

ARNet: Self-Supervised FG-SBIR with Unified Sample Feature Alignment and Multi-Scale Token Recycling

Jianan Jiang^{1,4}, Hao Tang^{2*}, Zhilin Jiang¹, Weiren Yu³, Di Wu^{1,4*}

¹Hunan University

²Peking University

³University of Warwick

⁴ExponentiAI Innovation

{jiangjn22, zhilin, dwu}@hnu.edu.cn, haotang@pku.edu.cn, weiren.yu@warwick.ac.uk

Abstract

Fine-Grained Sketch-Based Image Retrieval (FG-SBIR) aims to minimize the distance between sketches and corresponding images in the embedding space. However, scalability is hindered by the growing complexity of solutions, mainly due to the abstract nature of fine-grained sketches. In this paper, we propose an effective approach to narrow the gap between the two domains. It mainly facilitates unified mutual information sharing both intra- and inter-samples, rather than treating them as a single feature alignment problem between modalities. Specifically, our approach includes: (i) Employing dual weight-sharing networks to optimize alignment within the sketch and image domain, which also effectively mitigates model learning saturation issues. (ii) Introducing an objective optimization function based on contrastive loss to enhance the model’s ability to align features in both intra- and inter-samples. (iii) Presenting a self-supervised *Multi-Scale Token Recycling (MSTR) Module* featured by recycling discarded patch tokens in multi-scale features, further enhancing representation capability and retrieval performance. Our framework achieves excellent results on CNN- and ViT-based backbones. Extensive experiments demonstrate its superiority over existing methods. We also introduce Cloths-V1, the first professional fashion sketch-image dataset, utilized to validate our method and will be beneficial for other applications.

Code — <https://github.com/ExponentiAI/ARNet>

Introduction

Starting with (Yu et al. 2016), the triplet loss has gained prominence in the field of Fine-Grained Sketch-Based Image Retrieval (FG-SBIR). It encourages proximity among similar samples (Pos) in the embedding space while pushing different samples (Neg) farther apart, as shown in Fig. 1 (a), thereby achieving effective separation of the samples. However, the selection of suitable triplets necessitates further refinement to achieve optimal performance. Moreover, in some cases, the triplet loss might not provide sufficient gradient signals, affecting the model optimization process. To address these challenges, methods such as reinforcement learning (Muhammad et al. 2018; Bhunia et al. 2020, 2022a), meta-learning (Sain et al. 2021; Bhunia et al.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

*Corresponding Authors.

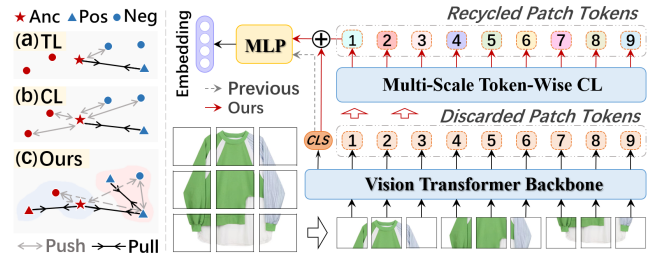


Figure 1: [Left] Comparison of our objective function with traditional methods based on Triplet Loss (TL) and Contrastive Loss (CL). [Right] Comparison of our method with previous methods for handling Patch Tokens in ViTs.

2022b), and out-of-sample learning (Sain et al. 2023) have emerged. These approaches aim to refine the model’s retrieval capabilities. However, such strategies may introduce instability during model training and pose challenges in tuning hyperparameters, such as gradient vanishing and hyperparameter sensitivity in (Sain et al. 2021; Bhunia et al. 2021). These challenges collectively compromise the overall model scalability, making it challenging for researchers to further expand their research. Furthermore, several methods (Sain et al. 2020; Yang et al. 2021; Pang et al. 2020) divide input sketches into strokes or layers, adding additional complexity. These might be less compatible with other datasets and less adaptable for broader exploration. In summary, triplet loss has achieved widespread adoption and spurred the development centered on its selection and optimization. However, this advancement has come at the cost of modeling complexity and a compromise in stability.

The introduction of Vision Transformer (ViT) (Dosovitskiy et al. 2020) has introduced a new architectural backbone and powerful performance in the field of computer vision. However, existing methods (Abhra et al. 2022; Sain et al. 2023) typically utilize the *[CLS] Token* as feature output, discarding the features of other *Patch Tokens* as shown in Fig. 1, or employing common aggregation methods such as AvgPool or MaxPool, etc. We found that there are differences between these discarded *Patch Tokens*, which may contain potential useful information. Therefore, reasonable recycling of useful features from these discarded tokens will improve the representation ability of the model.

In this study, we introduce the ARNet (Feature Alignment and Token Recycling). We experimentally discovered that a straightforward methodology can achieve favorable results and enhance the model’s versatility by switching any backbone. Additionally, recycling token features can improve the model’s feature representation capability. Specifically, we simultaneously optimize the feature distribution of intra- and inter-samples through structural design and loss function configuration. Our initial experiments revealed that single-branch encoders can tend to train saturation. Therefore, we introduced a dual weight-sharing network structure to enhance feature consistency within the sketch and image domains, ensuring effective learning of key fine-grained features. We also proposed a new objective optimization method for self-supervised training, promoting optimization of features in both intra- and inter-samples, as shown in Fig. 1 (c), which differs from traditional Triplet Loss (a) and Contrastive Loss (b). Furthermore, we introduced the plug-and-play *Multi-Scale Token Recycling (MSTR) Module* to recycle discarded *Patch Tokens* and used Multi-Scale Token-Wise Contrastive Learning to filter out similar features between tokens while retaining unique features. This approach enhances feature learning from individual samples, thereby improving the model’s representation ability.

