ASSA: Augmented-view contrastive and Stroke Self-Attention for On-The-Fly Fine-Grained SBIR

Tuan Nguyen Huu Long Dang Hoang Anh Nguyen Thi Tu Phuong Tu Minh
Posts and Telecommunications Institute of Technology, Vietnam
{tuannh, longdh, phuongtm}@ptit.edu.vn, anhntt.b22cn034@stu.ptit.edu.vn

Abstract

The On-the-fly Fine-grained Sketch-Based Image Retrieval (FG-SBIR) task aims to retrieve target images as the user draws, thereby reducing search time. This task faces two key challenges: (1) the large structural and visual gap between sketches and photos, which requires learning domain-invariant representations, and (2) variations in drawing order across users, which limit the effectiveness of sequential stroke modeling. To address these issues, we propose ASSA, a two-stage framework that combines Augmented-view Contrastive Learning and Stroke Self-Attention. The first stage enhances cross-domain alignment through dual geometric and photometric augmentations, while the second captures semantic correlations among strokes without relying on stroke order. Experiments on QMUL-Chair-V2 and QMUL-Shoe-V2 show that ASSA achieves up to 20% higher m@B and 15% higher wm@B than the best previous model. These results establish a new state-of-the-art in early retrieval performance for the on-the-fly FG-SBIR task.

1. Introduction

In recent years, the rapid proliferation of touchscreen and smart devices has enabled users to easily create hand-drawn sketches in daily life, work, and entertainment. Unlike textual queries, sketches provide an intuitive and abstract visual representation of objects, which makes them a natural medium for image retrieval tasks. Consequently, Sketch-Based Image Retrieval (SBIR) has emerged as a significant research topic in computer vision, aiming to bridge the gap between human-drawn sketches and natural photos through cross-domain feature learning. [9, 11, 15, 16, 25, 29, 31, 33]

Among existing SBIR frameworks, Fine-Grained SBIR (FG-SBIR) [2, 22, 23, 26] focuses on distinguishing between instances within the same category (e.g., retrieving a specific shoe or chair) rather than simply identifying broad object categories. Despite recent advances, the practical application of FG-SBIR remains limited because users must typically finish a complete sketch before any retrieval can occur. This process is not only time-consuming but also requires considerable drawing skill, making it less suitable for real-world scenarios.

To address these limitations, Bhunia et al.[1] introduced the concept of On-the-fly FG-SBIR, in which the system retrieves candidate images progressively as the user draws. This dynamic retrieval framework significantly reduces search latency and allows users to terminate the sketching process once the target image appears. Subsequent studies modeled the correlations among strokes based on their sequential order using sequence learning architectures, while additionally incorporating semantic or local–global features to enhance representation learning [7, 8, 19, 20]. However, this setting introduces new challenges: early-stage sketches often contain very limited structural information, and different drawing styles among users make it difficult to model sequential stroke dependencies.

Solving this problem, in this paper, we present ASSA (Augmented-view contrastive and Stroke Self-Attention) — a novel two-stage framework for the on-the-fly fine-grained SBIR task. Our approach integrates two complementary ideas: (1) an Augmented-view Contrastive Learning phase that enhances feature robustness by learning invariant representations across geometric and photometric transformations, and (2) a Stroke Self-Attention (SSA) module that captures semantic correlations among incomplete strokes without relying on their drawing order. Together, these components enable more effective feature alignment between sketches and photos and significantly improve retrieval accuracy in the early sketching stages. The early retrieval effectiveness of our model is visually illustrated in Figure 1, in comparison with other methods. The main contributions of this paper are summarized as follows:

- We proposed an Augmented-view Contrastive Learning phase that enhances feature robustness by constructing two complementary augmentation sets focusing on invariant features, enabling effective cross-domain alignment between sketches and photos.

Figure 1. Illustration of our method's capability to retrieve the target photo (top-10 list) with fewer strokes compared to SOTA methods such as RL-B[1], MGAL[7], LGRL[8] and PSRL[20]. The term T denotes the number of strokes.

- A progressive NT-Xent loss integration strategy is introduced, where the contrastive loss is applied after a warm-up with Triplet Loss, and the influence of loss weight (α and β) is systematically analyzed.
- We design a Stroke Self-Attention (SSA) module to capture semantic correlations among strokes without relying on sequential order, enhancing relational feature learning.
- Comprehensive experiments on two public benchmarks (QMUL-Chair-V2 and QMUL-Shoe-V2) verify that both phases significantly boost early retrieval, and their combination achieves state-of-the-art performance.

