Local Feature Patch Matching Enhanced Re-ranking Model for Clothing Image Retrieval

Simeng Cheng[1], Zhuangye Luo[2], Feng Zeng[3,*], Xiaowei Xie[4]

{234711094@csu.edu.cn, luozhuangye@csu.edu.cn, fengzeng@csu.edu.cn, showay@csu.edu.cn}

[1] School of Computer Science and Engineering, Central South University, Changsha, 410017, China

[2] School of Computer Science and Engineering, Central South University, Changsha, 410017, China

[3] School of Computer Science and Engineering, Central South University, Changsha, 410017, China

[4] School of Computer Science and Engineering, Central South University, Changsha, 410017, China

* Corresponding author

Abstract. Effective clothing image retrieval depends on robust global features and highly discriminative local features. We propose an effective re-ranking model to handle the issues of fine-grained difference discrimination and occlusion interference in clothing image retrieval. The global features are initially extracted by the Swin Transformer for a global-level retrieval, resulting in an initial set of candidate images. Then, a re-ranking network composed of multiple BiAttention layers is introduced to mine more discriminative local features from query–candidate image pairs. Each BiAttention module captures intra-image semantic information through self-attention, enhancing the representation capability for fine-grained differences. It also captures inter-image correlation by using bidirectional cross-attention to enhance the model's robustness to occlusion. In addition, the re-ranking network incorporates spatial relative attention to reinforce spatial position constraints. Positional importance weighting is further applied to matched feature blocks, so that the final similarity evaluation between image pairs appropriately concentrates on important feature regions. A large-scale clothing image dataset is built to validate the model's performance. Experimental results show that the proposed method is efficient in large-scale clothing image retrieval.

Keywords: image retrieval, re-ranking model, self-attention, cross-attention, Swin Transformer.

1 Introduction

There are two retrieval methods in e-commerce platforms: text-based and content-based image retrieval (CBIR). Text retrieval depends on the keywords input by users and manual annotations. The annotations often have ambiguity and are inconsistent, affecting the accuracy and expansion of the retrieval. CBIR allows users to upload pictures, then extracts visual features such as color, texture, shape, etc. The system searches for similar products from the image library according to

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365290

these features. This is a mainstream method in clothing image retrieval, especially in fine-grained tasks. CBIR is divided into two categories: category-based (CIR), which is to obtain images of the same class, and instance-based (IIR), which focuses on retrieving specific objects that are identical.

We focus on IIR in the context of clothing image retrieval. This task needs to deal with the challenges of fine-grained recognition. Specifically, clothing items have subtle inter-class differences and large intra-class variations due to factors such as perspective, scale, occlusion, deformation, background, and lighting. These challenges require feature representations to be both robust and highly discriminative.

We propose a re-ranking model that boosts retrieval accuracy via fine-grained local matching. The pipeline first retrieves top-k candidates with global features, and then reorders them by the more fine-grained similarity scores that are calculated through a local patch matching mechanism. The main contributions of this paper are as follows:

- We utilize the Swin Transformer to extract robust global features, thereby enhancing the accuracy of the initial retrieval results.
- We introduce a BiAttention module that simultaneously captures dependencies within an image and local correspondences between image pairs. This enhances the model's sensitivity to fine-grained differences in clothing images. It also reduces inter-class mismatches and improves the retrieval of local features in occluded images.
- To optimize local similarity calculations, we propose a spatial relative attention mechanism based on relative position encoding. Additionally, we propose a weighting strategy based on Gaussian kernels. These components align matched patches and emphasize discriminative regions during the reordering process, respectively.

2 Related work

Global image representation learning is applied in image retrieval. The core is to use CNNs to extract global semantic embeddings from images. For example, VGG captures multi-scale texture features by stacking convolutional layers [1]. ResNet enhances the performance of deep features with residual connections [2]. The method usually optimizes the embedding space by deep metric learning to cluster similar samples. The loss functions of representation include contrastive loss [3], triplet loss [4], and multi-similarity weighted loss [5]. However, global embeddings are easily affected by the background and posture changes, which may reduce the robustness of clothing image retrieval.

