From Sparse to Complete: Semantic Understanding Based on Stroke Evolution in On-the-fly Sketch-based Image Retrieval

Yingge Liu[1] , **Dawei Dai**[1] , **Xiangling Hou**[1] , **Shilin Zhao**[1] , **Guoyin Wang**[1,2]
[1] Chongqing Key Laboratory of Computational Intelligence, Key Laboratory of Big Data Intelligent Computing, Key Laboratory of Cyberspace Big Data Intelligent Security, Ministry of Education, School of Computer Science and Technology, Chongqing University of Posts and Telecommunications
[2] National Center for Applied Mathematics, Chongqing Normal University, Chongqing, China
yinggelyg@gmail.com, dai_dw@163.com,{s240201159, s240201159}@stu.cqupt.edu.cn, wanggy@cqnu.edu.cn

Abstract

In contrast with human sketching, which pre-conceptualizes outlines and features, conventional sketch retrieval models rely primarily rely on pixel-level processing and feature extraction, limiting their ability to capture early sketch intent. Consequently, these models are susceptible to subjective stroke noise, reducing retrieval accuracy. To address this issue, we propose a novel on-the-fly noise stroke retrieval framework designed to align with human sketch-drawing cognition. The proposed framework introduces two core innovations. (i) A stroke consistency detection module that effectively discriminates and suppresses noise strokes by quantifying the structural similarity between the current stroke and the target image, as well as its alignment with key skeletal components. (ii) An adaptive gated mixture of experts module that dynamically selects and integrates features from multiple expert networks during the early, sparse stages of sketching, thereby capturing relevant information with greater precision. Experimental results across diverse sketch datasets demonstrate that the proposed method effectively identifies and suppresses early noise strokes, significantly enhances sketch retrieval performance, and exhibits strong robustness across varying sketch styles.

1 Introduction

With the widespread adoption of interactive touchscreen devices, sketch-based image retrieval (SBIR) has emerged as an accessible and practical modality. In particular, the advent of on-the-fly frameworks[Bhunia *et al.*, 2020; Liang *et al.*, 2021; Liu *et al.*, 2022; Dai *et al.*, 2024b] has significantly lowered the entry barriers for users, enabling real-time sketching and retrieval, thereby reducing the required interaction time and obviates the need for complete sketches. However, a fundamental challenge in the early stages of on-the-fly SBIR is differentiating between keystrokes indicative of user intent and those that are unintentional or noisy strokes. To convey specific shapes, structures, or details, users may draw strokes that deviate from typical features expected by the model. Although these strokes may carry significant semantic information for the user, existing models may typically misclassify them as noise, negatively affecting the understanding of sketch intent and recognition accuracy. For example, as shown in Fig. 1, when users aim to represent the textural details of shoelaces, the drawn strokes may exhibit slight curvature or incomplete closures. These strokes encode significant semantic information about the texture. However, deep learning models that primarily rely on statistical regularities for feature matching may consider these strokes as noise that deviates from established category features. This phenomenon is particularly pronounced in objects with intricate details.

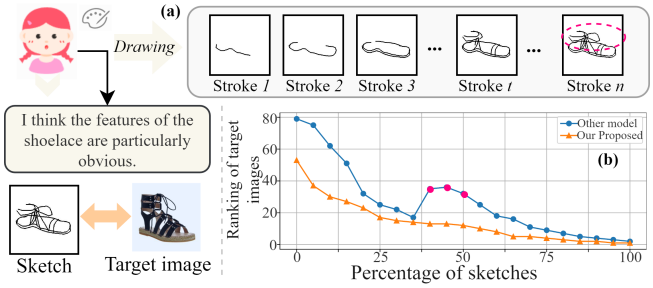


Figure 1: (a) The drawing process of sketch based on their understanding. The red dots in the line graph (b) indicate the moments when the user drew the shoelaces, at which point the sketch retrieval performance of the model significantly decreases. This is because the model treats the drawn shoelaces as non-critical features or even noise, leading to a substantial decrease in retrieval performance.

Recent studies have attempted to address noise strokes in SBIR [Koley *et al.*, 2024b; Bandyopadhyay *et al.*, 2024]. For example, Sain et al. [Sain *et al.*, 2021] proposed a style-agnostic model, Bhunia et al. [Bhunia *et al.*, 2022a] developed a stroke subset selector to filter noisy strokes, and Koley et al. [Koley *et al.*, 2024a] focused on handling abstract strokes. However, these approaches are largely oriented toward complete sketches. In the early stages of on-the-fly retrieval, where models receive only sparse strokes and lack contextual information, the identification of noise strokes and the extraction of effective features become significantly more challenging. The core challenges in on-the-

fly SBIR are twofold. (1) Accurately determining whether a given stroke is integral to the target features or an irrelevant distractor is challenging, particularly before the key features of the target object are fully rendered. This challenge is particularly pronounced for detail-rich objects such as facial sketches. (2) Extracting semantic features from sparse, incomplete strokes in the initial stages of sketching presents another challenge, as it is essential for effective matching during early retrieval.

To address the challenges of noise stroke interference in the early stages of on-the-fly retrieval, we propose a novel noise-filtering framework for the on-the-fly SBIR problem. Our method mitigates the challenges identified above by evaluating the significance of user-drawn sketch strokes at both pixel and structural levels. Our framework consists of two key components: the stroke consistency detection module (SCDM) and adaptive gated mixture of experts module (AGMoE). The SCDM assesses the importance of sketch strokes from both pixel-level and contour structure perspectives, ensuring the accuracy of detection for both simple and complex strokes, and preserving the crucial sketch information required for model parsing and learning. The objective of the AGMoE module is to extract semantic information from filtered strokes comprehensively, aiming to avoid information loss and semantic leakage, thereby enabling the extraction of robust and discriminative features, even in the early stages.

