superior early retrieval efficiency over the state-of-the-art baseline methods on two publicly available fine-grained sketch retrieval datasets.

Impact Statement — Widespread applicability of FG-SBIR is limited by the fact that drawing a complete sketch needs skill and takes more time. Recently, on-the-fly FG-SBIR was proposed to address this problem, in which image retrieval is performed after each stroke. In this paper, we considered a strong correlation exists among a sketch drawing episode generated by one photo; We are fully considering this correlation to learn the more efficient embedding space for the incomplete sketches, we first regard the a sketch drawing episode as a sequence, and then a learnable Bi-LSTM module and triplet loss function were designed to model the correlation. Experiments indicates that our method achieved superior early retrieval efficiency over the state-of-the-art baseline methods on two publicly available fine-grained sketch retrieval datasets.

Index Terms—Fine-grained SBIR, BiLstm, triplet loss

I. Introduction

Recently, with the rapid proliferation of various electronic touch screen devices, which makes hand-painted graphics (sketches) increasingly popular in people's daily lives, works, and entertainment, and thus more and more hand-painted data appear on the Internet. Different from text and labels, sketch is an abstract expression of what human beings see, which can effectively convey detail information that is difficult for the way of text and label. Consequently, sketch-related problems have become a topic of considerable research interest in the field of computer vision [1],[2]. Considering these fields, fine-grained sketch-based image retrieval (FG-SBIR) [3], [4], [5] [6], [7] has received particular attention because of its potential commercial applications, which address the problem of retrieving a particular photo for a given query sketch.

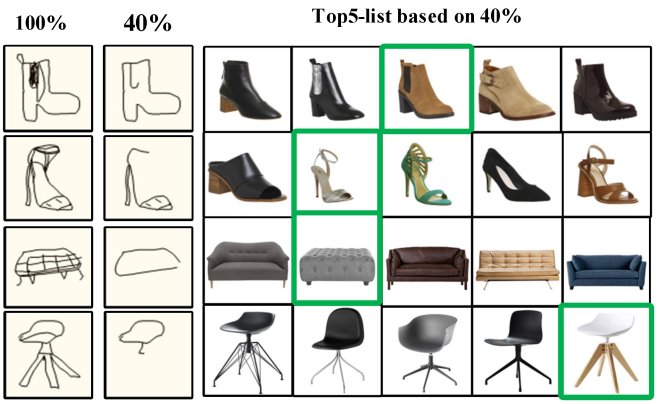


Fig. 1. Demonstrating our methods' ability to retrieve (top-5 list) the target photo using fewer strokes (approximately 40%).

Although the FG-SBIR method has made significant progress over the years, its practical application is hindered for the following reasons. First, sketching is time-consuming, often slower than clicking a tag for typing the keyword. Second, drawing a complete sketch requires skills. A new FG-SBIR method [8] with an on-the-fly setting was first proposed to overcome these problems.This method retrieves the target photo from an incomplete sketch with as few strokes as possible, which can be considered a cross-modal matching problem. Such a problem is solved by learning in a joint embedding space[9], where the feature vector is shared between the target photo and its sketch modalities[10]. Consequently, learning the efficient joint embedding space for incomplete or poorly drawn sketches and their corresponding target photo is the key in the on-the-fly FG-SBIR method[11].

In on-the-fly FG-SBIR, one photo can create a drawing episode containing many incomplete sketches[12]. The incomplete sketches in the drawing episode are not independent of each other and have a significant correlation. Therefore, we considered this correlation for learning the efficient embedding

This work was sponsored by Natural Science Foundation of Chongqing (No. cstc2019jcyj-msxmX0380) and China Postdoctoral Science Foundation (2021M700562).

Yingge Liu, Dawei Dai(Corresponding Author), Xiaoyu Tang, Shuyin Xia, Guoyin Wang are with College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 400065, China (email: dw dai@163.com).

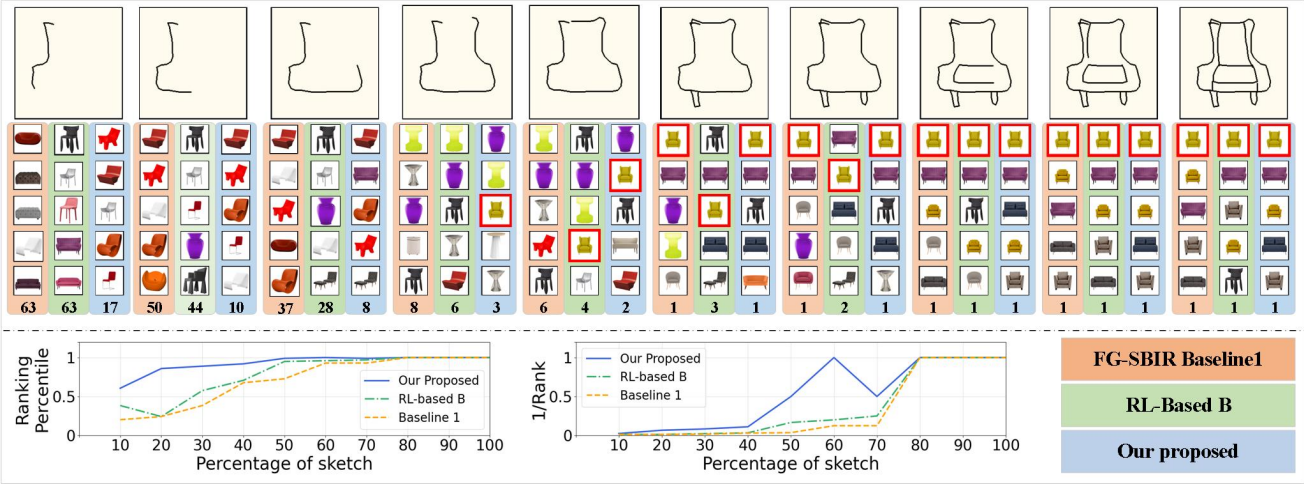


Fig. 2. Illustration of our proposed on-the-fly framework over two baselines. One is an FG-SBIR baseline (B1) (Szegedy et al. 2016) trained with completed sketches, and another is a reinforcement learning-based method (RL-based B) (Bhunia et al. 2020) trained with incomplete sketches. In this example, our method needs only 20% of the complete sketch to include the true match in the top-10 rank list, compared to 40% for RL-based B and B1 methods. The top-5 photo images retrieved by the three methods are also shown here. The number at the bottom denotes the paired (true match) photo's rank at every stage.