2. Related Works

2.1. Fine-grained SBIR

In the Fine-grained Sketch-Based Image Retrieval (FG-SBIR) task, each individual instance is treated as a distinct category [2, 17, 27]. To achieve such fine-grained discrimination, models must learn highly distinctive instance-level features instead of generic category representations. Yu et al. [32] first introduced deep learning into this domain through a triplet ranking network, enabling joint learning of sketch–photo embeddings. Subsequently, Liu et al. [18] explored graph embedding techniques to better capture structural relationships for fine-grained, scene-level retrieval, while Wang et al. [30] proposed a deep cascaded cross-modal ranking model that integrates complementary sketch and image information to enhance retrieval efficiency and top-K accuracy. Despite these advancements, the progress of FG-SBIR remains constrained by limited training data. To address data scarcity, Bhunia et al. [2] proposed a semi-supervised cross-modal retrieval framework that generates synthetic sketch–photo pairs using a sequential photo-to-sketch generation model. However, this approach still struggles to generalize well across diverse sketching styles. In practice, sketches vary widely in style and abstraction level, which significantly impacts retrieval performance. To mitigate this, Sain et al. [24] designed a style-agnostic deep SBIR model capable of dynamically adapting to user-specific styles, while Bhunia et al. [3] developed a model-agnostic meta-learning framework to further enhance the adaptability of FG-SBIR across sketch categories and user styles. Jianan et al [14] advanced this field by introducing a self-supervised FG-SBIR framework that unifies sample feature alignment with multi-scale token recycling, allowing the model to learn robust and semantically consistent representations without explicit sketch–photo pair annotations. However, these approaches still depend on complete sketches, making them less suitable for on-the-fly retrieval scenarios, where early and incomplete inputs are more realistic.

2.2. On-the-fly Finegrained SBIR

Traditional FG-SBIR frameworks assume that users provide a complete and detailed sketch before retrieval begins. However, this assumption limits their practicality, as drawing a full sketch is time-consuming and often unnecessary for identifying the target image. To address this limitation, Bhunia et al. [1] introduced the concept of On-the-fly FG-SBIR, where retrieval is performed dynamically as the user

draws. This study using reinforcement earning to optimize early retrieval — enabling the system to retrieve the correct photo with as few sketch strokes as possible by introducing a novel reward scheme that ensures consistent and efficient cross-modal ranking throughout the sketching process. Liu et al. [19] proposed an on-the-fly FG-SBIR model that treats the sketching process as a sequential problem, where a Bi-LSTM network is used to learn temporal correlations among incomplete sketches, thereby optimizing the embedding space and enhancing early retrieval performance. Dai et al. [7] proposed a Multi-Granularity Association Learning (MGAL) framework for the on-the-fly FG-SBIR task, in which the embedding space of incomplete sketches is guided to approximate that of subsequent sketches and the corresponding target photo. The LGRL (Local-Global Representation Learning) [8] framework introduces a joint representation learning approach that integrates both local and global features for on-the-fly FG-SBIR, enabling the model to capture fine-grained stroke details as well as overall sketch context. Meanwhile, the Prior Semantic-Embedding Representation Learning method [20] leverages prior semantic knowledge by embedding global semantic information into the feature space of incomplete sketches. These studies share a common characteristic — they either exploit the sequential dependencies among strokes during the sketching process or capture the local–global features of sketches. However, a major limitation arises from the fact that, although strokes are correlated, users may adopt highly diverse drawing orders to create a sketch, making stroke sequentiality inherently limited. Moreover, previous works often overlook the invariant features of real images, resulting in less effective feature extraction and consequently reduced early retrieval performance.

3. Methodology

We present ASSA (Augmented-view Contrastive and Stroke Self-Attention), a two-stage learning framework for on-the-fly fine-grained sketch-based image retrieval. The goal of ASSA is to jointly improve feature alignment between sketches and photos and enhance retrieval performance at early sketching stages. ASSA consists of two complementary parts: an Augmented-view Contrastive Learning stage that builds a discriminative and robust embedding space, and a Stroke Sel-Attention (SSA) module that models relationships among strokes without enforcing a fixed drawing sequence. An overall view of this approach is shown in Figure 2.

3.1. Augmented-view contrastive

Baseline. In this phase, the goal is to enhance the effectiveness of early image retrieval by enriching feature representations using Augmented-view for both sketch images and their corresponding target photos (Figure 2 - left). We design a model consisting of three branches without weight sharing. Specifically, the positive and negative photo branches share the same weights, while the complete sketch branch uses a separate set of weights. This approach has been shown to outperform the strategy of not sharing weights and using a 1×1 convolution followed by a softmax operation. [12]

First, each input image is passed through two different sets of augmentation, called $\mathcal{T}_1$ and $\mathcal{T}_2$ functions, to help the model learn features that are invariant to appearance changes. To enhance the generalization ability of visual feature learning models, the network needs to learn invariant features that are robust to common real-world variations. In particular, the model should be invariant to geometric transformations and invariant to color changes. So, $\mathcal{T}_1$ is the combination of two augmentations: Crop, Rotation and Flip - focus on shape-based invariant features, while $\mathcal{T}_2$ is the combination of Cutout, Color Jilter, Gray scale, and Gaussian blur - focus on Photometric-invariant features. These enhancements have been shown to be effective in previous studies [4–6] and are reused in our work. Here, although Cutout is a shape-based augmentation, combining it with other geometric transformations may cause loss of structural features and hinder shape invariance learning. Therefore, it is included in the $\mathcal{T}_2$ function in our study. The impact of varying the set of augmentations between the two functions, $\mathcal{T}_1$ and $\mathcal{T}_2$, is presented in Section 4.2.