To evaluate the generalization ability and robustness of the proposed model in handling complex strokes, we selected two categories of datasets: a set of simple sketch datasets (ChairV2 and ShoeV2) and detailed facial sketch dataset. Additionally, considering the diverse range of sketch styles encountered in practical applications, variations in user drawing habits, and the fact that existing datasets may not fully reflect real-world scenarios, we constructed a more challenging dataset based on the SKSF-A dataset [Yun *et al.*, 2024], encompassing seven distinct sketch styles. The strokes in the sketch sequences of this dataset closely approximate the actual user drawing process. Model performance is evaluated from both qualitative and quantitative perspectives. The experimental results demonstrate that the proposed model exhibits excellent retrieval performance in the early stages of sketch drawing and is suitable for handling sketches of different styles. Our contributions are summarized as follows:

(i) We propose a novel on-the-fly noise-filtering cross-modal retrieval framework. This framework effectively addresses the noise strokes generated during the drawing and retrieval process, thereby enhancing retrieval performance and user experience.

(ii) We design a stroke evaluation and feature extraction mechanism integrating the SCDM and AGMoE module to identify key sketch information by assessing stroke importance at the pixel and contour levels, ensuring the maximal preservation of crucial semantic information while removing noise.

(iii) We construct a challenging multi-style sketch dataset that aligns closely with practical application scenarios. This dataset serves as a benchmark for evaluate the generalization ability and robustness of models in practical settings, providing a more relevant evaluation tool for future research.

2 Related Work

Sketch-Based Image Retrieval. SBIR has witnessed significant advancements, particularly in the fine-grained SBIR (FG-SBIR), which matches sketches to specific image instances. [Guo *et al.*, 2017; Bhunia *et al.*, 2021a; Bhunia *et al.*, 2021b; Bhunia *et al.*, 2022b; Zuo *et al.*, 2024; Zhou *et al.*, 2024; He *et al.*, 2025]. Researchers have explored various strategies to enhance the performance of FG-SBIR, including leveraging deep Siamese triplet networks [Yu *et al.*, 2016] and further optimize them by with attention-based higher-order loss functions and text tags [Bhunia *et al.*, 2023; Fang *et al.*, 2024]. While some studies address noisy sketches and style variations[Sain *et al.*, 2021; Bandyopadhyay *et al.*, 2024], key challenges like sketch style modeling in dynamic scenes and effective noise discrimination in intricate sketches still need further investigation.

Mixture-of-Experts. The MoE model uses a gating network to activate multiple expert subnetworks for task-specific processing. Its ability to manage data diversity has made it prominent in large language models (LLMs) [Carion *et al.*, 2020; Li *et al.*, 2022; Radford *et al.*, 2021] following the rise of vision transformers [Rao *et al.*, 2022; Jain *et al.*, 2023; He *et al.*, 2024] in the visual domain (e.g., VMoE [Riquelme *et al.*, 2021], LlMoE [Mustafa *et al.*, 2022], and CuMo [Li *et al.*, 2024]). Given the inherent stylistic diversity in sketches, the MoE model can assign different styles to specific experts for tailored learning, enabling targeted modeling of sketch features. This study introduces an MoE module to leverage its advantages for handling diverse sketch styles, aiming to improve model understanding and generation in complex or noisy scenarios. This approach allows for finer-grained stylistic capture and efficient resource utilization, enhancing complex brushstrokes processing performance.

3 Problem Definition

On-the-fly SBIR aims to retrieve a target image as rapidly as possible using the fewest strokes. The drawing process is modeled as an ordered sketch sequence $\mathcal{S} = \{s_1, \ldots, s_t, \ldots, s_n\}$, where s_t is the tth stroke and n is the total stroke count. At time t ($1 \leq t \leq n$), the input sketch is s_t. Consider a sketch sequence space $\mathcal{D}^S$, where $\mathcal{S} \in \mathcal{D}^S$. Given an image database $\mathcal{D}^I = \{I_1, \ldots, I_m\}$ containing m images, the retrieval model dynamically evaluates the degree of matching between s_t and each $I_j \in \mathcal{D}^I$. A matching function $f(s_t, I_j) \in \mathbb{R}$ is defined to represent similarity. The model generates a dynamic ranking list $\mathcal{R}_t$, arranging images in descending order based on their match with s_t. At each time instance t of dynamic retrieval, the model evaluates the match between the current sketch s_t and each image $I_j \in \mathcal{D}^I$. A matching function $f(s_t, I_j)$ that outputs a scalar similarity measure is then defined. Based on this matching function, the model ranks the images in the database, generates a dynamic ranking list $\mathcal{R}_t$, and orders images by decreasing match score s_t. Our objective is to construct a dynamic retrieval model where the rank of $\mathcal{I} \in \mathcal{R}_t$ improves rapidly as the user adds strokes. Ideally, a minimum stroke count $t_{opt} \leq n$ exists such that for $t \geq t_{opt}$, the target image $\mathcal{I}$ achieves the top rank, meaning $Rank(\mathcal{I}, \mathcal{R}_t) = 1$. We aim to retrieve the

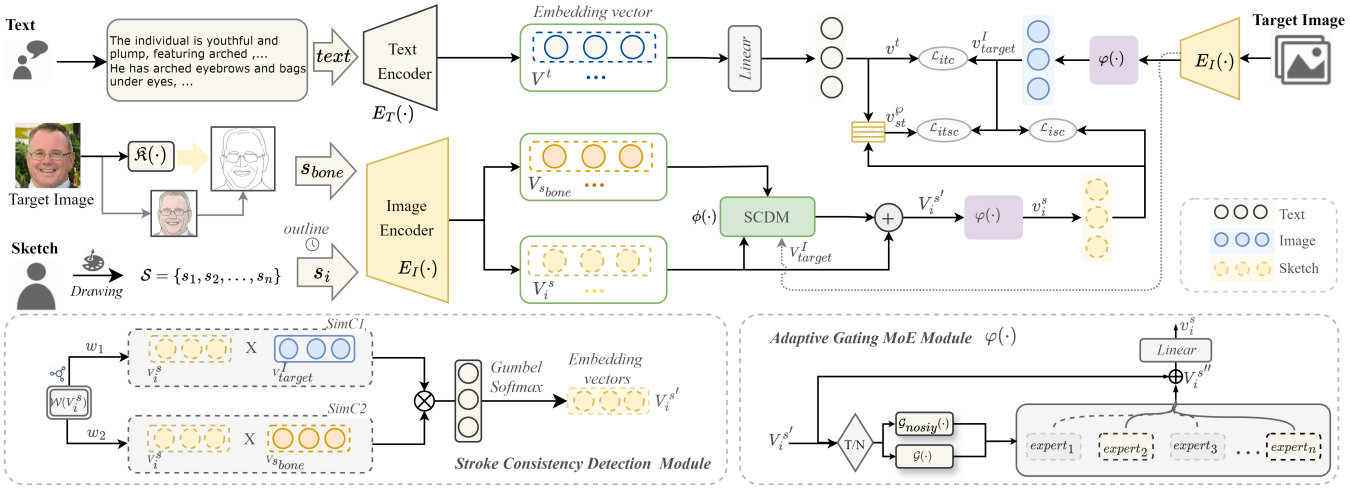


Figure 2: Model Architecture. Ours model essentially consists of two separate branches the image branch, encapsulating an image encoder, SCDM and AGMoE, and the text branch comprising of a text encoder and linear.