space for incomplete sketches. In this work, we first designed base model that a triplet network comprising a pre-trained neural model and learnable attention module to learn joint embedding sharing between the photo and its corresponding complete sketch, through which we can obtain the low-dimensional vector of photo and complete sketch that used to final retrieval, such base model was also used to extract the feature vector for the incomplete sketches; Second, assuming a strong temporal correlation between drawing episodes, we designed a learnable Bi-LSTM module and a triplet loss function after the attention module to learn better representations of incomplete sketches. We applied our method to two publicly available fine-grained sketch retrieval datasets and achieved superior early retrieval efficiency. As shown in Fig. 1, the target photo appears in the top-5 retrieval list with 40% strokes; practically, users can stop the search when the target photos appear in the result list.

Our contributions can be summarized as follows. (a) To learn an efficient representation for incomplete sketches, we regard the learning of drawing episodes as a sequence problem and propose an independent LSTM module to avoid the diversity of incomplete sketches confusing the base model. (b) We proposed several new realistic challenges with on-the-fly FG-SBIR framework and performed extensive experiments on the two public datasets to demonstrate the superiority of our proposed method.

II. Related Work

A. Category-level SBIR

In this type of retrieval task, if the retrieved photo and query sketch belong to the same category, the retrieval is considered successful. Previous methods usually used the manual descriptors to construct global or local joint representations for the photos and their associated sketches [13], [14]. Such as SIFT [15], HOG [16], edge local direction histograms [17], and learning key shapes [18]. In the recent years, deep neural network models have been used to extract more effective features. For example, Yu et al [19] proposed a multi-scale channel neural network that optimized the sketch based on the feature information in the sketch. Subsequently, classical ranking losses, such as contrast or triplet loss , have been used in SBIR problems to narrow the embedding space of sketch-photo pairs. To deal with the problem of large-scale image retrieval, researchers have highlighted many methods for constructing a hash index around the basic idea of a hash. For example, Liu et al [20] proposed a semi-heterogeneous deep architecture (deep sketch hashing), which aimed to reduce the computational cost of retrieval; Song et al [21] proposed an edge-guided cross-domain learning method to address the problem of mapping the sketch and image domains to the common space domain.

B. Fine-grained SBIR

Different from category-level SBIR, each instance can be regarded as a category FG-SBIR, which is more challenging comparing with category-level SBIR[22]. Yu et al [19] constructed a neural model using a triple ranking loss and combining the edge image and sketch into a mapping pair, which improved the efficacy of spatial mapping. Zhang et al [23] realized end-to-end image retrieval using heterogeneous networks to deal with the problem of information loss when extracting features from edge images. Wang et al [24] proposed a deep cascaded cross-modal ranking model that exploits all the beneficial multi-modal information in sketches and annotated images. Peng et al [6] proposed a unsupervised learning approach to model a universal manifold of prototypical visual sketch traits as a domain generalization problem to identify cross-category generalization for FG-SBIR. Zhu et al [25] designed a gradually focused bi-linear attention model to extract detailed information more effectively. Sain et al [26] proposed a cross-modal network method with hierarchical co-attention, and then uses cross-modal VAE and meta-learning to realize style agnostic SBIR. Du et al [27] proposed a progressive training strategy effectively to integrate features with different granularity. In the latest research by

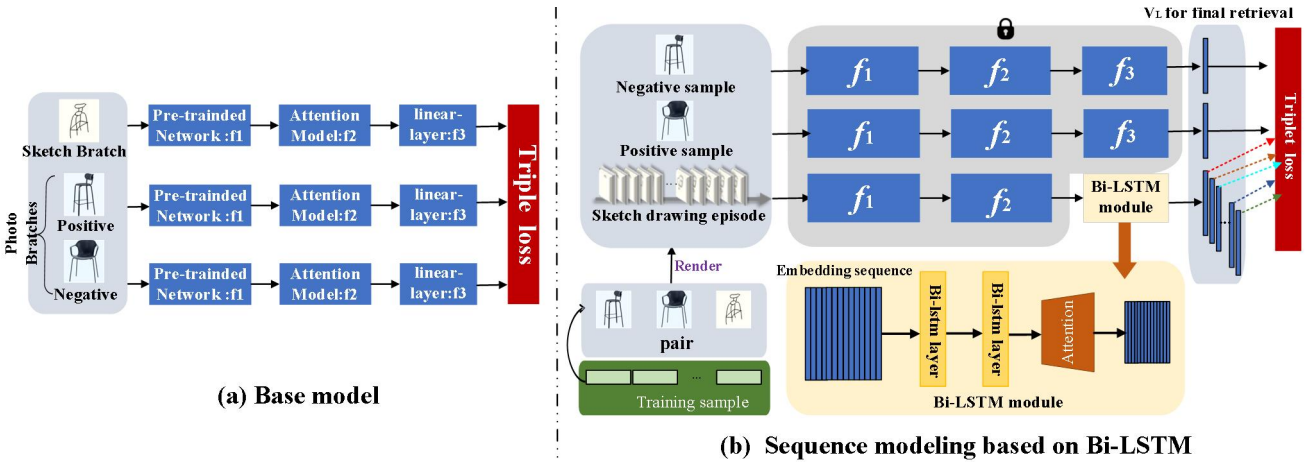


Fig. 3. Overview of our approach. (a) Base model: a classical FG-SBIR framework; (b) Our proposed for on-the-fly FG-SBIR. The locks signify that the weights are fixed during learning.