Recently, some research has introduced resort strategies to optimize the results of initial retrieval, including query expansion, geometric verification, context-based rearrangement, and difference-based methods. Query expansion is to enhance the expression of the original query by using top-ranked images. The model proposed by LAttQE[6] relies on aggregating images to enhance the expression of the query. SuperGlobal [7] employs GeM pooling to refine the global descriptor of the query and its top k neighbors. Geometric verification uses the spatial and geometric

context of local features to eliminate false matches. Some methods [8, 9, 10] utilize RANSAC-based strategies to estimate geometric transformation models for verifying local correspondences. CVNet [11] constructs cross-scale feature correlations and compresses them into image similarities for re-ranking. Contextual information among the top k nearest neighbors is leveraged to improve retrieval precision. CRL [12] represents the fixed-length retrieval list as a correlation matrix and employs a lightweight CNN to jointly learn contextual relations and pairwise relevance among the images. INRNet [13] introduces a bidirectional feature extractor and constructs a neighborhood data construction mechanism (NDCM) to reorganize data for re-ranking. Diffusion-based methods propagate similarity scores over a graph structure to better capture the intrinsic manifold geometry of the data and the intrinsic relationships among samples. CAS [14] applies similarity diffusion within local clusters and incorporates a reverse constraint term into the bidirectional diffusion objective. However, these re-ranking methods suffer from the lack of explicit local alignment mechanisms and limited capacity to model fine-grained visual similarity. Prior reranking methods refine rank via diverse optimizations. We likewise rerank the global top-k, but learn pair similarity by aligning local patches to optimize clothing-image matching.

3 Model

To address the challenges of fine-grained feature discrimination and severe occlusion in clothing image retrieval, we propose a re-ranking model based on local feature block matching. The model fully exploits local features within clothing images and integrates spatial relative attention with position-aware weighting strategies to significantly improve retrieval accuracy, as illustrated in Fig. 1.

3.1 Global Retrieval Module

As illustrated in Fig. 1, the global retrieval module takes a single image $I \in \mathbf{R}^{C\times H\times W}$ as input and adopts the Swin Transformer [15] as the backbone to extract global features $f_g \in \mathbf{R}^{C_g}$. The cosine similarity between the query and candidate images is used to rank the initial retrieval results.

The global retrieval module is trained with contrastive loss, which maximizes the cosine similarity between positive image pairs while minimizing it for negative pairs. The contrastive loss is defined as:

$$\mathscr{L}_c = \frac{1}{N}\sum_i\sum_j\Big[\sum_{y_i=y_j}[m_p - s_p]_+ + \sum_{y_i\neq y_j}[s_n - m_n]_+\Big], \tag{1}$$

where N is the batch size, y denotes sample categories, s_p and s_n represent cosine similarity of positive and negative pairs. m_p and m_n are the margins for positive and negative pairs.

3.2 Re-ranking Network

As shown in Fig. 1, the re-ranking network takes a query-candidate image pair (I^q, I^k) as input to deeply explore pairwise similarity. We utilize the feature maps $(f_2^q, f_2^k) \in \mathbf{R}^{\frac{H}{8}\times\frac{W}{8}\times 2C}$ extracted from the second stage of Swin Transformer, which provide a good balance between lo-

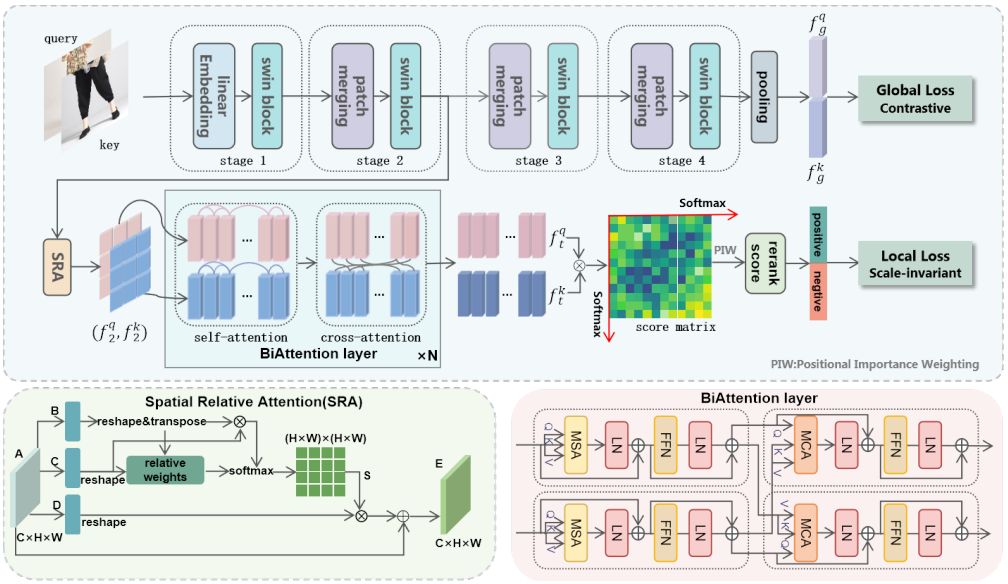

Fig. 1. Overview of the proposed global retrieval and re-ranking network. Global retrieval adopts the Swin Transformer as the backbone to extract global features and is trained with a contrastive loss.The re-ranking network extracts local features from the second stage of Swin Transformer. We apply spatial relative attention to enhancing spatial constraints.The BiAttention layer applies self-attention and cross-attention. The re-ranking network is trained with a scale-invariant loss.