The first component of this phase is an CNN network, here we using InceptionNetV3 [28], pre-trained on the ImageNet dataset, which is used to extract features from the input images. This process is formulated in Equation 1, where x_1 and x_2 denote the two augmented versions of the input, and B_1, B_2 represents the extracted features obtained after passing through CNN (for both photos and sketches).

$$\begin{aligned} B_1 &= CNN(x_1) \\ B_2 &= CNN(x_2) \end{aligned} \tag{1}$$

The second component is a self-attention block, which models the pairwise relationships among feature points in the output feature maps of CNN. As expressed in Equation 2, f_{att} is implemented using a standard Multi-Head Attention module, while denotes A_p is the Average Pooling operation.

$$\begin{aligned} V_1 &= A_p(B_1 + B_1 \times f_{att}(B_1)) \\ V_2 &= A_p(B_2 + B_2 \times f_{att}(B_2)) \end{aligned} \tag{2}$$

After that, we apply a simple linear layer follow by normalize method to reduce the high-dimensional feature vector (V_1, V_2) into a lower-dimensional representation (v_1, v_2). The output of each branch consists of two corresponding feature vectors: v_{p_1} and v_{p_2} (for positive photo branch), v_{n_1} and v_{n_2} (for negative photo branch), v_{s_1} and v_{s_2} (for

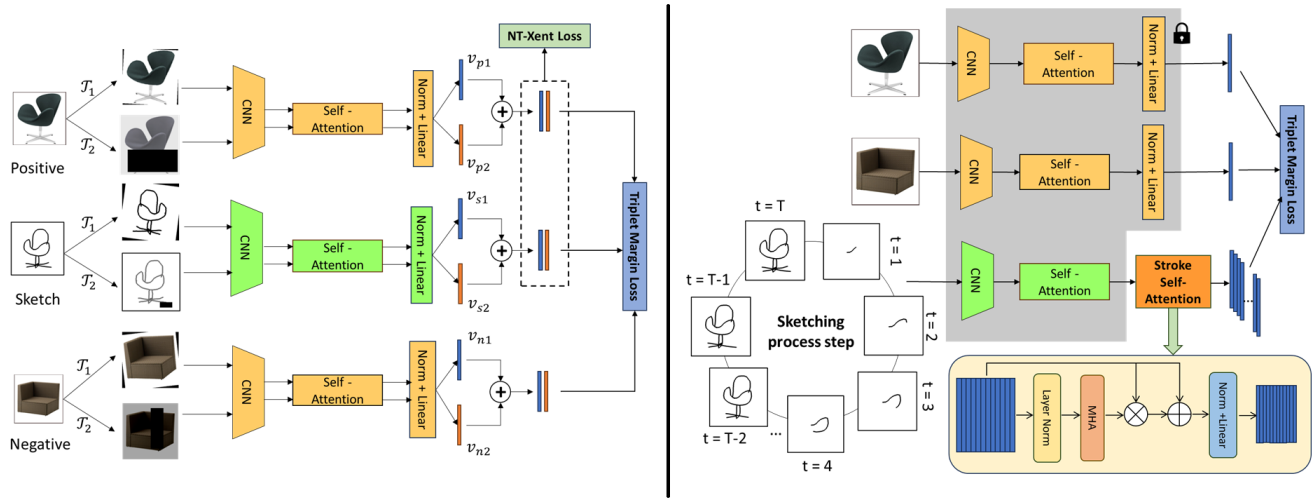

Figure 2. (1) Phase 1 – FG-SBIR model based on Augmented-view contrastive, using NT-Xent Loss and Triplet Margin Loss. (2) The Stroke Self-Attention is used to learn the relational features among incomplete strokes. The weights of the other part are frozen

complete sketch branch). The two output vectors from each branch are concatenated into a single vector, resulting in more informative representations of the input images. Consequently, three unique output vectors are obtained: positive vector (v_p), negative vector (v_n) and complete sketch vector (v_s).

Loss function. The Triplet loss (See 3) was employed to optimize a joint embedding space between photos and their corresponding sketches, where only the complete sketch was taken into account during this phase. To enhance training stability and ensure progressive feature alignment, the model is first pre-trained using only the Triplet loss to establish a discriminative embedding space. After convergence, the NT-Xent loss (See 4) is incorporated to further refine the learned representations by enforcing stronger intra-class compactness and inter-class separation. The NT-Xent loss is applied to v_p and v_s to pull them closer together, with the aim of improving retrieval performance.

$$\mathcal{L}_{triplet} = max\{d(v_s, v_p) - d(v_s, v_n) + \lambda, 0\} \quad (3)$$

$$\mathcal{L}_{NT-Xent}(i,j) = -log\frac{exp((v_i^T v_j)/\sigma)}{\sum_{k=1}^{2B} \mathbf{1}_{[k \neq i]} exp((v_i^T v_k)/\sigma)} \quad (4)$$

Thus, the loss function of Phase 1 is defined as:

$$\mathcal{L} = \alpha \mathcal{L}_{triplet} + (1-\alpha)\mathcal{L}_{NT-Xent} \quad (5)$$

Here, α and β are the weights of each loss function. The experimental results at 4.2 demonstrate that this strategy produces optimal performance.