Choi et al [28], Huang F et al [29], Collomosse et al [30], and Bhunia et al [8], an interactive framework was proposed to relieve sketch retrieval from being limited to static search.

C. *Incomplete Sketch*

Studies on incomplete sketch [30][31][32][33][34][35] have mainly focused on sketch synthesis and completion. The efficacy of the attention mechanism [36] and VAE [37]has been verified for sketch completion. Ha and Eck [38] proposed a sketch-RNN to generate sketches automatically of specified types based on pen strokes. Moreover, Chen et al[39] used a CNN to replace the Bi-LSTM encoder, which performed better than the latter in generating multiple types of sketches. Cao et al [40] proposed a CNN-based auto-encoder to capture pixel-level positional information of strokes to generate high-quality sketches and used condition vectors as distinguishers to improve the performance on multi-category sketches. Liu et al [41]uses conditional GANs to complete sketches at the image-to-image level to assist with recognition. Aksan et al[42] proposed a generative model for complex free-form structures, which regards drawings as a collection of strokes that form complex structures (e.g., flow charts), and can simulate the appearance of individual strokes and the compositional structure of larger diagrammatic drawings. Lin et al [43]proposed a bidirectional encoder representation from a transformer model to retrieve images from sketches. Ghosh et al [44]constructed a differentiable neural renderer to render strokes to improve the quality of recreated images; Bhunia et al proposed a new framework known as the on-the-fly FG-SBIR, in which the retrieval was conducted at every stroke drawn and a reinforcement learning-based method was used to optimize the embedding space of incomplete sketches, thereby improving on earlier retrieval performance.

III. Methodology

We consider the on-the-fly FG-SBIR as a sequence optimization problem and propose combining CNN and Bi-LSTM to address the sketch drawing episode. By learning efficient embedding spaces for the incomplete sketch, we expect to retrieve the target photo at the earliest stroke possible (see Fig. 2). An overview of our proposed model is shown in Fig. 3. We first train the state-of-the-art FG-SBIR model [45], [8] as the base model using a triplet loss function. Next, we maintain the photo branch and learn the sketch branch using a Bi-LSTM module and triplet loss function for the incomplete sketch sequence.

Formally, our base model learns an embedding function F that maps a complete sketch s and its target photo x to a d-dimensional feature vector v_L for the final retrieval, i.e., we obtain a list of vectors $V_L=\{F(x_i)\}_{i=1,\ldots,n}$ from a given gallery of n photos. For a given query of incomplete sketch s, we obtain its embedding vector using the proposed method (see Fig. 3(b)) and obtain the top-k retrieved photos from V_L based on the pairwise distance metric. If the target photo first appears in the top-k list at the current stroke, we consider the top-k accuracy true for that sketch. Because one photo can create a series of sketches, the sketch rendering operation proposed by Bhunia et al [8] is used to produce rasterized sketches.

A. *Base model*

First, we designed a neural model to learn the joint embedding space for the photo and its complete sketch. To verify the effectiveness of our framework, we use a triplet network [8], having three CNN model branches with shared weights corresponding to the input images, including a positive photo, a query sketch (complete one), and a negative photo, as shown in Fig. 3(a). The first CNN branch is a pre-trained neural model f1, which extracts the features of the input images. The InceptionNet [45] model trained on ImageNet is used as the function f1 (see Eq. (1)), where x and B indicate the input image and its corresponding feature maps, respectively. The second branch f2 allows learning the embedding vector of the complete sketch and its target photo, as shown in Eq. (2); two attention mechanisms (f_{att}), including spatial and channel [2] attention, can be used. The third branch f3 maps the high-dimensional vectors (VH) to low-dimensional vectors (V_L) for the final retrieval, using a simple linear mapping (A) in Eq. (3). We designed a triplet loss function to learn the joint embedding space shared between photos and sketches, and consider only the complete sketch in the base model. The terms

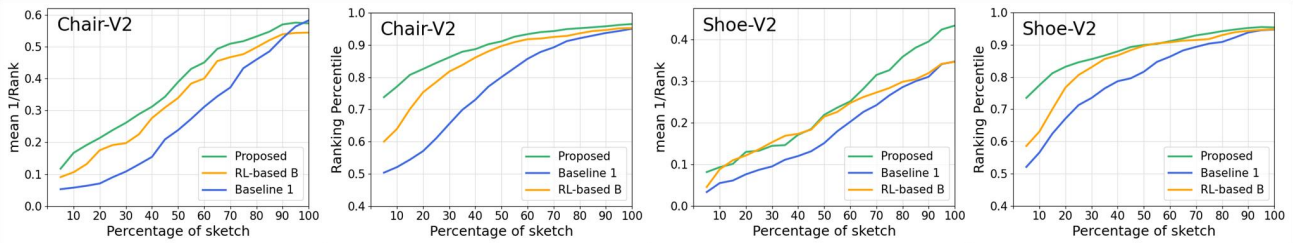


Fig. 4. Performance of our proposed method. Rather than showing the complete sketch (T = 20 sketch rendering steps), we visualize it as a percentage of the sketch. A higher value indicates a better early retrieval performance.

v_{SL} and v_p (v_n) in Eq. (4) indicate the low-dimensional vector of the complete sketch and target photo (non-target photo), respectively, obtained using Eq. (3).

$$B = f_1(x) \quad (1)$$

$$V_H = f_2(B) = Global\_pooling(B + B \cdot f_{att}(B)) \quad (2)$$

$$V_L = A \cdot V_H \quad (3)$$

$$J1(\theta) = max\left(\sum_0^n (d(v_{sL}, v_p) - d(v_{sL}, v_n) + \alpha, 0)\right) \quad (4)$$

B. Bi-LSTM module for the sketch drawing episode

Due to the creation of multiple incomplete sketches of each photo, a diversity that can confuse the base network is created. The base model only learns the shared embedding space between the photo and its complete sketch. Therefore, we designed a learnable module to optimize embedding spaces for incomplete sketches. For a given target photo, all incomplete sketches are part of its complete sketch, and a strong correlation exists between these incomplete sketches. Hence, the drawing episode was considered as a sequence optimization problem.