target image as quickly as possible, which translates to the minimization of t_{opt}. We attempt to optimize the matching function and ranking such that:

$$\underset{f}{argmin}\,\{t_{opt} | \forall t \geq t_{top}, Rank\,(\mathcal{I}, \mathcal{R}_t) = 1\} \qquad (1)$$

where $t_{top} \in [1, n]$, and $t_{opt}(f)$ denotes the minimum number of strokes under the matching function.

4 Proposed Methodology

Overview. We design a cross-modal, real-time retrieval system for noise filtering (See Fig. 2), comprising text and image branches for diverse query inputs. The text branch uses a text encoder and linear layer to reduce dimensionality, facilitating semantic matching with target images. The image branch handles images and sketches. For sketch queries, processing involves an image encoder, SCDM, and AGMoE module. Image queries use an image encoder and AGMoE. Dimensionally reduced sketch features are then fine-grained matched with corresponding image features. Text assists sketch matching, particularly for sparse early sketches, to mitigate retrieval randomness. Dimensionally reduced text and sketch features are fused into ($v^{\varphi}_{st} = fusion[v^s_i : v^t]$) for image matching. Specifically, we present three salient designs. (i) A novel SCDM analyzes sketch stroke weights at two complementary levels, improving noise filtering reliability. Simultaneously, AGMoE maximizes semantic information retention post-denoising. (ii) Textual data assists retrieval of early, sparse sketches, effectively mitigating initial retrieval randomness. (iii) Training integrates unimodal contrastive and auxiliary fusion losses for fine-grained matching between combined queries and target images.

4.1 Stroke Consistency Detection Module

The SCDM evaluates each stroke's importance from two complementary perspectives. First, we focus on pixel-level similarity between sketch strokes and target images. This direct comparison captures fine-grained correspondences to assess the contribution to detail representation. An image encoder $E_I(\cdot)$ extracts embeddings $V_i^s = E_I(S_i)$ for the current sketch stroke s_i and $V_I^p = E_I(T^p)$ for the target image T^p. Pixel-level consistency is then measured using the cosine similarity between these embeddings.

$$sim(\mathcal{S}, \mathcal{I}) = \frac{V_i^s \cdot V_I^p}{\|V_i^s\| \|V_I^p\|} \qquad (2)$$

Second, to enhance the perception of global structural information and assess stroke importance, we compare the current sketch with the target image's skeleton contour, S_{bone}. This aids in identifying whether strokes capture the target image's key structures. We calculate the cosine similarity between the embedding representations of the stroke and $S_{bone} = \Re(T^p)$, described as $V_{s_{bone}} = E_I(S_{bone})$, using an image encoder to measure structural consistency. Combining pixel- and structural-level similarities yields a comprehensive stroke importance score, sim.

$$sim = w_1 \cdot sim_{pix}(V_i, V_I^p) + w_2 \cdot sim_{str}(V_i, V_{s_{bone}}) \qquad (3)$$

Where, w_1 and w_2 are dynamically adjusted parameters. For simple line drawings, precise pixel-level matching is less critical, so w is dynamically adjusted based on the sketched object's complexity.

To transform the continuous similarity score sim into discrete stroke weights for differentiated processing. we use the Gumbel-Softmax function. This function differentiably samples from a categorical distribution, yielding a discrete, sparse weight distribution. Specifically, Gumbel-Softmax generates a corresponding weight value for each stroke based on its comprehensive similarity score. Strokes with higher similarity scores receive higher weights, indicating their crucial role in representing the target image, while strokes with lower similarity scores receive lower weights, suggesting they may be noise or redundant information.

4.2 Adaptive Gating MoE Module

We use MoE to better capture diverse sketch strokes for enhanced quality. To address MoE's limitations, we propose the AGMoE module, which improves model robustness through a Top-k gating mechanism and adaptive noise.

As shown in Fig. 2, the denoised sketch strokes $V_i^{s'}$ serve as the inputs for the AGMoE module, which primarily consists of N expert networks $Expert_i : \mathbb{R}^D \rightarrow \mathbb{R}^D, i = 1, 2, 3, \ldots, n$, a gating network $\mathcal{G} : \mathbb{R}^{D_{gate}} \rightarrow \mathbb{R}^N$, and parameters $w_{noise} \in \mathbb{R}^{D \times N}$, which are utilized for jointly adjusting noise. First, the input stroke X_{input} passes through the gating network $\mathcal{G}(\cdot)$ to generate logit values (l_{logits}) for each expert network, which are used to measure the degree of applicability of each expert to the current input stroke.

$$l_{logits} = \mathcal{G}(X_{input}) \in \mathbb{R}^{B \times T \times N} \tag{4}$$

Where, B, T, and N denote the batch size, stroke length, and number of expert networks, respectively. During training, we add adaptive noise to the gating network's output logits (l_{logits}) to enhance the module's exploration capabilities and increase expert selection diversity. This randomness encourages the model to explore various expert combinations during training, improving its generalization ability. When the training condition (*train* = True) is satisfied and noisy gating is enabled ($noisy_{gating}$ = True) , the logits are dynamically adjusted by the following formula:

$$\begin{cases} \sigma = softplus(0.1 \cdot l_{logits}) + \alpha, \\ L_{nosiy} = l_{logits} + \mathcal{N}(0, \alpha^2) \end{cases} \tag{5}$$