In summary, our contributions are summarized as follows:

- We propose a novel self-supervised FG-SBIR framework straightforwardly with the dual weight-sharing networks to enhance feature alignment within the samples domain and the objective optimization function to optimize intra- and inter-sample feature distributions concurrently.
- We introduce a plug-and-play module for ViT, the MSTR, which leverages the discarded patch tokens recycling and contrastive learning among tokens with multi-scale features to achieve better feature representation and enhanced performance.
- Extensive experiments verify the effectiveness of each component and the superior performance of our method on FG-SBIR over other state-of-the-art methods.
- Additionally, we introduce a new dataset named Clothes-V1, filling the gap in professional fashion clothing datasets in this field. Its multi-level quality can be valuable for other computer vision tasks as well.

Related Work

FG-SBIR. Fine-grained sketch-based image retrieval (FG-SBIR) involves retrieving images in a specific category based on a given sketch query. The evolution of FG-SBIR, from the introduction of triplet loss (Yu et al. 2016) to advancements in model learning capabilities through attention mechanisms (Song et al. 2017) and recent approaches such as reinforcement learning (Muhammad et al. 2018; Bhunia et al. 2020, 2022a), meta-learning (Sain et al. 2021; Bhunia et al. 2022b), and sketch layered feature representation (Bhunia et al. 2020; Sain et al. 2020; Yang et al. 2021), has been rapid. However, these approaches are based on triplet loss and improve model performance using intricate methodologies. In contrast, our approach takes a different perspective by designing a simple and universal frame-

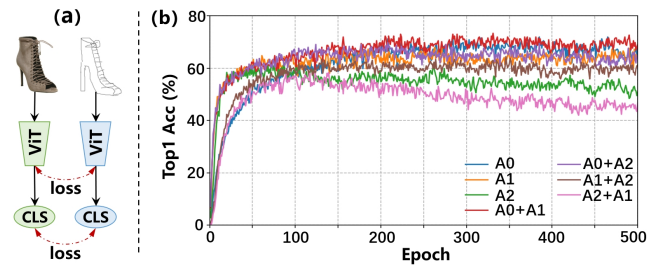


Figure 2: (a) A initial network for FG-SBIR. (b) Visualization of the training process with different augmentations.

work. This is suitable for any backbone, yielding improved retrieval results without a complex methodology design.

Contrastive Learning. Contrastive learning improves model representational ability by maximizing positive sample similarity and minimizing negative sample similarity. In earlier works, Wu et al. (2018) proposed the memory bank to store data features. Since then, methods like (Ye et al. 2019; Tian, Krishnan, and Isola 2020; Chen et al. 2020; Duan et al. 2024) used data augmentation, learned from multiple views as input or larger pos-neg sample pairs, etc. He et al. (2020) proposed the momentum encoder to address sample selection bias. Caron et al. (2020) used a clustering algorithm to reduce computational costs. Recently, Wang et al. (2021b); Tang et al. (2021, 2023) proposed pixel contrast for image segmentation or synthesis. These methods drive unsupervised representation learning to achieve superior performance. Similarly, we can leverage the concept of contrastive learning to FG-SBIR for better retrieval performance.

Vision Transformer. Vision Transformer (ViT) (Dosovitskiy et al. 2020) uses a self-attention mechanism like (Vaswani et al. 2017), leading to a more effective way to model global features compared to traditional CNNs. Based on this, methods such as (Han et al. 2021; Thiry et al. 2024; Liu et al. 2021; Wang et al. 2021a) proposed smaller patches, the shift-window mechanism, progressive shrinking pyramid, etc. to further enhance the capability of ViTs. However, only a few studies (Abhra et al. 2022; Sain et al. 2023) have explored ViT in the field of FG-SBIR. Moreover, these methods choose the *[CLS] Token* as feature output and discard other *Patch Tokens*, which will also result in discarding some potentially useful features. Therefore, we introduce the ViT into FG-SBIR and further improve model performance by recycling the discarded features.

Methodology

How to Keep Learning?

We begin with an initial structure as shown in Fig. 2 (a), using the prevalent pre-trained ViT-B/16 (Dosovitskiy et al. 2020) as our backbone and modifying the dimensions of the classification layer. However, we encountered the first challenge: the model’s performance exhibited saturation around the 150-epoch mark, and subsequent training iterations yielded negligible enhancements, as illustrated by

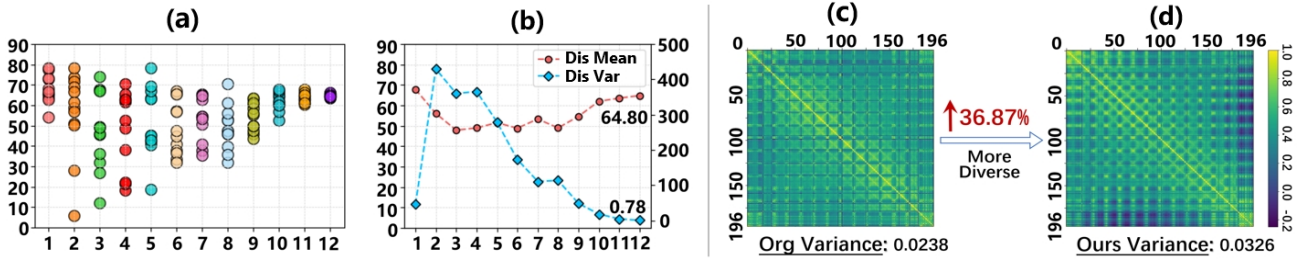


Figure 3: (a)(b) The average attention distance of ViT backbone in multi-heads (dots) w.r.t each layer. (c)(d) The comparison of similarity and variance between discarded 196 *Patch Tokens* in the output of the ViT backbone.