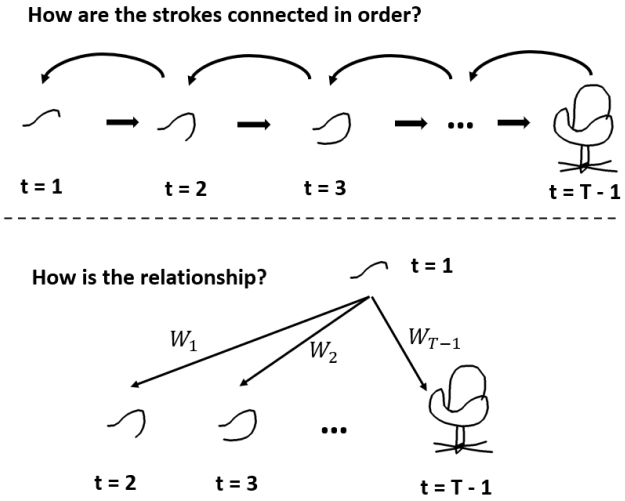

Figure 3. Comparison between sequential dependency modeling (top) and the proposed Stroke Self-Attention (SSA) mechanism (bottom). The sequential model connects strokes only in drawing order, whereas SSA captures pairwise relationships among all strokes without relying on temporal order.

3.2. Stroke Self-Attention

We use the results of phase 1 to continue learning features for incomplete sketches, with the goal of improving retrieval efficiency at an early stage. As shown in Figure 2 - right, the sketching process is represented as a sequence of incomplete sketches arranged in order. In the second training phase, the weights of f_1 and f_2 across all branches, as well as f_3 in the positive and negative branches, are kept fixed. The output of f_2 in the sketch branch is $V_s \in \mathcal{R}^{N \times D}$,

that N denotes the number of incomplete strokes and D is the high dimension vector of each stroke feature. These stroke-level embeddings are then fed into the proposed Sequential Sketch Attention (SSA) module, which captures the contextual relationships among strokes and refines their representations. As shown in Figure 3, SSA learns global correlations among strokes without relying on their sequential order.

Specifically, the input feature matrix V_s is first normalized via Layer Normalization, ensuring a stable feature distribution and improving convergence:

$$\hat{V_s} = LayerNorm(V_s) \tag{6}$$

Subsequently, $\hat{V_s}$ is processed by a Multi-Head Self-Attention (MHA) mechanism, which computes self-attention weights to model inter-stroke dependencies and emphasize structurally informative regions:

$$Z = MHA(\hat{V_s}) \tag{7}$$

To preserve original spatial semantics, a residual connection is incorporated, where the updated features are expressed as:

$$V_s = Z \odot V_s + V_s \tag{8}$$

Where, $\odot$ is element-wise. denotes element-wise multiplication. This design ensures a balance between newly learned relational information and the original representations. Finally, a linear projection layer maps the refined features into a compact embedding space, producing $v_s \in \mathcal{R}^{D_{out}}$, where D_{out} is is the lower-dimensional output. This structure allows the model to focus adaptively on semantically relevant strokes, while maintaining robustness to stroke order variations—thereby improving early-stage retrieval performance without overfitting to specific sketching patterns.

4. Experiments

Datasets. Currently, there exist only two publicly available datasets for the on-the-fly FG-SBIR task, namely the QMUL-Shoe-V2 and QMUL-Chair-V2 datasets, rasterized by Bhunia [32], which were employed to construct our experimental setup for performance evaluation. The QMUL-Shoe-V2 dataset consists of 6,730 sketches and 2,000 photos, of which 6,051 sketches and 1,800 photos were used for training, while the remainder served as the test set. Similarly, the QMUL-Chair-V2 dataset comprises 2,000 sketches and 400 photos, with 1,275 sketches and 300 photos used for training and the rest reserved for testing.

Implementation Details. Our model is implemented in the PyTorch framework on a single NVIDIA Tesla P100 GPU with 16 GB memory provided free by Kaggle. The input images are resized to 299 $\times$ 299, and the resulting embedding vectors have a dimensionality of 64, with the temperature parameter $\mathcal{T}$ set to 0.07 and margin = 0.3. The sketch images in the standard datasets were rendered with T = 20 steps. Our model training for 400 epochs, using a batch size of 24 (both phases) for ChairV2, the AdamW optimizer [21] with a learning rate of 1e-4 for the first 200 epochs and reduced to 1e-5 for the next 200 epochs, in both phases. In Phase 1, we control the balance between the Triplet Loss and the NT-Xent Loss using a single parameter α, where the total weight is defined as $(\alpha, 1-\alpha)$. We vary α within the range $[0, 1]$ with a step of 0.1 to investigate its influence on retrieval performance and determine the optimal trade-off between the two objectives. Additionally, as previously described, we incorporate the NT-Xent Loss after Phase 1 has been trained with the Triplet Loss for 200 epochs.

Evaluation Metric. In the context of the on-the-fly FG-SBIR task, we focus on whether the correct photo is ranked near the top of the retrieval results. To evaluate this, we compute the proportion of sketches whose corresponding images appear within the top-q positions, referred to as A@q. In addition, two complementary indicators, m@A (the average ranking percentile) and m@B (the inverse rank normalized by the number of sketches) [1], are employed to measure how effectively the system retrieves matches during the earlier stages of sketch completion. It should be emphasized that both m@A and m@B provide an overall assessment across the entire sketching sequence.

$$w_i = e^{-\frac{P_i}{P_n}} \tag{9}$$

$$\text{wm@A} = \frac{\sum_i^m \sum_j^n w_{ij}(1 + \frac{1-rank_{ij}}{m-1})}{m.n} \tag{10}$$

$$\text{wm@B} = \frac{\sum_i^m \sum_j^n w_{ij} \frac{1}{rank_{ij}}}{m.n} \tag{11}$$

To better evaluate early retrieval performance, we adopt the weighted variants of m@A and m@B, called wm@A and wm@B proposed by Dai et al [8] (see 9, 10 and 11). Here, p_i and p_n denote the number of strokes in the current sketch and the sketch at step n; n is the number of intermediate sketches per photo; m refers to the number of photos in the test set and $rank_{ij}$ represents the rank of the i-th target photo retrieved using the j-th sketch. Higher metric values indicate better performance. While m@A and m@B reflect the average performance across all sketching stages, wm@A and wm@B put greater emphasis on retrieval quality at the earlier stages.