As shown in Fig. 3(b), we designed a Bi-LSTM module to learn the temporal correlation in the drawing episode and optimize the embedding space. First, we train the base model to obtain the feature vector V_{HS} (high-dimensional, see Eq.(2)) for all incomplete sketches ($S = \{s_i\}$, i=1,2,...,T) in a drawing episode; second, we consider $\{V_{HS}[i]\}$ as a temporal sequence, and design a Bi-LSTM module (including several Bi-LSTM layers) and a triplet loss function (see Eq. (4)) to learn the embedding space of the incomplete sketch as close to its target photo (v_p). Here, the term v_{SL} indicates the low-dimensional vector of the incomplete sketch in a drawing sketch obtained Eq. (5); $\{V_H[i]\}$ indicates the sequence vector of a drawing sketch obtained using Eq. (2).

$$V_{SL} = f_{lstm}(\{V_H[i]\}), i = 0,1,2,\ldots,T \quad (5)$$

IV. Experiments

A. Dataset

We used QMUL-Shoe-V2 [46],[4],[6] and QMUL-Chair-V2 [4] datasets that were designed for FG-SBIR to train our base model, and their rasterized sketch images that had been specifically designed for the on-the-fly FG-SBIR problem were used to train our models and to evaluate their retrieval performance over different stages of a complete sketch drawing episode. QMUL-Shoe-V2 contained 6730 sketches and 2000 photos, of which 6051 and 1800, respectively, were used to train the models, and the rest were used to test our model. QMUL-Chair-V2 contained 2000 sketches and 400 photos, of which 1275 and 300 were used to train our model, and the rest were used to test our model.

B. Implementation details

We implemented our model in PyTorch on a 40 GB Nvidia A100 GPU. We adopted the Inception-V3 [45] network pre-trained on ImageNet datasets [47] as the backbone network to extract the features for both the photos and sketches. The channel of the extracted feature map for the sketch branch was 2048, and the hidden state of the sketch branch through the Bi-LSTM layer was 1024. For training the backbone network, we used the triplet loss function with a margin of 0.3. Next, we render the sketch image with T = 20 steps, keep f1 and f2 fixed to use the triplet loss function, and train the Bi-LSTM of the sketch branch for 500 epochs. Our retrieval model uses the Adagrad optimizer, with an initial learning rate of 0.01, a mini-batch size of 20 (Shoe-V2 is 120), and the embedding space dimension D of 64.

C. Evaluation Metric

Regarding the frame of the on-the-fly FG-SBIR, we prioritize the target photo appearing at the top of the list. Thus, the percentage of sketches with true match photos appearing in the top-q list (Acc.@q accuracy) was used to quantify the performance. Moreover, m@A (the ranking percentile) and m@B (1/rank versus percentage of sketch) (Bhunia et al. 2020) were used to capture the early retrieval performance for the incomplete sketches. Considering this context, a higher value of m@A and m@B indicates a better performance during the early sketch retrieval.

D. Baseline Methods

In 2020, Bhunia et al. proposed an on the fly FG_SBIR framework focusing on early retrieval, which uses reinforcement learning methods to optimize incomplete retrieval. Thus, based on previous studies, we choose the RL-based B method and some existing FG-SBIR baselines applied to the problem of on-the-fly FG-SBIR to verify the contribution of our proposed method. Five FG-SBIR baselines (B1, B2, B3, B4, and TS) mentioned in [8] were also selected for our experiments.

B1: The framework is the same with our base model [41, 49] that trained only with a triplet loss using complete sketch (Fig.3(a)). B2: The framework is also the same with our base model, but trained on all intermediate sketches as training data. B3: 20 models (the same with base model) were trained on every percentage of

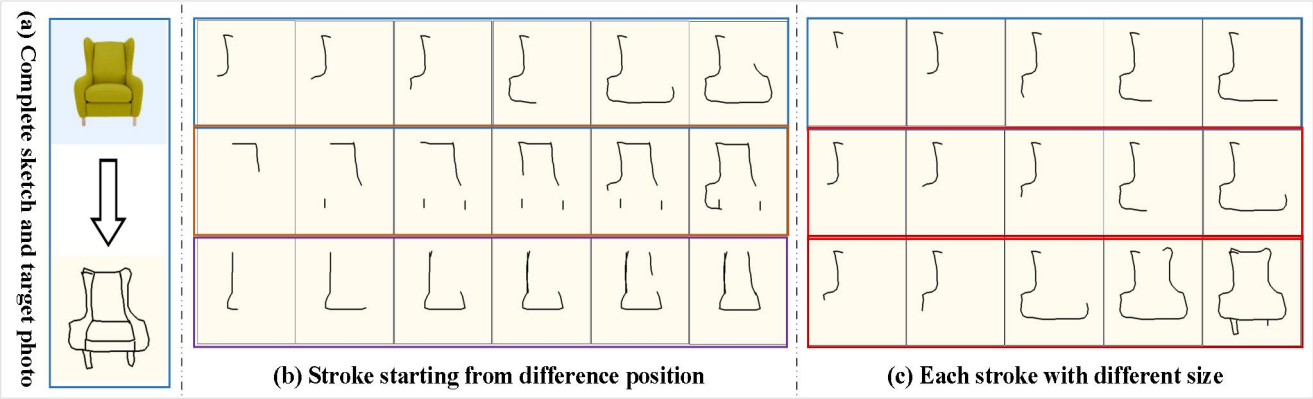


Fig. 5. Different painting styles among users may affect the model performance. (a) One complete sketch and its target photo; (b) Different sketch drawing episodes for one target photo; (c) Each stroke drawn in different sizes.