Where $\alpha = 10^{-2}$ is a constant ensuring the noise's standard deviation is always greater than zero. The softplus function correlates noise variance with the magnitude of l_{logits} achieving adaptive noise injection. Depending on whether noisy gating is enabled during the training phase, the final l'_{logits} are calculated as follows:

$$l'_{logits} = \begin{cases} \mathcal{L}_{nosiy}, & If \quad noisy_{gating} \,\&\, train \quad (6) \\ l_{logits}, & Otherwise \quad (7) \end{cases}$$

Subsequently, to enhance computational efficiency and direct the model's attention to the most relevant experts, we introduce the Top-k gating mechanism. To ensure that softmax normalization is computed only on the selected Top-k logits, we set the values of the unselected logits to negative infinity, resulting in a sparse vector l_{sparse}. By applying softmax normalization to l_{sparse}, we obtain the gating weights $\mathcal{G}(\cdot)$ representing the importance of each selected expert.

$$\mathcal{G} = \text{softmax}(l_{sparse}) \in \mathbb{R}^{B \times T \times K} \tag{8}$$

Each expert network is a simple multilayer perceptron with two linear layers and a rectified linear unit activation function.

$$\begin{cases} \tilde{x} = Relu(w_1 \cdot x + b_1), \\ Expert_i(x) = Dropout(w_2 \cdot \tilde{x}) \end{cases} \tag{9}$$

For each selected $Expert_i, i \in \{1, 2, 3, ..., k\}$, its output for input $\tilde{x}$ is computed as $Y_i = Expert_i(\tilde{x}) \in \mathbb{R}^{B \times T \times D}$. The outputs of all k selected experts are then stacked.

$$Y = [Y_1, Y_2, \cdots, Y_k] \in \mathbb{R}^{B \times T \times K \times D} \tag{10}$$

Dataset	Sketch		Image		Other	
	Train	Test	Train	Test	Style	Text
QMUL-Chair-V2	1,275	725	300	100	3	✗
QMUL-Shoe-V2	6,051	679	1,800	200	3	✗
ChairV2-U1	✗	1894	✗	100	N	✗
ShoeV2-U1	✗	3781	✗	200	N	✗
FS2K-SDE1	53,950	22,500	1,079	450	1	✓
FS2K-SDE2	16,700	7,150	334	143	1	✓
User-U1	✗	3,474	✗	110	N	✓
SFSK-A	✗	938	✗	134	7	✗

Table 1: Several publicly available sketch datasets.

Finally, we apply the gating weights to weight the output of each expert, thereby obtaining the final weighted output.

$$Y_{weighted} = \mathcal{G} \otimes Y \tag{11}$$

Where, $\otimes$ denotes element-wise multiplication. The gating network scales each expert's output based on its corresponding importance.

4.3 Expert Contrastive Learning

This section outlines our model's training, which incorporates two auxiliary tasks designed for noisy stroke filtering during interactive drawing.

(1) *Cross-modal alignment*. To account for input modality variations, specialized alignment tasks are performed by different experts on specific modalities. We design a cross-modal alignment learning task to explore the specialized knowledge and skills of each expert for specific alignment tasks, thereby facilitating fine-grained learning and optimization. Image-text contrastive learning (ITC loss [Radford *et al.*, 2021]) ensures close alignment of text and images within the feature space, optimizing their semantic consistency. Sketch image contrastive learning (ISC loss [Radford *et al.*, 2021]) maps sketches and their corresponding facial images to a shared feature space.

(2) *Bimodal alignment*. Leveraging the complementarity of multimodal information further enhances the robustness of the model. When one modality's information is missing or excessively noisy, the model can rely on the other, preventing significant overall performance impact. ITSC loss calculates fused features as follows:

$$\mathcal{L}_{ITSC} = -\sum_{i=1}^{N} \log \frac{\exp(sim(v_{st}^{\wp}, v_i^I)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(v_{st}^{\wp}, v_j^I)/\tau)} \tag{12}$$

Where, $v_{st}^{\wp}$ is the embedding vector after fusing sketch and text, and τ is a temperature parameter.

5 Experiments

Datasets. To select appropriate datasets based on sketch stroke complexity, we use eight publicly available datasets (See Table 1) [Song *et al.*, 2018; Muhammad *et al.*, 2018; Pang *et al.*, 2019; Bhunia *et al.*, 2020; Dai *et al.*, 2022; Liu *et al.*, 2024; Dai *et al.*, 2024b]. Furthermore, to test our model's generalizability across diverse styles, we expand the SKSF-A dataset [Yun *et al.*, 2024], which encompasses seven

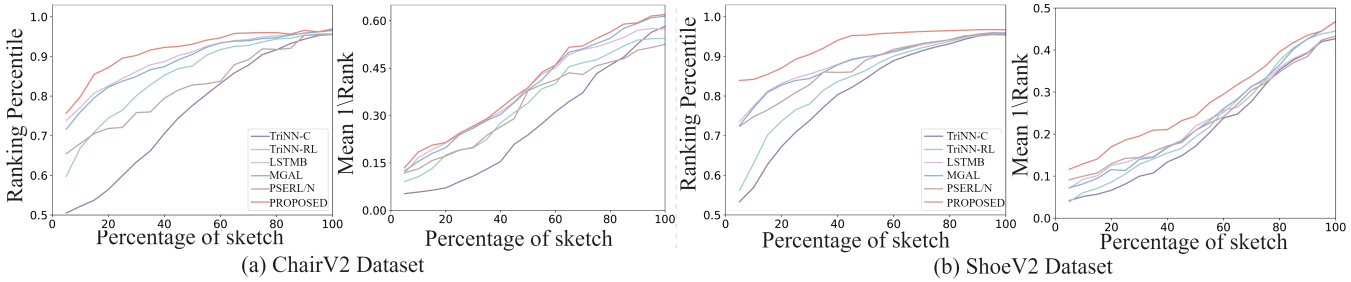


Figure 3: Compared to the baseline model, our method shows a clear performance advantage in early sketch retrieval. The x axis represents indicates the percentage of strokes, and the y axis indicates the retrieval metrics, with higher values indicating stronger retrieval performance.

sketch styles, to simulate scenarios with the stylistic diversity and cluttered strokes encountered in practical applications.