the A0 curve in Fig. 2 (b) *. We tried experimenting with conventional solutions like data augmentation, output feature normalization, etc. Unfortunately, these methods did not show improvements. In terms of data augmentation, methods like A1 and A2 appeared to have no benefit in the generalization capability or even drive the model into overfitting territory, as shown in Fig. 2 (b) by the A1 and A2 curves.

Thus, the question of whether the model’s ability to learn better feature representations from input samples was impeded by its single-network architecture. Inspired by (Bromley et al. 1993), we surpass this limitation by introducing the dual weight-sharing networks. This change in approach enabled the model to optimize the feature distribution of both intra- and inter-samples, yielding more representative and discriminative features. Furthermore, through an alternative preprocessing approach for inputs, we have a noteworthy observation: the inclusion of A0 is beneficial to the weight-sharing network and consistently yields better outcomes. Intriguingly, the elaborate data augmentation strategy A2 is incompatible with the nuances of fine-grained datasets, dampening the model’s generalization abilities. The A0+A1 curve in Fig. 2 (b) demonstrates the model’s enduring generalization capacity within the 500 epochs with the absence of pronounced feature degradation.

Are Discarded Patch Tokens Useless?

In previous studies (Dosovitskiy et al. 2020), the features at the *[CLS] Token* were chosen as the features extracted by ViT, and the features at other *Patch Tokens* were discarded. So are these discarded features really useless?

The Multi-Head Self-Attention module is the core component of the ViT, aiding in the analysis of information and attention distribution learned by the model. For the n -th head, the attention matrix can be expressed as:

$$\text{Att}^n = \text{softmax} \left(\mathbf{Q}^n (\mathbf{K}^n)^T / \sqrt{d_{\mathbf{K}^n}} \right) \cdot \mathbf{V}^n, \quad (1)$$

where $n \in \{1, \dots, N\}$, N is the number of multi-heads with each head computing the attention independently, followed by concatenation and projection to form the final output. \mathbf{Q} (query), \mathbf{K} (key), and \mathbf{V} (value) are linear projections of the input feature, obtained by applying three learnable weight matrices W^Q , W^K , and W^V , respectively.

*A0: no augmentation. A1: image scaling, rotation, and cropping operations. A2: similar to Moco (He et al. 2020).

As expressed in Eq. (2), we calculate the mean **Mean** and variance **Var** of the average attention distance of each head. Here, Att_{ij} is the attention weight of the i -th token to the j -th token, $|i - j|$ is the distance between the i -th token and the j -th token, and μ_i is the average attention distance of the i -th token. M is the number of patch tokens.

$$\mu_i = \sum_{j=1}^M \text{Att}_{ij}^n |i - j|, \quad \text{Mean}^n = \frac{1}{M} \sum_{i=1}^M \mu_i, \quad (2)$$

$$\text{Var}^n = \frac{1}{M} \sum_{i=1}^M \sum_{j=1}^M (\text{Att}_{ij}^n |i - j| - \mu_i)^2.$$

As illustrated in Fig. 3, we compute the average attention distance of multi-heads attention for each layer within the ViT backbone, along with the *Mean* and *Variance* of attention distance for each layer. A lower *Mean* indicates a model preference for local features, while a lower *Variance* indicates more consistent feature representations. We observe that the model’s attention map tends to be consistent as the number of model layers increases (Fig. 3 (a)), but it is not completely consistent (Fig. 3 (b)). There are always some differences between the *Patch Tokens* in each head (Fig. 3 (c)), and these differences likely contain useful semantic information. By recycling these discarded *Patch Tokens*, we can extract the unique features contained in each token and integrate them into the *[CLS] Token*, achieving better feature representation. Fig. 3 (c) and (d) present the feature similarity graphs with and without the *Patch Tokens* recycling, respectively. Our method increases feature differences by 36.87%, thereby ensuring that the unique features of each token are retained and utilized effectively.

Feature Alignment and Token Recycling

We present the framework of our method in Fig. 4, which consists of three main components: (i) The *Basic Model* that achieves high retrieval performance with a straightforward structure by feature alignment. (ii) The *Multi-Scale Token Recycling (MSTR)* Module designed to recycle potentially valuable features from discarded *Patch Tokens*. (iii) The *Loss Function* utilized for optimizing model training.