cal detail and contextual information. Then, spatial relative attention is applied to enhance positional relevance based on relative positions. The feature maps are then reshaped into 1D sequences $(\tilde{f}_2^q, \tilde{f}_2^k) \in \mathbf{R}^{\frac{HW}{64} \times 2C}$ in spatial. These sequences are passed through a BiAttention module to produce more discriminative pairwise representations. Based on these representations, local patch-level matching is conducted. Finally, positional importance weighting is applied to compute a refined similarity score. This score is used to adjust the original global ranking for improved retrieval performance.

BiAttention Module. Inspired by LoFTR [16], we adopt a BiAttention module. This module extracts and aggregates key information both within individual images and between image pairs, as illustrated in Fig. 1. Self-attention layer establishes long-range dependencies to capture complex internal structures of the image. Different regions in the image may carry distinct semantic information. Self-attention enables the model to learn the inter dependencies among these regions, resulting in more discriminative features. For local features $X \in \mathbf{R}^{\frac{HW}{64} \times 2C}$, the self-attention computation is formulated as follows:

$$\{Q,K,V\} = \{XW^Q, XW^K, XW^V\}, \tag{2}$$

where $W^Q, W^K and W^V$ are learnable weight matrices. Then, we can use the scaling dot product to calculate self-attention:

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{3}$$

Multi-Head Self-Attention (MSA) enables the model attend to multiple representation subspaces in parallel.

$$\text{MSA}(Q,K,V) = \text{Concat}(head_1, head_2, ..., head_h)W. \tag{3}$$

The output passes through a feed-forward network (FFN). Layer normalization (LN) and residual connections are applied after MSA. The formula of self-attention layer is:

$$Z = \text{LN}\big(\text{MSA}(Q,K,V)\big) + Q, \tag{4}$$

$$Z' = \text{FFN}(Z) + Z. \tag{5}$$

The cross-attention mechanism strengthens correlation between query and candidate images by focusing on the most relevant parts of the images. It enables fine-grained matching. As shown in Fig. 1, the inputs to the cross-attention layers are derived from the self-attention outputs of image pair (Z^k, Z^q) or (Z^q, Z^k), implementing bidirectional interactions. The cross-attention layer is computed as follows:

$$\{Q_q, K_q, V_q\} = \{Z^q W^{Q_q}, Z^q W^{K_q}, Z^q W^{V_q}\}, \tag{6}$$

$$\{Q_k, K_k, V_k\} = \{Z^k W^{Q_k}, Z^k W^{K_k}, Z^k W^{V_k}\}, \tag{7}$$

$$z_q = \text{LN}\big(\text{MCA}(Q_q, K_k, V_k)\big) + Q_q, \tag{8}$$

$$z'_q = \text{FFN}(z_q) + z_q, \tag{9}$$

$$z_k = \text{LN}\big(\text{MCA}(Q_k, K_q, V_q)\big) + Q_k, \tag{10}$$

$$z'_k = \mathrm{FFN}(z_k) + z_k. \tag{11}$$

The BiAttention module is executed for N iterations. The first input of self-attention is $(\tilde{f}_2^q, \tilde{f}_2^k) \in \mathbf{R}^{64^{HW} \times 2C}$ and the final output after N iterations is denoted as (f_t^q, f_t^k).

Local Feature Patch Matching. Based on the feature mapping (f_t^q, f_t^k) obtained from BiAttention, we establish a local feature patch matching mechanism. It addresses the limitations of global features in fine-grained scenarios and severe occlusion. It applies threshold-based matching to local feature patches between query and candidate images. This enhances the model to distinguish subtle inter-class differences and suppress the similarity of images from different categories. For same-category images affected by occlusion, the matching regions strengthen local similarity. This compensates for missing information and improves their re-ranking rank. First, calculate the confidence matrix between features f_t^q and f_t^k:

$$S(i,j) = f_t^q(i) \cdot f_t^k(j), \tag{12}$$

where $i, j \in \mathbf{R}^{\frac{HW}{64}}$ are indices of the flattened patches of (f_t^q, f_t^k), respectively. Then, dual-softmax is applied on the matrix to obtain a bidirectional confidence matrix:

$$C(i,j) = \frac{\exp(S(i,j))}{\sum_{j'} \exp(S(i,j'))} \cdot \frac{\exp(S(i,j))}{\sum_{i'} \exp(S(i',j))}. \tag{13}$$