Implementation Details. The model is trained using RTX4090 GPUs. Batch sizes are eight for simple stroke training and 16 for facial sketch training. The model trains for 50 epochs on eight public datasets using the AdamW optimizer with a 0.05 weight decay factor. The AGMoE module employs eight expert networks with a top-k value of four. A cosine-annealing learning rate schedule is adopted, with the learning rate varying between 3e-5 and zero. During training, the ChairV2 and ShoeV2 datasets have 20 total sketch strokes and 64 output dimensions. The total number of strokes in the user-drawn dataset varied.

Evaluation Metrics. (i) We employ the standard FG-SBIR evaluation metric, Acc.@q, which quantifies the percentage of sketches where the true paired photograph appears within the top-q retrieval results. (ii) *On-the-fly*: Given our prioritization of the target images appearing at the top of the retrieval lists, we select the metrics m@A (rank percentile) and m@B (percentage of sketches with 1/rank) [Bhunia *et al.*, 2020] to gauge average retrieval performance across all sketching stages. To evaluate early-stage sketching performance, we use w@MA and w@MB. Higher values of these metrics indicate better performance.

Competitors. TriNN-C [Yu *et al.*, 2016] and TriNN-SR [Bhunia *et al.*, 2020] use a Sketch-a-Net backbone with standard triplet loss. TriNN-RL(_TS) [Bhunia *et al.*, 2020] uses an on-the-fly retrieval framework based on representation learning (RL). LSTMB [Liu *et al.*, 2022] is a Bi-LSTM module optimizing partial sketch sequences. PSERL/N [Liu *et al.*, 2024] optimizes sketch representations using prior knowledge. We compare the model without labels. Cross-Hier [Sain *et al.*, 2020] utilizes a cross-modal hierarchy with expensive paired embedding. StyleMeUp [Sain *et al.*, 2021] utilizes model-agnostic meta learning training. NSR [Koley *et al.*, 2024a] is a lightweight, portable, and interpretable seamless plugin. Seql [Dai *et al.*, 2023] is a sequential learning method that integrates convolutional neural network and lstm modules. MGAL [Dai *et al.*, 2022] and MGRL [Wang *et al.*, 2025] are based on multi-grained RL, using Sketch-a-Net as backbone. MITRL [Dai *et al.*, 2024b] is a multimodal retrieval model trained using LLMs. We also compare its performance against CLIP [Radford *et al.*, 2021], BLIP [Li *et al.*, 2022] and FVIP [Dai *et al.*, 2024a].

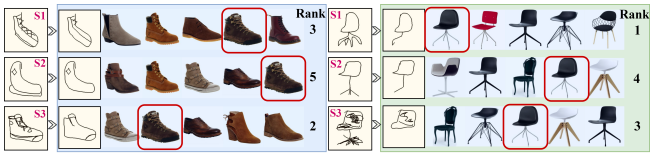


Figure 4: Left: three styles of the same sketch. Target image's state upon its first appearance in the Top-5 list.

Methods	ChairV2		ShoeV2	
	Acc.@1	Acc.@5	Acc.@1	Acc.@5
TriNN-C [Yu *et al.*, 2016]	48.71	76.37	-	65.59
TriNN-SR [Yu *et al.*, 2016]	45.32	74.31	-	61.79
TriNN–RL_TS [Bhunia *et al.*, 2020]	-	73.47	-	62.67
TriNN–RL [Bhunia *et al.*, 2020]	51.20	73.80	30.80	65.10
Cross-Hier [Sain *et al.*, 2020]	62.40	79.10	36.20	67.80
StyleMeUp [Sain *et al.*, 2021]	62.80	**79.60**	36.40	68.10
NSR(P-SLA) [Koley *et al.*, 2024a]	56.50	77.10	36.50	69.30
NSR(SLA) [Koley *et al.*, 2024a]	54.90	76.60	36.10	67.80
Proposed	**62.94**	79.52	**37.94**	**69.98**

Table 2: Comparison of the retrieval performance of our proposed method with other models on complete sketches.

5.1 Comparisons With SOTA Results

To evaluate model performance in terms of identifying noisy strokes, we conduct experiments on simple sketch and face sketch datasets. The results are detailed below.

Performance Analysis on Simple Lines. Given the inherent lack of textual information in the ChairV2 and ShoeV2 datasets, the text branch's input prompt during model training is set to "a photo of f'{$color$} chair/shoe ' ". We chose color as the textual prompt due to sketches' inherent black-and-white characteristics.

(1) *Partial/Early Retrieval.* As illustrated in Fig. 3, the retrieval performance of all methods exhibits an upward trend as the sketch is progressively completed and tends to stabilize when the sketch is near completion. Compared with other models, our model demonstrates a significant advantage in the initial stages of sketch drawing, as it can rapidly and accurately retrieve the target image with very few strokes. These results indicate that the proposed method can effectively identify keystrokes in the early stages of sketching and overcome potential interference strokes during the drawing process. To address the challenges of diverse styles, Fig. 4 presents a retrieval comparison of two groups of sketches. Our model

Model	FS2K-SDE1						FS2K-SDE2					
	m@A	m@B	w@mA	w@mB	Acc.@1	Acc.@5	m@A	m@B	w@mA	w@mB	Acc@.1	Acc@.5
TriNN-C [Yu *et al.*, 2016]	84.77	32.69	50.40	15.83	-	91.33	77.83	24.59	46.00	12.47	-	94.41
TriNN-SR [Yu *et al.*, 2016]	94.16	28.58	58.18	15.83	-	64.22	89.77	34.14	54.99	18.96	-	69.23
TriNN–RL [Bhunia *et al.*, 2020]	84.42	22.76	51.52	12.21	43.50	51.78	85.65	26.70	51.91	14.59	54.26	69.23
SeqL [Dai *et al.*, 2023]	96.22	45.48	59.57	24.56	74.01	90.00	90.22	41.55	54.85	22.22	65.02	**95.82**
MGRL [Wang *et al.*, 2025]	**98.80**	**78.92**	**61.69**	**46.20**	**78.52**	97.11	**96.65**	**69.19**	**60.12**	**40.80**	**78.02**	95.10
MITRL [Dai *et al.*, 2024b]	98.50	70.80	61.25	40.58	73.25	**97.55**	95.81	67.90	59.01	39.09	77.60	94.40
Proposed	**99.70**	**81.10**	**62.34**	**48.97**	**82.13**	**97.57**	**97.93**	**78.68**	**60.78**	**46.44**	**80.25**	**96.22**
Improve(+)	0.91%	2.76%	1.05%	5.99%	4.59%	+0.02	1.32%	13.71%	1.09%	13.82%	2.85%	1.17%
Clip-based* [Radford *et al.*, 2021]	98.01	57.27	60.97	32.58	75.76	78.68	96.02	57.32	59.38	33.36	48.95	71.04
Blip-based* [Li *et al.*, 2022]	97.64	46.72	60.80	27.01	68.22	75.44	97.10	64.06	60.24	37.50	51.58	87.64
FVIP-based* [Dai *et al.*, 2024a]	98.14	56.70	61.16	32.58	76.48	82.45	95.41	59.40	58.99	34.37	49.65	81.77
MITRL [Dai *et al.*, 2024b]	**99.70**	**80.48**	**62.33**	**48.47**	**81.49**	**98.66**	**97.82**	**70.02**	**60.82**	**41.06**	**59.78**	**96.50**
Proposed*(T&S)*	**99.77**	**84.51**	**62.39**	**51.23**	**87.11**	**98.79**	**99.16**	**81.93**	**61.90**	**49.27**	**84.91**	**98.60**
Improve(+)	+0.07	5.0%	+0.03	7.7%	6.9%	0.13%	0.93%	9.5%	1.16%	11.5%	7.6%	0.746%