Comparison methods. The comparative experiments involve multiple baseline and state-of-the-art methods.

Table 1. Comparative results are reported against multiple baseline approaches. In particular, A@5 and A@10 denote the top-5 and top-10 retrieval accuracies evaluated on the complete sketches (at t = T). Meanwhile, m@A, m@B, wm@A, and wm@B are employed to assess the retrieval performance over the entire sketching sequence.

	QMUL-Chair-V2						QMUL-Shoe-V2					
	m@A	m@B	wm@A	wm@B	A@5	A@10	m@A	m@B	wm@A	wm@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
B4	81.24	29.85	-	-	75.14	87.69	81.02	19.50	-	-	62.34	77.24
B5	83.22	36.19	49.05	19.07	-	-	85.46	23.35	50.89	12.02	-	-
B6	87.11	38.20	52.56	20.43	-	-	86.55	24.80	51.79	14.11	-	-
TS	76.01	27.64	-	-	73.47	85.13	77.12	17.13	-	-	62.67	76.47
RL-B	85.44	35.09	51.05	18.41	76.34	89.65	85.38	21.44	50.06	11.96	65.77	79.63
LSTM-B	89.54	38.57	54.06	21.17	79.87	91.33	88.70	24.00	53.57	12.38	62.16	77.02
MGAL-B	88.95	39.16	53.60	21.25	81.73	92.56	88.58	24.24	53.49	12.79	65.31	78.22
PSRL-B	92.79	41.18	56.86	24.05	-	84.12	95.58	33.42	58.76	19.69	-	78.53
LGRL-B	91.47	45.21	55.41	25.07	86.37	**94.73**	89.99	25.87	54.51	13.58	60.81	75.07
Our	**96.98**	**64.32**	**60.31**	**40.47**	**86.69**	93.19	**96.20**	**45.39**	**59.27**	**29.34**	**67.12**	**82.13**

Table 2. Comparative performance across different feature embedding configurations

Dimension	Model	QMUL-Chair-V2				QMUL-Shoe-V2			
		m@A	m@B	wm@A	wm@B	m@A	m@B	wm@A	wm@B
32	RL-B	82.61	34.67	-	-	82.94	19.61	-	-
	LSTM-B	87.86	33.08	52.96	18.21	88.35	23.89	53.45	12.89
	MGAL-B	86.52	34.07	51.88	18.47	88.26	24.76	53.28	13.08
	PSRL-B	93.73	43.42	57.45	25.44	95.43	33.94	58.65	20.00
	LGRL-B	91.10	43.92	55.31	25.20	89.04	24.31	53.84	13.03
	Our	**96.63**	**59.53**	**59.88**	**36.61**	**95.81**	**39.83**	**59.23**	**26.58**
64	RL-B	85.44	35.09	51.05	18.41	85.38	21.44	50.06	11.96
	LSTM-B	89.54	38.57	54.06	21.17	88.70	24.00	53.57	12.38
	MGAL-B	88.95	39.16	53.60	21.25	88.58	24.24	53.49	12.79
	PSRL-B	92.79	41.18	56.86	24.05	95.58	33.42	58.76	19.69
	LGRL-B	91.47	45.21	55.41	25.07	89.99	25.87	54.51	13.58
	Our	**96.98**	**64.32**	**60.31**	**40.47**	**86.69**	**45.39**	**59.27**	**29.34**
128	RL-B	84.71	34.49	-	-	84.61	20.81	-	-
	LSTM-B	88.71	35.71	53.23	19.37	89.31	25.90	54.06	13.91
	MGAL-B	88.98	38.13	-	-	88.21	25.79	-	-
	PSRL-B	93.87	44.52	57.55	25.82	95.24	32.84	58.56	19.34
	LGRL-B	90.89	43.08	54.68	23.73	89.90	25.87	54.51	13.58
	Our	**96.45**	**62.29**	**60.03**	**38.57**	**95.52**	**38.96**	**58.92**	**24.33**

Table 3. Retrieval performance after incorporating the NT-Xent Loss following N epochs of training in phase 1

N epochs	QMUL-Chair-V2				QMUL-Shoe-V2			
	m@A	m@B	wm@A	wm@B	m@A	m@B	wm@A	wm@B
0	94.88	54.64	58.63	34.12	90.51	39.51	56.51	26.35
100	94.96	58.61	59.46	37.05	94.16	40.98	58.15	26.53
150	96.44	61.87	59.97	39.10	94.25	42.20	57.98	27.71
200	**96.98**	**64.32**	60.31	**40.47**	**96.20**	45.39	59.27	**29.34**
250	96.94	63.80	**60.32**	40.02	96.07	**46.21**	**59.39**	29.32
300	96.86	62.39	60.13	37.87	95.80	45.54	59.18	28.94