Basic Model. We begin by randomly loading sketch-image sample pairs from the dataset and performing data augmentation. The augmented and original samples are then

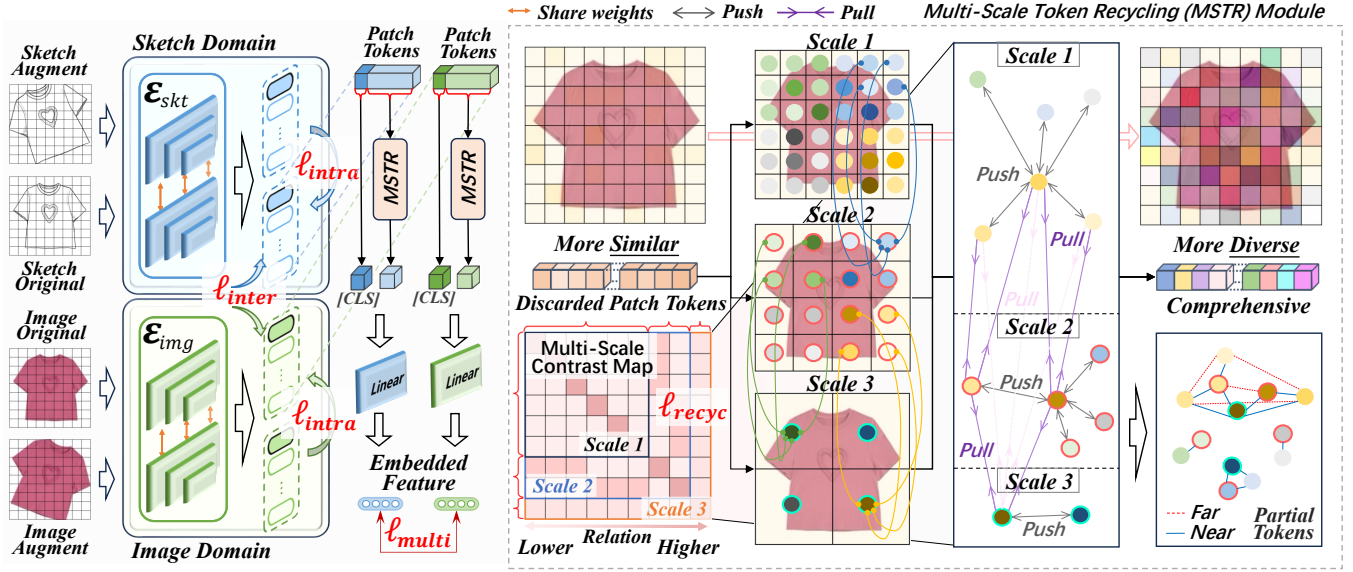


Figure 4: An overview of our method, which enhanced feature alignment from both intra- and inter-sample perspectives. We propose the *MSTR Module* to further improve the Encoder’s feature representation ability by recycling discarded *Patch Tokens*.

input into the weight-sharing network \mathcal{E} as previously described. Finally, we utilize a linear layer to map the *[CLS] Token* of sketch and image features to a joint embedding space. Specifically, denoting the sketches and images within our dataset as \mathcal{S} and \mathcal{I} respectively, with the number of samples in each minibatch designated as m , we succinctly represent our data-loading process as $\mathcal{B}_m \Rightarrow \{s_j \in \mathcal{S}, i_j \in \mathcal{I} \mid s_j \text{ and } i_j \in \mathcal{R}^{224 \times 224 \times 3}\}$, and $j = 1, 2, \dots, m$. Initiating the feature extraction process, we begin by applying data augmentation to the input samples \mathcal{B}_m , generating the sample pairs denoted as $\mathcal{B}'_m \Rightarrow \{(s_j, s'_j, i_j, i'_j) \mid s_j \in \mathcal{S}, i_j \in \mathcal{I}, s'_j \in \mathcal{S} \text{ and } i'_j \in \mathcal{I} \text{ are augmented samples}\}$. Subsequently, the augmented samples undergo patch encoding and positional encoding before being fed to the encoders \mathcal{E}_{skt} and \mathcal{E}_{img} , obtaining the sketch features $\mathcal{F}_{skt} \Rightarrow \{f_{skt} \leftarrow \mathcal{E}_{skt}(s), f'_{skt} \leftarrow \mathcal{E}_{skt}(s')\}$ and image features $\mathcal{F}_{img} \Rightarrow \{f_{img} \leftarrow \mathcal{E}_{img}(i), f'_{img} \leftarrow \mathcal{E}_{img}(i')\}$ for the input samples \mathcal{B}_m . Finally, the resulting features \mathcal{F}_{skt} and \mathcal{F}_{img} are channeled through a linear layer, mapping them into the $\mathcal{R}^{1 \times 512}$ dimension within a joint embedding space. This process yields the embedded sketch features \mathcal{F}_{skt}^{out} and image features \mathcal{F}_{img}^{out} for retrieval in a straightforward way.

MSTR Module. Based on the previous analysis and the *Basic Model*, we further propose the *Multi-Scale Token Recycling (MSTR) Module*, which can achieve a better feature representation of samples by recycling potentially useful features from discarded *Patch Tokens*. Specifically, for the feature \mathcal{F} obtained by the ViT backbone, which contains the features of retained *[CLS] Token* \mathcal{F}_{CLS} and discarded *Patch Tokens* \mathcal{F}_{PT} . We use the *MSTR Module* to extract potentially useful features from the \mathcal{F}_{PT} , then output and add them to the \mathcal{F}_{CLS} . Other operations remain consistent with the *Basic Model*, making *MSTR* a plug-and-play module.