By setting a similarity threshold τ, we use the mutual nearest neighbor (MNN) algorithm to obtain matching feature patches between image pairs. We record the spatial positions and the match configuration confidence of each feature patch. The algorithm requires that query and candidate patches be each other's most similar match. This prevents one patch from matching multiple patches in the other image. The formula for the set of matched feature patches M_c is as follows:

$$M_c = \{(i,j) \mid j{=}\arg\max C(i,\cdot), i{=}\arg\max C(\cdot,j), C(i,j) > \tau\}. \tag{14}$$

To enhance retrieval performance, precise local correspondences are established during re-ranking. Patches that satisfy the mutual nearest neighbor (MNN) criterion are retained. The other mismatched patches are assigned zero similarity.The similarity between two patches is:

$$\mathrm{Sim}(i,j) = \begin{cases} 0, & (i,j) \notin M_c \\ C(i,j), & (i,j) \in M_c \end{cases}. \tag{15}$$

Then, the total similarity between the image pair is:

$$\mathrm{Sim}(I^q, I^k) = \sum_{i,j \in \mathbf{R}^{\frac{HW}{64}}} \mathrm{Sim}(i,j) = \sum_{(i,j) \in M_c} C(i,j). \tag{16}$$

Spatial Relative Attention (SRA). The relative positions of matched local patches influence matching reliability. Based on the intuition that smaller spatial displacement implies better correspondence under viewpoint variation, we propose a Spatial Relative Attention (SRA) mechanism.

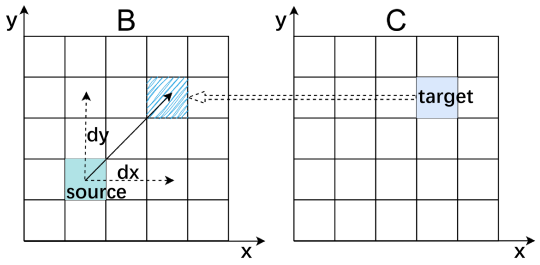

Fig. 2. Example of relative weights. dx and dy represent the relative distance from the source point of feature map B to the target point of feature map C, respectively.

Built on a spatial attention module[17], SRA incorporates directional-aware embeddings of horizontal and vertical offsets as relative positional encodings to model geometric consistency.

As shown in Fig. 1, given a local feature map $A \in \mathbf{R}^{H\times W\times C}$, we compute feature maps $B = conv_1(A)$ and $C = conv_2(A)$, then reshape to $\mathbf{R}^{C\times N}$, where $N = H \times W$. Next, compute the attention score:

$$s(i,j) = \text{softmax}(B_i C_j + s_h(i,j) + s_w(i,j)), \tag{19}$$

where $s(i,j)$ represents the influence of position i on the relative position of j. $s_h(i,j)$ and $s_w(i,j)$ are learnable horizontal and vertical relative position embeddings, respectively. As shown in Fig. 2, we compute $s_w(i,j)$=$PE(dx)$ and $s_h(i,j)$=$PE(dy)$, and PE is positional encoding. Feature map D=$conv_3(A)$ is also computed and reshaped, while multiplying by a scale parameter α. The final output is obtained by:

$$\mathrm{E}_j = \alpha \sum s(i,j) D_i + A_j. \tag{20}$$

Positional Importance Weighting (PIW). Patches matching at the center of two images are more important than those at the edge. To address varying importance in local regions, we design an adaptive center weight allocation mechanism. Patch positions are weighted by a 2-D Gaussian kernel to generate a spatial weight matrix. The formula is as follows:

$$G(x,y) = \exp\left(-\frac{(x-u_x)^2 + (y-u_y)^2}{2\sigma^2}\right), \tag{19}$$

where (x,y) is the position coordinate and (u_x,u_y) is the center coordinate. σ is the standard deviation of the Gaussian function, which controls the rate of weight decay. Flattening the weight matrix yields w_k, which weights the confidence values of matched feature patches of images (I^q, I^k):