B1[32][26] represents a standard FG-SBIR model trained with triplet loss using only complete sketches. **B2** extends B1 by including all intermediate sketches to improve generalization. **B3** trains twenty separate models, each specialized for a certain sketch completion level. **B4**[10] combines triplet and ranking losses within a unified deep framework to optimize retrieval across T sketching stages. **B5** adopts the LGRL[8] backbone but is trained only on complete sketches, while **B6** follows the ASSA baseline trained with all stroke information. In **TS**[13], incomplete sketches are first reconstructed via an image-to-image translation network before retrieval. **RL-B**[1] applies reinforcement learning to refine sketch embeddings obtained from the CNN backbone. **LSTM-B**[19] employs a bidirectional LSTM to model stroke sequences, and **MGAL-B**[7] enhances partial sketch embeddings via multi-level association learning. **LGRL-B**[8] learns joint local–global representations from incomplete sketches and target photos, whereas **PSRL-B**[20] further integrates semantic embeddings to strengthen representations and

Table 4. Retrieval performance when varying the balance parameter α in $[0, 1]$, where the loss weights are defined as $(\alpha, 1-\alpha)$ for Triplet Loss and NT-Xent Loss, respectively

α	$1-\alpha$	QMUL-Chair-V2				QMUL-Shoe-V2			
		m@A	m@B	wm@A	wm@B	m@A	m@B	wm@A	wm@B
1	0	96.84	62.94	60.14	39.81	95.76	43.57	59.03	27.76
0.9	0.1	96.98	**64.32**	**60.31**	**40.47**	**96.20**	**45.39**	**59.27**	**29.34**
0.8	0.2	**97.15**	64.12	60.24	39.39	92.82	42.65	57.27	27.70
0.7	0.3	96.95	61.99	60.27	39.67	92.95	43.08	57.10	28.11
0.6	0.4	96.92	60.54	60.22	38.50	93.04	41.95	57.53	27.59
0.5	0.5	96.94	63.89	60.22	39.42	92.52	40.30	56.95	27.01
0.4	0.6	96.97	61.86	60.19	39.43	95.37	39.19	58.82	25.69
0.3	0.7	96.53	61.81	59.86	37.30	95.84	37.89	59.29	25.23
0.2	0.8	96.85	61.44	59.97	38.58	95.86	40.19	59.08	25.73
0.1	0.9	96.76	59.00	60.11	36.71	95.28	38.83	58.95	25.83
0	1	95.62	56.85	59.65	34.16	90.10	39.42	56.43	26.30

Table 5. Comparison of the performance of other state-of-the-art (SOTA) methods and the baseline with and without augmented-view on the Chair-V2 and Shoe-V2 datasets.

	Use Augmented	QMUL-Chair-V2				QMUL-Shoe-V2			
		m@A	m@B	wm@A	wm@B	m@A	m@B	wm@A	wm@B
RL-B	✗	85.44	**35.09**	51.05	18.41	85.38	21.44	50.06	11.96
RL-B	✓	**86.37**	33.56	**52.47**	**18.69**	**86.18**	**23.06**	**51.43**	**12.85**
LSTM-B	✗	89.54	38.57	54.06	21.17	88.70	24.00	53.57	12.38
LSTM-B	✓	**95.23**	**47.99**	**59.32**	**30.07**	**94.79**	**38.72**	**58.59**	**25.02**
MGAL-B	✗	88.95	39.16	53.60	21.25	88.58	24.24	53.49	12.79
MGAL-B	✓	**90.40**	**40.27**	**55.74**	**23.33**	**90.08**	**25.80**	**54.61**	**14.49**
PSRL-B	✗	92.79	41.18	56.86	24.05	95.58	33.42	58.76	19.69
PSRL-B	✓	**96.12**	**49.78**	**59.35**	**31.06**	**95.94**	**42.95**	**59.14**	**28.56**
ASSA (Our)	✗	96.15	56.69	59.63	36.02	95.51	43.12	59.16	27.76
ASSA (Our)	✓	**96.98**	**64.32**	**60.31**	**40.47**	**96.20**	**45.39**	**59.27**	**29.34**

boost early retrieval accuracy.

4.1. Results and Analysis

Table 1 presents a comparison between our proposed method (ASSA) for the on-the-fly SBIR task and the state-of-the-art (SOTA) methods on two datasets: QMUL-Chair-V2 and QMUL-Shoe-V2. The proposed method significantly outperforms all state-of-the-art (SOTA) methods in this task in terms of early sketch retrieval performance.

(i) QMUL-Chair-V2: Our model establishes a new state-of-the-art performance, achieving substantially higher early retrieval accuracies than the two strongest competitors for the on-the-fly FG-SBIR task, namely PSRL and LGRL. Specifically, in terms of the m@B and wm@B metrics, our method surpasses LGRL by approximately 20% and 15%, respectively, validating our hypothesis that capturing relational correlations among strokes is more beneficial than modeling their sequential dependencies. Furthermore, the A@5 and A@10 metrics evaluate retrieval accuracy based on complete, well-drawn sketches, and thus serve as secondary indicators in the context of on-the-fly SBIR. Although our approach exhibits a marginal decrease in A@10 compared to prior methods, this does not undermine the overall effectiveness of the proposed framework. Since the primary objective of on-the-fly FG-SBIR is to rapidly retrieve the target image with as few strokes as possible, performance during the early sketching phase remains the most critical factor.