For the input feature \mathcal{F}_{PT} , we recognize that after the

self-attention mechanism of ViT, the features among these tokens become *More Similar*. To uncover potentially useful information, we leverage the fact that each token originates from a different position in the image, thus containing unique semantic information. Our goal is to make the final \mathcal{F}_{PT} *More Diverse*, ensuring each token retains unique and useful features, thereby improving the overall feature representation. We first reduce the dimension of \mathcal{F}_{PT} to obtain three scale features, providing a richer multi-level feature representation \mathcal{F}_{ML} . We then use two Transformer layers (Vaswani et al. 2017) to sequentially encode the \mathcal{F}_{ML} . Finally, the resulting embedded feature is dimensionally reduced and added with \mathcal{F}_{CLS} .

During the model training process, we reduce the dimension of \mathcal{F}_{ML} to filter out and concatenate the key feature \mathcal{F}_{ML}^{map} . Furthermore, we constructed a *Multi-Scale Contrast Map (MAP)* to guide the learning of the *MSTR Module*. For tokens within the same scale, we maximize the distance between them to highlight their differences. For tokens between different scales, we maximize the distance between patches at different positions and minimize the distance between patches at the same or adjacent positions, ensuring that we capture unique features while not ignoring the inherent similarities among tokens:

$$\ell_{recyc} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \left((\mathcal{F}_{ML}^{map} \cdot \mathcal{F}_{ML}^{mapT})_{ij} - \text{MAP}_{ij} \right)^2. \quad (3)$$

Loss Function

Inspired by (Chen et al. 2020), we introduce contrastive loss as our basic objective function:

$$\mathcal{L}(\mathcal{F}) = -\frac{1}{N} \sum_{i=1}^N \log \frac{\sum_{z=1}^{\frac{N}{L}-1} \exp((\mathcal{F}_i \cdot \mathcal{F}_{N-z+i}^T)/\tau)}{\sum_{j=1}^L \mathbb{I}_{[i \neq j]} \exp((\mathcal{F}_i \cdot \mathcal{F}_j^T)/\tau)}, \quad (4)$$



Figure 5: Comparison of our dataset with other datasets.

where \mathcal{F} denotes an input feature vector derived from n sample vectors and concatenate Cat along the batch size dimension, its construction follows $\mathcal{F} = \text{Cat}(\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n)$. N is the batch size, L is the length of the concatenated feature vector, and τ is a temperature parameter. The indicator function $\mathbb{I}[i \neq j] \in \{0, 1\}$ equals 1 when $i \neq j$. (\cdot) is the dot product to obtain the similarity matrix between two vectors.

Basic Model Optimize. The complete loss function of our Basic Model (ARNet Basic) comprised of: (i) *Multi-Modal Loss* ℓ_{multi} minimizes the distance between the sketch and its corresponding image within the joint embedding space. (ii) *Inter-Sample Loss* ℓ_{inter} encourages the similarity between the sketch features and image features output by their Encoders \mathcal{E} , which can further optimize their ultimate distribution within their joint embedding space. (iii) *Intra-Sample Loss* ℓ_{intra} learns the distinctive features inherent to the original images and their augmented counterparts in a self-supervised way, enhancing the capability of feature representation. In general, it is represented as follows:

$$\mathcal{L}_B = \underbrace{\mathcal{L}(\mathcal{F}_{skt}^{out}, \mathcal{F}_{img}^{out})}_{\ell_{multi}} + \underbrace{\mathcal{L}(f_{skt}, f_{img})}_{\ell_{inter}} + \underbrace{\mathcal{L}(f_{skt}, f'_{skt}) + \mathcal{L}(f_{img}, f'_{img})}_{\ell_{intra}}. \quad (5)$$

Full Model Optimize. Based on the Basic Model, we incorporate the MSTR Module to achieve better performance. As a result, the loss function of our full model (ARNet Full) can be represented as:

$$\mathcal{L}_+ = \mathcal{L}_B + \ell_{recyc}. \quad (6)$$

In this way, our approach can facilitate the model to simultaneously focus on information both intra- and inter-samples with multi-perspectives, while incorporating recycled potentially valuable features.

Experiments and Results

Experimental Settings

Clothes-V1. We invited four professional designers (over five years) and three junior designers (over one year or design novice) to complete its creation. This diverse team composition ensures a wide range of sketch styles and levels of

Method	Clothes-V1(%)		
	Acc.@1	Acc.@5	Acc.@10
Triplet-SN	64.36	85.82	92.73
Triplet-Att-SN	70.18	83.64	91.64
B-Siamese	84.73	97.82	99.27
OnTheFly	63.27	90.18	92.73
EUPS-SBIR	89.47	94.73	97.36
ARNet Basic	94.12	98.91	99.27
ARNet Full	95.27	98.55	99.27

Table 1: Comparative results of our model against other methods on Clothes-V1.

expertise, making the dataset more representative of real-life scenarios. *Its specialized nature and diverse collection of fashion sketches* make it a valuable resource for various tasks, including FG-SBIR, image generation, image translation, and fashion-related research.

Specifically, Clothes-V1 has 1200 (500) sketches (images), containing 925 (380) and 275 (120) for the training and validation set. Each image corresponds to at least one sketch and at most five sketches. The dataset is further categorized into types such as sweaters, shirts, jackets, dresses, cheongsams, etc. Please refer to our GitHub for more details.

Datasets. We utilize two widely used FG-SBIR datasets *QMUL-Chair-V2* and *QMUL-Shoe-V2* (Yu et al. 2016) along with our proposed self-collected dataset Clothes-V1 to evaluate the performance of our proposed framework. The *QMUL-Chair-V2* includes 964 (300) and 311 (100) sketches (images) for the training set and validation set, and the *QMUL-Shoe-V2* includes 5982 (1800) and 666 (200) sketches (images) for the training set and validation set.