TABLE I
COMPARATIVE RESULTS OF THE PROPOSED METHOD WITH DIFFERENT BASELINE METHODS (* INDICATES CITING DATA FROM BHUNIA)

	Chair-V2				Shoe-V2			
	m@A	m@B	A@5	A@10	m@A	m@B	A@5	A@10
B1*	77.18	29.04	76.47	88.13	80.12	18.05	65.69	**79.69**
B2*	80.46	28.07	74.31	86.69	79.72	18.75	61.79	76.64
B3*	76.99	30.27	76.47	88.13	80.13	18.46	65.69	79.69
B4*	81.24	29.85	75.14	87.69	81.02	19.50	62.34	77.24
TS*	76.01	27.64	73.47	85.13	77.12	17.13	62.67	76.47
RL-based B*	85.44	35.09	76.34	89.65	85.38	21.44	**65.77**	79.63
Ours	**89.54**	**38.57**	**79.87**	**91.33**	**88.70**	**24.00**	62.16	77.02

TABLE II
COMPARATIVE RESULTS WITH VARYING FEATURE-EMBEDDING.

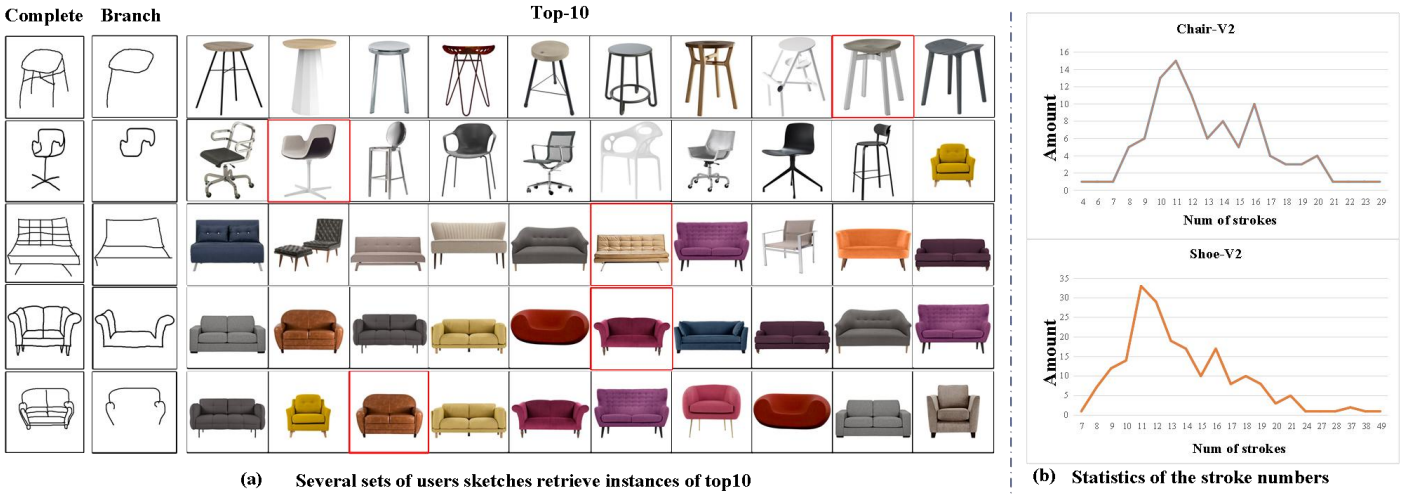
	Chair-V2				Shoe-V2			
	m@A		m@B		m@A		m@B	
	RL-B*	Ours	RL-B*	Ours	RL-B*	Ours	RL-B*	Ours
32	82.61	**87.86**	**34.67**	33.08	82.94	**88.35**	19.61	**23.98**
64	85.44	**89.54**	35.09	**38.57**	85.38	**88.70**	21.44	**24.00**
96	-	**88.47**	-	**35.49**	-	**89.40**	-	**25.50**
128	84.71	**88.71**	34.49	**35.71**	84.61	**89.31**	20.81	**25.90**
256	81.39	**89.19**	31.37	**36.75**	80.69	**87.59**	19.68	**20.42**

sketches (such as 5%, 10%, ..., 100%) respectively. B4: Bhunia et al. impose combination of triplet loss and ranking loss that approximated by a generalized deep network at T different instants of the sketch. TS: For a given incomplete sketch, a image-to-image translation model [15] was first used to generate the complete sketch, and then it is fed to an baseline model for photo retrieval. RL-based B: Bhunia et al. adopt reinforcement learning method to optimize the embedding space of the incomplete sketches that obtained from the a CNN model by a ranking loss.