(ii) QMUL-Shoe-V2: The improvement in accuracy on this dataset is comparable to that on the QMUL-Chair-V2 dataset, demonstrating the stability of our method when applied to larger-scale data. This result highlights the promising potential of our model for object retrieval tasks on large datasets, such as those commonly found in e-commerce environments.

Furthermore, we also evaluated the performance of the proposed method by varying the dimensionality of the feature embedding space. The results in Table 2 indicate that our approach consistently outperforms other state-of-the-art methods across most embedding dimensions. In addition, the experimental results show that changes in the embedding space have little impact on the early retrieval perfor-

Table 6. Comparison of different augmentation strategies on the Chair-V2 and Shoe-V2 dataset.

Aug. ID	$\mathcal{T}_1$	$\mathcal{T}_2$	QMUL-Chair-V2				QMUL-Shoe-V2			
			m@A	m@B	wm@A	wm@B	m@A	m@B	wm@A	wm@B
Augment - 1	**(1) - (6)**	**(7)**	96.24	58.41	59.91	37.91	95.86	41.12	59.18	26.83
Augment - 2	**(1) - (5)**	**(6) - (7)**	**96.99**	61.28	59.99	37.23	95.72	42.85	59.23	27.97
Augment - 3	**(1) - (4)**	**(5) - (7)**	96.80	62.97	60.10	38.12	95.95	41.63	59.23	27.34
Augment - 4	**(1) - (3)**	**(4) - (7)**	96.98	**64.32**	**60.31**	**40.47**	**96.20**	**45.39**	**59.27**	**29.34**
Augment - 5	**(1) - (2)**	**(3) - (7)**	96.40	59.67	59.92	36.10	95.87	43.37	59.18	28.62
Augment - 6	**(1)**	**(2) - (7)**	96.46	58.75	59.60	34.79	95.17	41.54	58.76	60.51

mance of our method.

4.2. Ablation Study

Impact of Triplet Loss and NT-Xent Loss: We analyze the impact of combining Triplet Loss and NT-Xent Loss on early retrieval performance by adjusting their respective weights (α and $1-\alpha$) during training. Here, α and $1-\alpha$ represent the weighting coefficients for Triplet Loss and NT-Xent Loss, respectively, and we vary α in $[0, 1]$ with a step of 0.1 to study its impact. As shown in Table 4 we observe that: **(i)** the early retrieval performance improves consistently as both losses are jointly optimized, with the best results obtained when α is around 0.9. **(ii)** When α becomes low (the contribution of Triplet Loss decreases), the performance metrics tend to decline. This suggests that Triplet Loss still plays a dominant role in maintaining the relational structure among samples, while NT-Xent Loss contributes to enhancing the global discriminative power of the learned representations. Additionally, we conduct a brief study on the integration strategy of NT-Xent Loss after a warm-up period with Triplet Loss. Results in Table 3 show that introducing NT-Xent after a moderate warm-up (around 200 epochs) yields the best performance, while excessively long warm-up periods provide no further benefit.

Influence of Augmented-view and SSA module: We conducted an analysis to examine the impact of phase 1 and phase 2 of our proposed method on early retrieval performance. For the SOTA methods that share the same phase 1 framework, including RL-B, LSTM-B, MGAL-B, and PSRL-B, we replaced their original phase 1 with our augmented-view strategy. The results presented in Table 5, show that in most cases, using our proposed approach leads to a clear improvement in retrieval performance, particularly in the m@B and wm@B metrics. This indicates that when the feature representation becomes more diverse and informative, the ability to correctly retrieve the target image from early sketch inputs is significantly enhanced. Furthermore, we also evaluated the effect of the SSA module on relational feature learning by using the original phase 1 of previous methods while applying our phase 2. The results demonstrate that even without the influence of our phase 1, the independent use of phase 2 still enables our model to outperform previous approaches. These findings suggest that employing a correlation learning model that does not rely on the sequential order of sketch strokes provides better early retrieval performance than methods that depend on sequential learning.

Effect of different augmentation strategies: We measured the impact of using different combinations of augmentation operations in each view, $\mathcal{T}_1$ and $\mathcal{T}_2$. As described in Section 3.1, the augmentations used in our study are indexed as follows: Crop**(1)**, Flip**(2)**, Rotation**(3)**, ColorJitter**(4)**, Gray**(5)**, Gaussian Blur**(6)**, and Cutout**(7)** Since the total number of possible combinations among these augmentations is extremely large, we did not evaluate every combination but instead measured the performance based on the number of augmentations applied in each view. The results obtained when using augmentation operations from **(1)** to **(7)** are shown in Table 6. It can be observed that when the number of augmentations between the two views is imbalanced, the early retrieval performance reaches its lowest values and gradually improves as the two views become balanced. The difference in performance is quite noticeable, ranging from 1% to 6%, depending on the specific early retrieval metrics. It is important to note that there are many possible augmentation types, each affecting the feature learning process differently. Therefore, in our experiments, we only selected commonly used augmentations that have been proven effective in previous studies. Nevertheless, changing the specific augmentation operations does not significantly affect the state-of-the-art performance of our proposed method compared to previous approaches.