Table 3: Comparative results with different baseline methods. Comparison of the model's performance with and without textual input. Data with horizontal lines indicate the next highest performance.

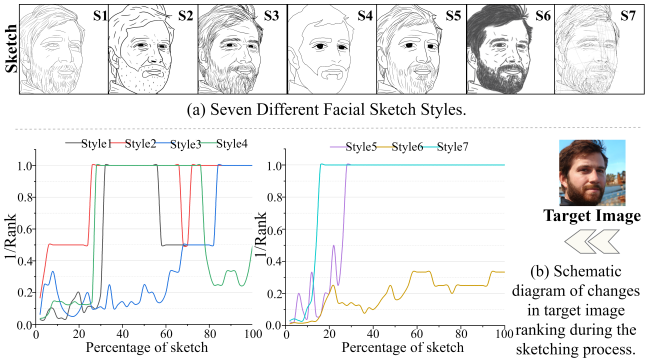
exhibits the ability to maintain efficient retrieval performance in the presence of such variations.

(2) *Complete sketch Retrieval.* As Table 2 shows, our performed excellently on datasets. On the ChairV2 dataset, our method demonstrated significant improvements compared with most existing methods. In terms of Acc.@5, the performance of our model is slightly lower than that of the StyleMeUp model but still highly competitive. On the ShoeV2 dataset, our model's advantages are more significant, showing notable improvements over the best baselines. This is likely because ChairV2 has higher similarity than ShoeV2. When considering the results comprehensively, our model outperform all other models.

Performance Analysis on Facial Sketches. Table 3 compares our model with previous models. Our method clearly outperforms others in the early retrieval stage. Notably, our method achieves significant performance gains over baseline models trained with triplet loss. This is because in the initial stages of sketching, as a result of the sparsity of strokes and frequent presence of noise, simpler methods (e.g., reducing sketch-positive sample distance) are susceptible to noise, potentially misleading the model's training direction. Additionally, incorporating textual information substantially improves retrieval performance. This improvement stems from textual information's ability to mitigate inherent randomness caused by limited sketch details. After training our approach with the FVIP language model, we observed a marked enhancement in its performance. The results demonstrate that effectively identifying interfering strokes during the drawing process and accurately matching them with the target image is crucial for improving the model's early retrieval performance.

5.2 Diverse Sketch Style Analysis

Sketch drawing styles vary significantly between individuals, with more detailed facial sketches showing even greater diversity due to increased complexity. To investigate our model's ability to manage multiple styles effectively, we extended the SKSF-A dataset to include seven distinct sketch styles, as shown in Fig.5. We present seven distinct stylistic sketches and demonstrate the retrieval process for each. The results indicate that retrieval performance exhibits significant



(a) Seven Different Facial Sketch Styles.

(b) Schematic diagram of changes in target image ranking during the sketching process.

Figure 5: Examples of sketches in seven styles and the retrieval process (Style1 as S1).

Types	m@A	m@B	w@mA	w@mB	Acc@.1	Acc@.5
Style1	93.08	46.91	57.23	26.53	44.77	80.59
Style2	94.97	51.69	58.63	29.53	50.74	81.34
Style3	94.43	48.36	58.29	27.62	50.74	82.08
Style4	92.52	42.85	56.81	24.09	44.02	82.83
Style5	95.86	54.33	59.42	31.77	52.98	82.83
Style6	86.00	32.77	52.39	18.41	38.80	71.64
Style7	92.56	47.93	56.79	27.22	44.02	75.37
Average Value	92.77	46.40	57.08	26.45	46.58	79.52

Table 4: Model retrieval performance across different sketch styles.

fluctuations across different styles. Therefore, maintaining the model's robustness against stroke noise and stylistic variations is crucial for ensuring retrieval stability.

Table 4 presents the model's retrieval performance across different sketch styles. Notably, the m@A metric maintained excellent level across all styles, generally exceeding 90%. This demonstrates the model's effective recognition capability for sketches of varying styles and its excellent ability to rank candidate results. However, we observe the model's m@B value is lowest for Style6 and peaked for Style5, exhibiting a significant 65.79% difference. The model's performance in terms of the Acc.@1 and Acc.@5 metrics aligns with the trend observed for the m@A, with Acc.@5 showing

Abs.Methods	ChairV2			FS2K-SDE1		
	m@A	m@B	Acc.@5	m@A	m@B	Acc.@5
AGMoE_N	85.92	32.61	58.26	93.60	77.38	88.35
SCDM	87.21	35.00	65.22	96.61	79.64	95.24
SCDM_AGMoE_N	88.97	36.92	70.39	98.01	80.23	96.75
SCDM_AGMoE_N_ITSC	89.62	38.20	72.26	98.95	81.90	97.24
SCDM_AGMoE	90.17	40.94	76.86	99.01	83.37	98.13
Ours-full	92.83	41.24	78.89	99.77	84.51	98.79

Table 5: Ablation Experiment.

robust performance across all styles. When comparing early retrieval performance with complete sketch retrieval, Style6 performed poorest in both m@B and Acc.@1. These results indicate that sketches in style6 may possess unique visual characteristics that pose a challenge to the model's precise identification. In summary, the experimental results reveal inherent performance differences in the model's handling of different sketch styles, underscoring the importance of considering sketch style during model training to mitigate its impact on recognition performance.