E. Performance Analysis

The performance of our proposed method on the problem of on-the-fly SBIR was shown in Fig.4 against the baseline methods (B1 and RL-based B). We can note that (1) All methods can perform better and better with the gradual completeness of the incomplete sketch; (2) Our proposed and RL-based B perform better in the early retrieval than that of B1. This is because no mechanism in B1 is designed to optimize the incomplete sketches or early retrieval. (3) The performance of our proposed at early retrieval is significantly improved against with that of RL-based B; Compared to the B1, our method designs a special module to learn the efficient embedding space for incomplete sketches. Moreover, considering the RL-based method, we optimize the embedding space of incomplete sketches by differentiable Bi-LSTM module, instead of the non-differentiable ranking loss

In addition to the five baselines (B1, B2, B3, B4, TS, and RL-based B), the quantitative results are shown in Table 1,* in the table refers to the data in Bhunia[8]; we can note that (1) Our proposed used a LSTM-based module considering the correlation between incomplete sketches to learn efficient embedding representation; And our proposed outperforms all baselines by a significant margin at the early sketch retrieval. (2) In B1, author trained a base model (See Fig.a) using only complete sketches, and no mechanism is designed to optimize the incomplete sketches; And thus, such model perform poor at early retrieval. (3) In B2, author trained a base model (Fig.a) on all intermediate sketches as training data to optimize the embedding representation, but the model still perform pool, the fundamental reason is that the diversity of incomplete sketches can obviously confuse the base model. (4) Compared with B2, B3 goes to another extreme, in which each base model was trained on every percentage of sketches; Despite a slight improvement in performance at easily retrieval, such method can weaken the generalization ability of the model. (5) TS try to complete the incomplete sketch first by a image-to-image translation model, and then perform the retrieval through the base model; The performance of this method depends on the performance of image-to-image translation model, in which there are still many problems to be solved in this kind of model, so it can not meet the needs of the problem of on-the-fly SBIR. (6) In B4 and RL-based B, authors explored to learn the efficient embedding representation by optimizing a non-differentiable ranking loss, both which can improve the performance of easily retrieval. We further evaluated the performance of our proposed with a varying feature embedding dimension. The results in Table 2 show that our proposed method outperforms RL-based B at each dimension.



(a) Several sets of users sketches retrieve instances of top10 (b) Statistics of the stroke numbers

Fig. 6. Hand-drawn sketches from users. (a) Several sets of users sketches retrieve instances of top10; (b) Statistics of the stroke numbers of each complete sketch in the "users" data.

TABLE Ⅲ
COMPARISON RESULTS WITH DIFFERENT STARTING STROKES

		Chair-V2			Shoe-V2		
	model	1	2	4	1	2	4
m@A	RL-B	85.44	85.67	85.63	85.38	83.20	83.83
	ours	**89.54**	**89.20**	**89.54**	**88.70**	**88.62**	**88.74**
m@B	RL-B	35.09	34.10	33.53	21.44	22.67	23.65
	ours	**38.57**	**36.92**	**36.86**	**24.00**	**23.93**	**24.88**
A@5	RL-B	76.34	75.23	75.23	65.77	62.91	65.91
	ours	**79.87**	74.61	**78.63**	62.16	62.76	63.60
A@10	RL-B	89.65	88.23	88.85	79.63	76.42	77.77
	ours	**91.33**	**89.78**	**90.71**	77.02	75.25	76.42

TABLE Ⅳ
COMPARISON RESULTS WITH DIFFERENT NUMBER OF RENDERED SHEETS.

	Chair-V2				Shoe-V2			
	m@A		m@B		m@A		m@B	
	RL-B	Ours	RL-B	Ours	RL-B	Ours	RL-B	Ours
Steps-10	86.53	**90.06**	35.22	**39.75**	83.88	**89.28**	22.80	**25.07**
Steps-15	85.63	**89.73**	34.90	**38.99**	83.15	**88.87**	22.08	**24.08**
Steps-20	85.44	**89.54**	35.09	**38.57**	85.38	**88.70**	21.44	**24.00**
Steps-25	85.22	**89.43**	34.17	**38.36**	82.59	**88.50**	21.55	**23.65**
Steps-30	85.01	**89.26**	33.67	**38.14**	82.41	**88.36**	21.33	**23.50**

TABLE Ⅴ
COMPARATIVE RESULTS ON THE USERS' DATA.

	Source	Method	m@A	m@B	A@5	A@10
Chair-V2	Test set	RL-B*	85.44	35.09	76.34	89.65
		Ours	**89.54**	**38.57**	**79.87**	**91.33**
	Practice	RL-B	88.28	34.70	71.00	**86.00**
		Ours	**90.72**	**35.98**	**72.00**	84.00
Shoe-V2	Test set	RL-B*	85.38	21.44	65.77	79.63
		Ours	**88.70**	**24.00**	62.16	77.02
	Practice	RL-B	87.47	30.76	71.50	85.50
		Ours	**89.88**	**31.29**	65.50	81.00

V. FURTHER ANALYSIS

Each photo corresponds to only one complete sketch drawing episode in the above experiments and has the following challenges for practical implementation, as shown in Fig. 5. Sketches reveal significant differences among users; e.g., for the same image, different users would draw from different starting positions, generating different drawing episodes that influence the algorithm performance (see Fig. 5(b)). Moreover, a single stroke can vary in shape and size between users (see Fig. 5(c)), and thus can contain different information.

A. Different starting position for drawing a sketch

In this section, we generate sketch drawing episodes from different starting position, where each photo corresponds to several sketch drawing episodes and analyze its influence on early retrieval results. Here, each photo generates four sketch drawing episodes. The first was generated based on the default order of strokes (named order-0), and the other three (order-1, order-2, order-3) were obtained by randomly modifying the starting positions of order-0. In the above sections, all models were trained on sketches generated using the default order.

In practice, the order of strokes is highly diverse and far more than four. To verify the influence of such diversity on model learning, we designed three tasks (Tasks 1, 2, and 3). Task 1 indicates the original experiment in which each photo generates only one sketch drawing episode from order-0. In Task 2, we extended the training and testing data by combining orders-0 and 1. In Task 3, we further extended the training and testing data by combining orders-0, 1, 2, and 3. Training Tasks 2 and 3 were implemented as in Task 1. From the results in Table 3, we observe that (1) different styles increase the diversity of data. In a moderate range, this diversity is equivalent to performing data augmentation; Diversity may also increase the difficulty of learning. For Shoe-V2, the data are rich, and the expanded data still obey the original data distribution; thus, the proposed and baseline methods have improved model performance. For Chair-V2, the data is relatively small, and the expanded data expands the boundary of the original data distribution, reducing the model performance; (2) our proposed method outperforms the baseline methods by a significant margin.