Implementation Details. Our model is implemented using the PyTorch framework and is based on the ViT-B/16-1K model (Dosovitskiy et al. 2020). The input size is 224×224 , the final embedding vector’s dimension is 512, and the temperature parameter τ is 0.07. We train the model on a single NVIDIA 32GB Tesla V100 GPU, using a batch size of 16 and the Adam optimizer (Kingma and Ba 2014) with a learning rate of $6e-6$ and weight decay of $1e-4$, and the training process lasts for 500 epochs. In addition, we incorporated other pre-trained models VGG16 (Simonyan and Zisserman 2014), ResNet50 (He et al. 2016), InceptionV3 (Szegedy et al. 2016), and Swin-ViT (Liu et al. 2021) to further verify the adaptability and efficacy of our framework.

Evaluation Metrics. We utilize the widely used R@1, R@5, and R@10 to quantify the retrieval performance, which represents the probability of the correct image appearing in the first 1, 5, and 10 retrieved results, respectively.

Competitors. We compare our approach with previous 12 state-of-the-art methods: *Triplet-SN* (Yu et al. 2016), *Triplet-Att-SN* (Song et al. 2017), and *B-Siamese* (Sain et al. 2020) are basic methods. *CMHM-SBIR* (Sain et al. 2020) proposes a multi-level structure. *OnTheFly* (Bhunia et al. 2020) employs multi-stage sketch input. *SketchAA* (Yang et al. 2021)

Method	QMUL-Chair-V2			QMUL-Shoe-V2		
	R@1	R@5	R@10	R@1	R@5	R@10
Triplet-SN	33.75	65.94	79.26	18.62	43.09	59.31
Triplet-Att-SN	37.15	67.80	82.97	22.67	51.20	65.02
B-Siamese	40.56	71.83	85.76	20.12	48.95	63.81
CMHM-SBIR	51.70	80.50	88.85	29.28	59.76	74.62
OnTheFly	39.01	75.85	87.00	35.91	66.78	78.54
SketchAA	52.89	-	94.88	32.22	-	79.63
Semi-Sup	60.20	78.10	90.81	39.10	69.90	87.50
StyleMeUp	62.86	79.60	91.14	36.47	68.10	81.83
Adpt-SBIR	-	-	-	38.30	76.60	-
Part-SBIR	63.30	79.70	-	39.90	68.20	82.90
NT-SBIR	64.80	79.10	-	43.70	74.90	-
EUPS-SBIR	71.22	80.10	92.18	44.18	70.80	84.68
ARNet Basic	73.31	93.24	97.15	40.11	67.54	79.29
ARNet Full	75.45	94.66	97.87	42.91	72.95	81.72

Table 2: Comparative results of our model against other methods on QMUL-Chair-V2 and QMUL-Shoe-V2.

proposes a multi-granularity modeling. *Semi-Sup* (Bhunia et al. 2021) introduces a semi-supervised framework. *Style-MeUp* (Sain et al. 2021) proposes a meta-learning framework. *Adpt-SBIR* (Bhunia et al. 2022b) introduces a model-agnostic framework. *Part-SBIR* (Chowdhury et al. 2022) proposes cross-modal associations. *NT-SBIR* (Bhunia et al. 2022a) uses sketch stroke selection. *EUPS-SBIR* (Sain et al. 2023) employs model aided learning.

Performance Analysis

FG-SBIR. As presented in Tables 1 and 2, our proposed method exhibits superior performance compared to other baselines on three datasets. **(i) QMUL-Chair-V2:** Our model achieves new SOTA results and its Top5 accuracy can surpass EUPS-SBIR’s Top10 accuracy. For Top1 accuracy, it outperforms nearly 10% compared to NT-SBIR and Part-SBIR. **(ii) QMUL-Shoe-V2:** Although our model does not secure the top ranking, it still demonstrates encouraging potential with third position, boasting a slight gap in three metrics compared to the existing methods. This can be attributed to the dataset’s larger size and the more relatively irregular sketch strokes compared to the chair and clothes, leading to more complex mappings and being closely clustered within the joint embedding space, which may affect the optimization process. **(iii) Clothes-V1:** The results present a clear advantage of our model in Top1 and Top5 accuracy, along with an impressive nearly 100% accuracy in Top10. **(iv) ARNet Full:** Based on the ARNet Basic, our model achieved excellent results on three datasets with steady improvement by our proposed *MSTR Module*. These experimental results demonstrate our model’s remarkable retrieval capabilities in a straightforward way, as well as the effectiveness of our approach for recycling discarded Patch Tokens.

Retrieval Performance. We present the comparative retrieval performance in Fig. 6. The corresponding images retrieved by our method achieve better ranking results in



Figure 6: Retrieval results between our method (blue) and Triplet-Att-SN (green) on QMUL-Chair-V2.