5.3 Component Ablation Analysis

To evaluate the contribution of each component, we conducted ablation experiments on four datasets. For the ChairV2 datasets, we employed the CLIP model for weight initialization, whereas facial sketches are pre-trained using FVIP. The AGMoE(_N) sub-table represents the presence or absence of added noise in the AGMoE modules. *Ours-full* represents the training configuration using all components.

The results in Table 5 indicate that the model incorporating the SCDM outperforms the AGMoE module model without the SCDM or added noise. It can be concluded that the SCDM enhances the recognition accuracy of the model, although its impact varied slightly across different metrics. For the ChairV2 datasets, models utilizing the SCDM consistently exhibited improvements across all metrics, regardless of whether noise is added. This result indicates that the MoE architecture can effectively capture and process diverse sketch strokes, ensuring a high degree of consistency in stroke quality and style. These ablation experiments confirm the positive impact of both the SCDM and AGMoE module on model performance, with their combination yielding optimal results. Across all four datasets, the proposed method demonstrated superior performance, validating the effectiveness of our framework.

5.4 Analysis Under Practical Applications

Significant variations in user drawing habits, including skill, style, and stroke length, contribute to increased stroke complexity. To simulate realistic sketching scenarios, we use three datasets with diverse styles created by numerous volunteers. Distinct sketch styles are presented in Fig. 6. Evidently, sketching styles immediately exhibit noticeable differences. However, our method demonstrates the ability to retrieve target images accurately, even with a limited number of strokes.

As shown in in Table 6, the retrieval performance of all compared methods declin across the datasets, indicating that

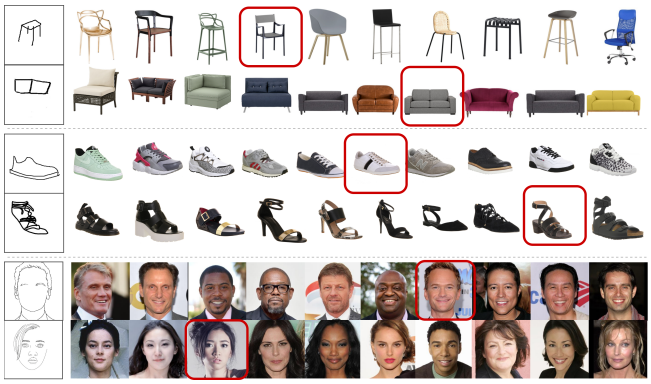


Figure 6: The target image first ranked within the top ten retrieval results during sketching. The red box highlights the target image.

Dataset	Model	m@A	m@B	w@mA	w@mB
ChairV2-U1	TriNN-RL	88.28	34.70	53.33	20.47
	LstmB	90.72	35.98	53.90	19.75
	MGAL	91.28	39.05	51.87	15.97
	PSERL/N	90.83	36.27	48.20	14.23
	Proposed	**91.98**	**40.81**	**54.69**	**21.68**
ShoeV2-U1	TriNN-RL	87.47	30.76	52.75	18.54
	LstmB	89.88	31.29	53.43	16.75
	MGAL	90.10	31.75	53.47	16.76
	PSERL/N	86.29	30.04	41.71	14.28
	Proposed	**90.25**	**32.14**	**53.90**	**16.97**
Face-U1	TriNN-C	85.38	31.29	29.21	10.94
	TriNN-SR	85.10	34.37	52.01	18.09
	TriNN-RL	79.59	29.09	47.36	15.18
	SeqL	93.27	38.91	31.68	13.51
	MGRL	73.72	25.78	45.18	14.10
	MITRL	90.02	56.79	54.93	28.94
	Proposed	**94.05**	**58.35**	**56.28**	**31.65**

Table 6: Analysis of generalization ability in practical applications. Data with horizontal lines indicate the next highest performance.

both sketch style and stroke order influence retrieval effectiveness. Regardless of whether the sketches are simple line drawings or complex facial sketches, our method exhibited superior retrieval performance on diverse and complex datasets. This outstanding performance can be attributed to the effective identification of early noisy strokes, ensuring semantic consistency throughout the sketching process and maintaining strong early retrieval performance.

6 Conclusion

We propose a novel on-the-fly noise filtering cross-modal retrieval framework designed to address the challenges of noisy stroke interference and effective feature extraction in the early stages of instant retrieval. Extensive experimental results thoroughly validate that our proposed method can effectively identify noisy strokes in various sketches, including simple line drawings and complex sketches, thereby significantly improving the initial retrieval performance. Furthermore, we have construct a multi-style dataset to comprehensively evaluate the generalization ability and robustness of the model under complex stroke scenarios.

Acknowledgements

This work is sponsored by National Nature Science Foundation of China (No.62221005 and U2336212), Nature Science Foundation of Chongqing (No.2023NSCQ-MSX0054), Intelligent Policing Key Laboratory of Sichuan Province (No.ZNJW2025KFMS005) and Doctor Student Innovative Talent Program of Chongqing University of Posts and Telecommunications (No.BYJS202405). Dawei Dai is the corresponding author.

References

[Bandyopadhyay *et al.*, 2024] Hmrishav Bandyopadhyay, Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. What sketch explainability really means for downstream tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10997–11008, 2024.

[Bhunia *et al.*, 2020] Ayan Kumar Bhunia, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. 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*, pages 9779–9788, 2020.

[Bhunia *et al.*, 2021a] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More photos are all you need: Semi-supervised learning for fine-grained sketch based image retrieval. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 4247–4256, 2021.

[Bhunia *et al.*, 2021b] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Vectorization and rasterization: Self-supervised learning for sketch and handwriting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5672–5681, 2021.

[Bhunia *et al.*, 2022a] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 999–1008, 2022.

[Bhunia *et al.*, 2022b] Ayan Kumar Bhunia, Aneeshan Sain, Parth Hiren Shah, Animesh Gupta, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Adaptive fine-grained sketch-based image retrieval. In *European Conference on Computer Vision*, pages 163–181. Springer, 2022.