B. *Different sizes for each stroke*

In this section, we discuss cases where different users draw strokes of different sizes, implying that different strokes can contain different information. In the original sketch data, each stroke accounts for the same proportion (1/20) of the complete sketch; Here, we test the performance of the proposed and baseline methods on the generated test data where the drawing episode has strokes of same proportion (1/Step). The results in Table 4. show that (1) For the test sets with less than 20 steps, each stroke contains more information comparing with the original sketches, the early retrieval of our proposed has improved, while for baseline model it is uncertain; For the test sets with more than 20 steps, each stroke contains less information comparing with the original sketches, the early retrieval of our proposed has decreased. (2) Our proposed method outperforms the baseline by a significant margin.

C. *Practical test*

In this section, we aim to verify the performance of the proposed and baseline models in practice and then provide suggestions for future research. We invited 40 users (graduate students) to submit a total of 300 sketch drawing episodes (including 100 chair photos and 200 shoe photos) that we termed "users" data. Before drawing, we showed the users the target photo and explained several basic drawing rules. One instance drawn by different users is shown in Fig. 6. As shown in Table 5, the performances of the proposed and baseline methods were verified. The results confirm that (1) the proposed method notably outperformed the baseline methods (RL-based B) on the real hand-drawn sketches; (2) the performances of the proposed and RL-based methods in practical testing significantly improved compared with the test set. This is because the strokes used to complete a sketch (user data) were far fewer than those used in model training (see Fig. 6(b)); i.e., each stroke contains more details or information.

VI. Conclusion and Future Work

The recent increase in hand-painted data has made FG-SBIR problems a topic of interest in computer vision. However, finishing a sketch takes time and skill, hindering its widespread adoption. On-the-fly FG-SBIR was proposed to overcome the aforementioned barriers, where retrieval was performed after each stroke drawing. We considered the sketch drawing episode as a sequence optimization problem and designed a Bi-LSTM module to optimize the embedding space for incomplete sketches, and uncovered some challenges in practical applications. The experiments verified that the proposed method provides a considerable improvement.

The framework of On-the-fly FG-SBIR is new topic and faces many challenges in practical applications. For example, (1) sketch and photo belong to non-homologous data, and incomplete sketch can only provide a small amount of local information; Therefore, how to narrow the cross domain gap between them has always been the key problems in the problem of on-the-fly SBIR. (2) Users' painting styles show great differences, that is, there are great differences within the category; Therefore, One of key point is how to deal with these intra-class differences. (3) Users' painting styles show great differences, that is, there are great differences within the category; Therefore, how to deal with these intra-class differences is also a key problem. Our future work will focus on the above problems.

References

[1] Collomosse, J.; Bui, T.; and Jin, H. 2019. Livesketch: Query perturbations for guided sketch-based visual search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2879-2887.

[2] Dey, S.; Riba, P.; Dutta, A.; Llados, J.; and Song, Y. Z. 2019. Doodle to search: Practical zero-shot sketch-based image retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2179-2188.

[3] Yu, Q.; Liu, F.; Song, Y. Z.; Xiang, T.; Hospedales, T. M.; and Loy, C. C. 2016. Sketch me that shoe. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 799-807.

[4] Song, J.; Pang, K.; Song, Y. Z.; Xiang, T.; and Hospedales, T. M. 2018. Learning to sketch with shortcut cycle consistency. In Proceedings of the IEEE conference on computer vision and pattern recognition, 801-810.

[5] Xue J , Zhou Y , Jiang Z , et al. 2020. A Multiple Triplet-Ranking Model for Fine-Grained Sketch-Based Image Retrieval. In 2019 IEEE Visual Communications and Image Processing (VCIP).

[6] Pang, K.; Li, K.; Yang, Y.; Zhang, H.; Hospedales, T. M.; Xiang, T.; and Song, Y. Z. 2019. Generalising fine-grained sketch-based image retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 677-686.

[7] Sain, A.; Bhunia, A. K.; Yang, Y.; Xiang, T.; and Song, Y. Z. 2020. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. arXiv preprint arXiv:2007.15103.

[8] Bhunia, A. K.; Yang, Y.; Hospedales, T. M.; Xiang, T.; and Song, Y. Z. 2020. Sketch less for more: On-the-fly fine-grained sketch-based image retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9779-9788.

[9] Lei, H. , Chen, S. , Wang, M. , X He, and Li, S. . 2021. A new algorithm for sketch-based fashion image retrieval based on cross-domain transformation. In Wireless Communications and Mobile Computing, 2021.1-14.

[10] Dai, D. , Tang, X. , Xia, S. , Liu, Y. , Wang, G. , and Chen, Z. 2022. Multi-granularity association learning framework for on-the-fly fine-grained sketch-based image retrieval.

[11] Ribeiro, L. , Tu, B. , Collomosse, J. , and Ponti, M. . 2020. Sketchformer: transformer-based representation for sketched structure. In IEEE.

[12] Cao, Z. , Cui, S. , and Zhang, C. .2021. Dcr: disentangled component representation for sketch generation. In Pattern Recognition Letters, 145(8).

[13] Qi, Y.; Song, Y. Z.; Xiang, T.; Zhang, H.; Hospedales, T.; Li, Y.; and Guo, J. 2015. Making better use of edges via perceptual grouping. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1856-1865.

[14] Tolias, G.; and Chum, O. 2017. Asymmetric feature maps with application to sketch based retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2377-2385.

[15] Lowe, D. G. 1999. Object recognition from local scale-invariant features. In Proceedings of the seventh IEEE international conference on computer vision, 1150-1157. IEEE.

[16] Hu, R.; and Collomosse, J. 2013. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. Computer Vision and Image Understanding, 117(7): 790-806.