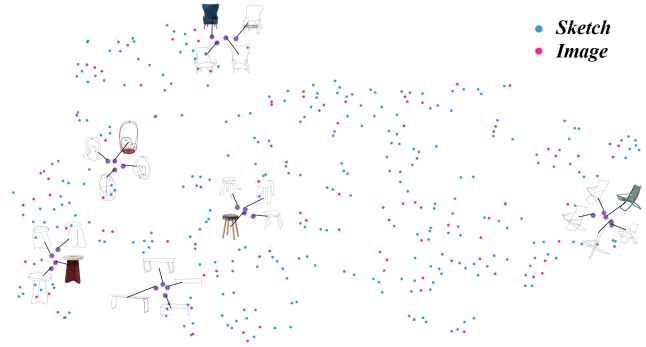


Figure 7: t-SNE visualizations of sketches (blue) and images (pink) feature embeddings on QMUL-Chair-V2.

Acc.@5 (top). The top-retrieved image in the Acc.@10 also exhibits a significant similarity to the provided sketch (bottom). Furthermore, in the second row, despite the ground truth being in second place, the first one resembles the provided sketch from a human visual perspective. These results highlight our stronger retrieval performance.

Feature Embedding. We utilize the t-SNE (Van der Maaten and Hinton 2008) to visualize the distribution of encoded sketches and images in the joint embedding space. As depicted in Fig. 7, it illustrates a relatively even distribution across two modalities and a close distribution between images and their corresponding sketches. This indicates the success of our method in extracting and encoding key information, effectively optimizing sample feature distribution and yielding better retrieval results.

Ablation Study

Framework. Table 3 represents that our dual weight-sharing networks can significantly enhance the overall performance of the model across various mainstream backbone networks. In essence, our framework takes into account both the single feature distribution and the joint feature distribu-

	Method	Backbone	ℓ_{inter}	Acc.@1	Acc.@5	Acc.@10
Framework	Sing	VGG16	×	22.42	56.23	74.38
	Sing	InceptionV3	×	39.86	78.29	90.04
	Sing	ResNet50	×	36.66	74.73	86.48
	Sing	ViT-B/16	×	59.43	89.32	95.02
	Sing	Swin-ViT	×	62.63	93.24	97.51
	Doub	VGG16	×	31.32	64.77	81.14
	Doub	InceptionV3	×	42.71	80.43	90.39
	Doub	ResNet50	×	42.35	72.95	85.41
	Doub	ViT-B/16	×	58.36	90.75	95.37
	Doub	Swin-ViT	×	63.35	90.75	95.37
	Doub	VGG16	✓	46.26	79.00	86.83
	Doub	InceptionV3	✓	46.26	82.56	91.82
	Doub	ResNet50	✓	48.04	81.14	89.32
	Doub	ViT-B/16	✓	73.31	93.24	97.15
	Doub	Swin-ViT	✓	73.31	93.95	98.22

	Method	Augment	Loss	Acc.@1	Acc.@5	Acc.@10
Dataloader	Sing	A0	C	59.43	89.32	95.02
	Sing	A1	C	68.33	90.39	95.73
	Sing	A2	C	61.92	91.04	96.44
	Doub	A0+A1	C	73.31	93.24	97.15
	Doub	A0+A2	C	69.75	92.17	95.37
	Doub	A1+A2	C	64.41	87.19	93.24
	Doub	A2+A1	C	58.36	85.41	91.82
	Trp-S	A0	T	57.65	89.83	96.09
	Trp-S	A1	T	55.52	87.90	95.37
	Trp-S	A2	T	51.60	81.85	90.75
	Trp-I	A0	T	24.20	51.96	68.68
	Trp-I	A1	T	16.01	41.64	55.87
	Trp-I	A2	T	17.44	41.28	58.36

Table 3: Ablation Studies on our framework and different ways to load datasets on QMUL-Chair-V2. ((i) Doub: Utilized our dual weight-sharing networks. (ii) Sing: no utilization. (iii) Trp-S/I: Use the sketch/image sample as Triplet-loss anchor. (iv) C/T: Contrastive loss/Triplet loss.)

tion of sketches and images, resulting in better feature representations. Furthermore, we observed that the ℓ_{inter} can enhance the overall model’s performance and accelerate the convergence trajectory. This can be attributed to its ability to further enhance the feature alignment between sketches and images, contributing to the extraction of more distinctive and representative features by dual weight-sharing networks.

Dataloader. We compare three different data processing strategies using two data loading methods based on the Contrast loss C and the Triplet loss T in Table 3. The results demonstrated that the C-based method consistently outperforms the T-based method under the same conditions. This superiority is attributed to our loss function’s comprehensive in-sample and inter-sample optimization, which provides greater gradient contrast information in each mini-batch optimization, thereby enhancing overall performance.

	Rcyc	Map	LayT	LayS	Dim	Acc.@1	Acc.@5	Acc.@10
×	×	×	×	×	73.31	93.24	97.15	
max	×	×	×	×	72.95	94.66	98.22	
mean	×	×	×	×	71.89	93.59	97.15	
cat	×	×	×	×	65.48	92.53	97.51	
✓	×	1	3	16	73.67	94.31	97.51	
✓	✓	1	3	16	74.38	94.66	97.15	
✓	✓	3	3	16	71.89	93.24	96.44	
✓	✓	2	3	16	75.45	94.66	97.87	
✓	✓	2	4	16	72.95	92.53	97.87	
✓	✓	2	2	16	73.67	95.02	95.02	
✓	✓	2	3	8	72.60	93.59	97.15	
✓	✓	2	3	24	73.31	94.31	97.51	
✓	✓	2	3	48	72.60	93.59	97.15	

Table 4: Ablation Studies on our MSTR Module on QMUL-Chair-V2. (max, mean, and cat: Use MaxPool, AvgPool, and Concatenation for Discarded Patch Tokens.)