[Bhunia *et al.*, 2023] Ayan Kumar Bhunia, Subhadeep Koley, Amandeep Kumar, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch2saliency: learning to detect salient objects from human drawings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2733–2743, 2023.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[Dai *et al.*, 2022] Dawei Dai, Xiaoyu Tang, Yingge Liu, Shuyin Xia, and Guoyin Wang. Multi-granularity association learning for on-the-fly fine-grained sketch-based image retrieval. *Knowledge-Based Systems*, 253:109447, 2022.

[Dai *et al.*, 2023] Dawei Dai, Yutang Li, Liang Wang, Shiyu Fu, Shuyin Xia, and Guoyin Wang. Sketch less face image retrieval: A new challenge. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Dai *et al.*, 2024a] Dawei Dai, Shiyu Fu, Yingge Liu, and Guoyin Wang. Vision-language joint representation learning for sketch less facial image retrieval. *Information Fusion*, 112:102535, 2024.

[Dai *et al.*, 2024b] Dawei Dai, Yingge Liu, Shiyu Fu, and Guoyin Wang. Multimodal image-text representation learning for sketch-less facial image retrieval. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[Fang *et al.*, 2024] Kaipeng Fang, Jingkuan Song, Lianli Gao, Pengpeng Zeng, Zhi-Qi Cheng, Xiyao Li, and Heng Tao Shen. Pros: Prompting-to-simulate generalized knowledge for universal cross-domain retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17292–17301, 2024.

[Guo *et al.*, 2017] Longteng Guo, Jing Liu, Yuhang Wang, Zhonghua Luo, Wei Wen, and Hanqing Lu. Sketch-based image retrieval using generative adversarial networks. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1267–1268, 2017.

[He *et al.*, 2024] Weikang He, Yunpeng Xiao, Tun Li, Rong Wang, and Qian Li. Interest hd: An interest frame model for recommendation based on hd image generation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):14356–14369, 2024.

[He *et al.*, 2025] Weikang He, Yunpeng Xiao, Mengyang Huang, Xuemei Mou, Rong Wang, and Qian Li. A pattern-driven information diffusion prediction model based on multisource resonance and cognitive adaptation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25. ACM, July 2025.

[Jain *et al.*, 2023] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023.

[Koley *et al.*, 2024a] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. How to handle sketch-abstraction

in sketch-based image retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16859–16869, 2024.

[Koley *et al.*, 2024b] Subhadeep Koley, Ayan Kumar Bhunia, Deeptanshu Sekhri, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. It's all about your sketch: Democratising sketch control in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204–7214, 2024.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2024] Jiachen Li, Xinyao Wang, Sijie Zhu, Chia-Wen Kuo, Lu Xu, Fan Chen, Jitesh Jain, Humphrey Shi, and Longyin Wen. Cumo: Scaling multimodal llm with co-upcycled mixture-of-experts. *arXiv preprint arXiv:2405.05949*, 2024.

[Liang *et al.*, 2021] Shuang Liang, Weidong Dai, and Yichen Wei. Uncertainty learning for noise resistant sketch-based 3d shape retrieval. *IEEE Transactions on Image Processing*, 30:8632–8643, 2021.

[Liu *et al.*, 2022] Yingge Liu, Dawei Dai, Xiaoyu Tang, Shuyin Xia, and Guoyin Wang. Bi-lstm sequence modeling for on-the-fly fine-grained sketch-based image retrieval. *IEEE Transactions on Artificial Intelligence*, 4(5):1178–1185, 2022.

[Liu *et al.*, 2024] Yingge Liu, Dawei Dai, Kenan Zou, Xiufang Tan, Yiqiao Wu, and Guoyin Wang. Prior semantic-embedding representation learning for on-the-fly fg-sbir. *Expert Systems with Applications*, page 124532, 2024.

[Muhammad *et al.*, 2018] Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning deep sketch abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8014–8023, 2018.

[Mustafa *et al.*, 2022] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576, 2022.

[Pang *et al.*, 2019] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Generalising fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 677–686, 2019.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Rao *et al.*, 2022] Yongming Rao, Wenliang Zhao, Yansong Tang, Jie Zhou, Ser Nam Lim, and Jiwen Lu. Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Advances in Neural Information Processing Systems*, 35:10353–10366, 2022.

[Riquelme *et al.*, 2021] Carlos Riquelme, Joan Puigcerver, Basil Mustafa, Maxim Neumann, Rodolphe Jenatton, André Susano Pinto, Daniel Keysers, and Neil Houlsby. Scaling vision with sparse mixture of experts. *Advances in Neural Information Processing Systems*, 34:8583–8595, 2021.

[Sain *et al.*, 2020] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Cross-modal hierarchical modelling for fine-grained sketch based image retrieval. *The British Machine Vision Conference*, pages 2–7, 2020.

[Sain *et al.*, 2021] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8504–8513, 2021.

[Song *et al.*, 2018] Jifei Song, Kaiyue Pang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning to sketch with shortcut cycle consistency. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 801–810, 2018.

[Wang *et al.*, 2025] Liang Wang, Dawei Dai, and Shiyu Fu. Multi-granularity representation learning for sketch-based dynamic face image retrieval. *Applied Intelligence*, 55(1):54, 2025.

[Yu *et al.*, 2016] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch me that shoe. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 799–807, 2016.

[Yun *et al.*, 2024] Kwan Yun, Kwanggyoon Seo, Chang Wook Seo, Soyeon Yoon, Seongcheol Kim, Soohyun Ji, Amirsaman Ashtari, and Junyong Noh. Stylized face sketch extraction via generative prior with limited data. In *Computer Graphics Forum*, page e15045. Wiley Online Library, 2024.

[Zhou *et al.*, 2024] Yanghong Zhou, Dawei Liu, and PY Mok. Zero-shot sketch based image retrieval via modality capacity guidance. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1780–1787, 2024.

[Zuo *et al.*, 2024] Ran Zuo, Haoxiang Hu, Xiaoming Deng, Cangjun Gao, Zhengming Zhang, Yukun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenediff: Generative scene-level image retrieval with text and sketch using diffusion models. 2024.