[17] Saavedra, J. M. 2014. Sketch based image retrieval using a soft computation of the histogram of edge local orientations (s-helo). In 2014 IEEE international conference on image processing (ICIP), 2998-3002. IEEE.

[18] Saavedra, J. M.; Barrios, J. M.; and Orand, S. 2015. Sketch based Image Retrieval using Learned KeyShapes (LKS). In Proceedings of the British Machine Vision Conference (BMVC), 1(2): 7.

[19] Yu, Q.; Yang, Y.; Song, Y. Z.; Xiang, T.; and Hospedales, T. 2015. Sketch-a-net that beats humans. arXiv preprint arXiv:1501.07873.

[20] Liu, L.; Shen, F.; Shen, Y.; Liu, X.; and Shao, L. 2017. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2862-2871.

[21] Song, Y., Lei, J., Peng, B., Zheng, K., Yang, B., and Jia, Y. 2019. Edge-guided cross-domain learning with shape regression for sketch-based image retrieval. IEEE Access, 7, 32393-32399.

[22] Zheng, Y. , Yao, H. , and Sun, X. . 2020. Deep semantic parsing of freehand sketches with homogeneous transformation, soft-weighted loss, and staged learning. In IEEE Transactions on Multimedia, PP(99), 1-1.

[23] Zhang, H.; Zhang, C.; and Wu, M. 2017. Sketch-based cross-domain image retrieval via heterogeneous network. In 2017 IEEE Visual Communications and Image Processing (VCIP), 1-4. IEEE.

[24] Wang, Y., Huang, F., Zhang, Y., Feng, R., Zhang, T., and Fan, W. 2020. Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 100, 107148.

[25] Zhu, M., Chen, C., Wang, N., Tang, J., and Bao, W. 2019. Gradually focused fine-grained sketch-based image retrieval. Plos one, 14(5), e0217168.

[26] Sain, A.; Bhunia, A. K.; Yang, Y.; Xiang, T.; and Song, Y. Z. 2021. Stylemeup: Towards style-agnostic sketch-based image retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8504-8513.

[27] Du, R., Chang, D., Bhunia, A. K., Xie, J., Ma, Z., Song, Y. Z., and Guo, J. 2020. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In European Conference on Computer Vision, 153-168.

[28] Choi, J., Cho, H., Song, J., and Yoon, S. M. 2019. Sketchhelper: Real-time stroke guidance for freehand sketch retrieval. IEEE Transactions on Multimedia, 21(8), 2083-2092.

[29] Huang, F., Canny, J. F., and Nichols, J. 2019. Swire: Sketch-based user interface retrieval. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 1-10.

[30] Collomosse, J., Bui, T., and Jin, H. 2019. Livesketch: Query perturbations for guided sketch-based visual search. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2879-2887.

[31] Chopra, S., Jain, G. ,Chopra, S., and Parihar, A. S. .2020. TransSketchNet: Attention-based Sketch Recognition using Transformers. In 24th European Conference on Artificial Intelligence (ECAI) 2020.

[32] Liu, H., Jiang, B., Xiao, Y., and Yang, C. (2019). Coherent semantic attention for image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 4170-4179.

[33] Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., and Ebrahimi, M. 2019. Edgeconnect: Generative image inpainting with adversarial edge learning. arXiv preprint arXiv:1901.00212.

[34] Li, J., Gao, N., Shen, T., Zhang, W., Mei, T.,and Ren, H. 2020. SketchMan: Learning to Create Professional Sketches. In Proceedings of the 28th ACM International Conference on Multimedia, 3237-3245.

[35] Xie, M., Xia, M., Liu, X., Li, C., and Wong, T. T. 2021. Seamless manga inpainting with semantics awareness. ACM Transactions on Graphics (TOG), 40(4), 1-11.

[36] Yu, J.; Lin, Z.; Yang, J.; Shen, X.; Lu, X.; and Huang, T. S. 2018. Generative image inpainting with contextual attention. In Proceedings of the IEEE conference on computer vision and pattern recognition, 5505-5514.

[37] Zheng, C.; Cham, T. J.; and Cai, J. 2019. Pluralistic image completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1438-1447.

[38] Ha, D.; and Eck, D. 2017. A neural representation of sketch drawings. arXiv preprint arXiv:1704.03477.

[39] Chen, Y.; Tu, S.; Yi, Y.; and Xu, L. 2017. Sketch-pix2seq: a model to generate sketches of multiple categories. arXiv preprint arXiv:1709.04121.

[40] Cao, N.; Yan, X.; Shi, Y.; and Chen, C. 2019. AI-sketcher: a deep generative model for producing high-quality sketches. In Proceedings of the AAAI Conference on Artificial Intelligence, 33(01):2564-2571.

[41] Liu, F.; Deng, X.; Lai, Y. K.; Liu, Y. J.; Ma, C.; and Wang, H. 2019. Sketchgan: Joint sketch completion and recognition with generative adversarial network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5830-5839.

[42] Aksan, E.,Deselaers, T.,Tagliasacchi,A.,and Hilliges, O.2020.Cose: Compositional stroke embeddings. arXiv preprint arXiv:2006.09930.

[43] Lin, H., Fu, Y., Xue, X., and Jiang, Y. G. 2020. Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6758-6767.

[44] Ghosh, A., Zhang, R., Dokania, P. K., Wang, O., Efros, A. A., Torr, P. H., and Shechtman, E. 2019. Interactive sketch & fill: Multiclass sketch-to-image translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 1171-1180.

[45] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2818-2826.

[46] Muhammad, U. R.; Yang, Y.; Song, Y. Z.; Xiang, T.; and Hospedales, T. M. 2018. Learning deep sketch abstraction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8014-8023.

[47] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252.

[48] D Wang, Sapkota, H. , Liu, X. , and Yu, Q. . 2021. Deep reinforced attention regression for partial sketch based image retrieval.