5. Conclusion

We introduced ASSA, a two-stage framework that integrates augmented-view contrastive learning and stroke self-attention for on-the-fly fine-grained SBIR. The proposed approach learns invariant representations while capturing semantic correlations among strokes without relying on their sequential order. Experiments on QMUL-Chair-V2 and QMUL-Shoe-V2 confirm that ASSA achieves state-of-the-art performance, particularly in early retrieval stages. Future work will explore extending our method to larger and more diverse datasets and incorporating interactive guidance for real-time sketch-based retrieval.

References

[1] 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 *CVPR*, 2020. 1, 2, 5, 6

[2] 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 *CVPR*, 2021. 1, 2

[3] Ayan Kumar Bhunia, Aneeshan Sain, Parth Shah, Animesh Gupta, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Adaptive fine-grained sketch-based image retrieval. In *ECCV*, 2022. 2

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[5] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR*, 2019.

[6] Ekin D. Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. In *CVPR*, 2020. 3

[7] Dawei Dai, Xiaoyu Tang, Shuyin Xia, Yingge Liu, Guoyin Wang, and Zizhong Chen. Multi-granularity association learning for on-the-fly fine-grained sketch-based image retrieval. *Knowledge-Based Systems*, 253:109447, 2022. 1, 2, 3, 6

[8] Dawei Dai, Yingge Liu, Yutang Li, Shiyu Fu, Shuyin Xia, and Guoyin Wang. Lgrl: Local-global representation learning for on-the-fly fg-sbir. *IEEE Transactions on Big Data*, 10(4):543–555, 2024. 1, 2, 3, 5, 6

[9] Sounak Dey, Pau Riba, Anjan Dutta, Josep Lladós, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, 2019. 1

[10] Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. Sodeep: a sorting deep net to learn ranking loss surrogates. In *CVPR*, 2019. 6

[11] Cusuh Ham, Gemma Canet Tarres, Tu Bui, James Hays, Zhe Lin, and John Collomosse. Cogs: Controllable generation and search from sketch and style. In *ECCV*, 2022. 1

[12] Tuan Nguyen Huu and Quynh Dao Thi Thuy. Unshared weight combine with self-attention at base model for on-the-fly fine-grained sbir. *SIViP*, 19, 2025. 3

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 6

[14] Jianan Jiang, Hao Tang, Zhilin Jiang, Weiren Yu, and Di Wu. Arnet: Self-supervised fg-sbir with unified sample feature alignment and multi-scale token recycling. In *AAAI*, 2024. 2

[15] 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 *CVPR*, 2024. 1

[16] Fengyin Lin, Mingkang Li, Da Li, Timothy Hospedales, Yi-Zhe Song, and Yonggang Qi. Zero-shot everything sketch-based image retrieval, and in explainable style. In *CVPR*, 2023. 1

[17] Hangyu Lin, Yanwei Fu, Peng Lu, Shaogang Gong, Xiangyang Xue, and Yu-Gang Jiang. Tc-net for isbir: Triplet classification network for instance-level sketch based image retrieval. In *ACMM*, 2019. 2

[18] Fang Liu, Changqing Zou, Xiaoming Deng, Ran Zuo, Yu-Kun Lai, Cuixia Ma, Yong-Jin Liu, and Hongan Wang. Scenesketcher: Fine-grained image retrieval with scene sketches. In *CVPR*, 2020. 2

[19] 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, 2023. 1, 3, 6

[20] Yingge Liu, Dawei Dai, Kenan Zou, Xiaoyu Tan, Yiqiao Wu, and Guoyin Wang. Prior semantic-embedding representation learning for on-the-fly fg-sbir. *Expert Systems with Applications*, 255:124532, 2024. 1, 2, 3, 6

[21] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. In *ICLR*, 2017. 5

[22] Kaiyue Pang, Ke Li, Yongxin Yang, Honggang Zhang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. Generalising fine-grained sketch-based image retrieval. In *CVPR*, 2019. 1

[23] Kaiyue Pang, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *CVPR*, 2020. 1

[24] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. Stylemeup: Towards style-agnostic sketch-based image retrieval, 2021. 2

[25] Yuming Shen, Li Liu, Fumin Shen, and Ling Shao. Zero-shot sketch-image hashing. In *CVPR*, 2018. 1

[26] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 2017. 1, 6

[27] Jifei Song, Qian Yu, Yi-Zhe Song, Tao Xiang, and Timothy Hospedales. Deepspatial semantic attention for fine-grained sketch-based image retrieval. In *CVPR*, 2017. 2

[28] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2015. 3

[29] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch your own gan. In *ICCV*, 2021. 1

[30] Yanfei Wang, Fei Huang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. Deep cascaded cross-modal correlation learning for fine-grained sketch-based image retrieval, 2020. 2

[31] Sasi Kiran Yelamarthi, M. Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch-based image retrieval. In *ECCV*, 2018. 1

[32] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy Hospedales, and Chen Change Loy. Sketch me that shoe. pages 799–807, 2016. 2, 5, 6

[33] Zhaolong Zhang, Yuejie Zhang, Rui Feng, Tao Zhang, and Weiguo Fan. Zero-shot sketch-based image retrieval via graph convolution network. In *AAAI*, 2020. 1