MSTR Module. As shown in Table 4, we conducted ablation experiments on the MSTR module from four perspectives: recycling discarded Patch Tokens (Rcyc), constructing MAP (Map), the number of Transformer Layers (LayT), the number of feature Scale Layers (LayS), and the feature dimension used for Map (Dim). The results indicate that simple methods like MaxPool, AvgPool, and Concatenation not only fail to improve but may even negatively impact model performance. Using the map can enhance contrastive learning between tokens, thus improving model performance. When examining LayT, 2 layers outperformed both 1 and 3 layers. This is mainly because 1 layer fails to encode feature information adequately, while 3 layers result in over-encoding and negative effects. For validating LayS, we evaluated models with 2 and 4 layers compared to our 3-layer setting. The results demonstrate that using 3 layers improves Top1 accuracy by approximately 2% and 3% compared to the settings of 2 and 4 layers, while also achieving balanced performance in Top5 and Top10 accuracy. Regarding Dim, we observed a similar phenomenon that 8 cannot encapsulate all unique information, while 24 and 48 include too much irrelevant information, weakening the feature representation of discarded *Patch Tokens*. These findings collectively demonstrate the effectiveness of our MSTR module.

Conclusion

In this paper, we present ARNet, a novel FG-SBIR solution, along with a general and diverse clothing dataset, Cloth-V1. Our approach incorporates structure and loss function design, enabling the model to concurrently focus on intra- and inter-sample feature alignment. Our proposed method exhibits strong scalability, accommodating various mainstream backbones as feature extraction networks, and offers straightforward implementation. Moreover, we introduce the MSTR module to further improve our basic model representation ability by recycling discarded Patch Tokens. Extensive experiments demonstrate that our ARNet achieves superior retrieval performance across several benchmarks.

References

- Abhra et al. 2022. Cross-Modal Fusion Distillation for Fine-Grained Sketch-Based Image Retrieval. In *BMVC*.
- Bhunia, A. K.; Chowdhury, P. N.; Sain, A.; Yang, Y.; Xiang, T.; and Song, Y.-Z. 2021. 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*, 4247–4256.
- Bhunia, A. K.; Koley, S.; Khilji, A. F. U. R.; Sain, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2022a. Sketching without worrying: Noise-tolerant sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 999–1008.
- Bhunia, A. K.; Sain, A.; Shah, P. H.; Gupta, A.; Chowdhury, P. N.; Xiang, T.; and Song, Y.-Z. 2022b. Adaptive fine-grained sketch-based image retrieval. In *European Conference on Computer Vision*, 163–181. Springer.
- 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.
- Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature verification using a “siamese” time delay neural network. *Advances in neural information processing systems*, 6.
- Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.
- Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.
- Chowdhury, P. N.; Bhunia, A. K.; Gajjala, V. R.; Sain, A.; Xiang, T.; and Song, Y.-Z. 2022. Partially does it: Towards scene-level fg-sbir with partial input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2395–2405.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Duan, B.; Tang, H.; Sun, C.; Zhu, Y.; and Yan, Y. 2024. Mining and Unifying Heterogeneous Contrastive Relations for Weakly-Supervised Actor-Action Segmentation. In *WACV*.
- Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; and Wang, Y. 2021. Transformer in transformer. In *NeurIPS*.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- 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.
- Pang, K.; Yang, Y.; Hospedales, T. M.; Xiang, T.; and Song, Y.-Z. 2020. Solving mixed-modal jigsaw puzzle for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10347–10355.
- Sain, A.; Bhunia, A. K.; Koley, S.; Chowdhury, P. N.; Chattopadhyay, S.; Xiang, T.; and Song, Y.-Z. 2023. Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6873–6883.
- 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*.
- 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.
- Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Song, J.; Yu, Q.; Song, Y.-Z.; Xiang, T.; and Hospedales, T. M. 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE international conference on computer vision*, 5551–5560.
- 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.
- Tang, H.; Qi, X.; Sun, G.; Xu, D.; Sebe, N.; Timofte, R.; and Van Gool, L. 2021. Edge guided gans with contrastive learning for semantic image synthesis. In *ICLR*.
- Tang, H.; Sun, G.; Sebe, N.; and Van Gool, L. 2023. Edge guided GANs with multi-scale contrastive learning for semantic image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Thiry, G.; Tang, H.; Timofte, R.; and Van Gool, L. 2024. Towards Online Real-Time Memory-based Video Inpainting Transformers. In *CVPRW*.
- Tian, Y.; Krishnan, D.; and Isola, P. 2020. Contrastive multiview coding. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 776–794. Springer.
- Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021a. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 568–578.
- Wang, W.; Zhou, T.; Yu, F.; Dai, J.; Konukoglu, E.; and Van Gool, L. 2021b. Exploring cross-image pixel contrast for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7303–7313.
- Wu, Z.; Xiong, Y.; Stella, X. Y.; and Lin, D. 2018. Un-supervised Feature Learning via Non-Parametric Instance Discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Yang, L.; Pang, K.; Zhang, H.; and Song, Y.-Z. 2021. Sketchaa: Abstract representation for abstract sketches. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10097–10106.
- Ye, M.; Zhang, X.; Yuen, P. C.; and Chang, S.-F. 2019. Un-supervised Embedding Learning via Invariant and Spreading Instance Feature. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.
- 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.