$$\mathrm{Sim}(I^q, I^k) = \sum_{(i,j)\in M_c} w_i w_j C(i,j). \tag{20}$$

Scale-Invariant Loss. The similarity score between two images is the sum of confidence values over matched patches. It has not been normalized. Therefore, the conventional contrastive

loss is not suitable. We propose a scale-invariant loss that focuses on the relative similarity between positive and negative samples, using a dynamic scaling factor:

$$\mathscr{L}_{\mathscr{S}} = \text{ReLU}\left(\frac{s_n}{\text{mean}(s_p)+\varepsilon} - \gamma\right), \tag{21}$$

where s_p and s_n denote the similarities of positive and negative samples, respectively. γ is the dynamic scaling factor, indicating the tolerance ratio threshold. ε is a small constant to prevent division-by-zero. and Relu is the activation function.

4 Experiment

4.1 Dataset and Evaluation Metrics

The large-scale clothing image dataset was constructed based on the clothing dataset from Raycloud Technology Company (www.raycloud.com). Specifically, 52,120 images corresponding to 3,436 categories were selected for training, while 52,306 images corresponding to 3,398 categories were used for testing. 1,000 images were randomly chosen as the query set, while the remaining 51,306 images serving as the candidate set. The performance of the model was evaluated using Recall and mean Average Precision(mAP).

Recall@K. If at least one of the top K retrieval results has the same product ID, the score is 1, otherwise 0. Calculate the average recall rate of all query graphs:

$$\text{Recall@}K = \frac{1}{r}\sum_{q=1}^{r} \text{score}, \tag{22}$$

where r is the total number of query images, q is the current query image.

mAP@K. There are visually similar images in different categories, and the retrieval performance is evaluated by mAP. Calculate the Average Precision (AP) for each query, and the mean of the AP over all queries yields the mAP:

$$\text{mAP} = \frac{1}{Q}\sum_{q=1}^{Q} \text{AP}(q), \tag{23}$$

$$\text{AP} = \frac{1}{r}\sum_{K=1}^{r} \big(p(K)\cdot rel(K)\big), \tag{24}$$

where Q is the number of query images,q is the current query image. r is the number of top candidate images, K denotes the ranking position, $p(K)$ is the precision of the top K results, and $rel(K)$ is the score of the image at the K-th position.

4.2 Experimental Details

Experiments were conducted using PyTorch framework on an NVIDIA GeForce RTX 3080 Ti graphics. The Adam optimizer was employed with a learning rate of $3e^{-5}$ over 40 epochs. Data augmentation techniques included horizontal flipping, random rotation, and aspect ratio distortion, with all input images resized to 224×224 pixels.

Global Retrieval Module. The global retrieval module utilized the Swin Transformer as the backbone, which output feature dimension is 784. A contrastive loss function was applied with margins 0.5 for positives and 0.8 for negatives.

Re-ranking Network. The re-ranking network used the frozen second stage features of Swin Transformer. 4 Bi-Attention modules were incorporated, with 8 attention heads and a feature patch matching threshold of 0.2. A Gaussian kernel with σ=0.6 was used in PIW. A dynamic scaling factor of 0.7 was applied in the Scale-Invariant loss function.

4.3 Comparative Experiments

As shown in Table. 1, the proposed re-ranking network achieves superior retrieval performance on the clothing dataset. We first compare our model with five global retrieval baselines on their retrieval results. GeM [18] and ArcFace [19] are implemented on ResNet-50 and ResNet-101, respectively. GeM applies learnable generalized-mean pooling to boost global descriptors, while ArcFace shrinks intra-class variance for robust global feature learning. IRT [20] fine-tunes the transformer through metric learning and incorporates a differential entropy regularizer. However, these methods struggle to capture fine-grained differences in clothing image retrieval and remain sensitive to background clutter and pose variations. Compared with R50-GeM, R101-GeM, R50-ArcFace, R101-ArcFace, and IRT, our model improves Recall@1 by 4.7%, 4.0%, 1.7%, 1.5%, and 1.3%, and enhances mAP@100 by 15.43%, 14.78%, 8.46%, 7.93%, and 3.58%, respectively.

In the case of the same global feature retrieval, we compare the re-ranking model with four re-ranking methods. Global representation denotes our proposed global retrieval approach. αQE[18] relies on the average retrieval expansion features and ignore the local details and structural information between images. DL-ENDR[21] optimizes image similarity based on diffusion graph structures, but its diffusion paths are susceptible to noise when the initial retrieval results contain errors. SuperGlobal[7] uses feature enhancement to strengthen the global representation but lacks explicit alignment. CAS[14] emphasizes cross-scale information fusion but lacks clear region-level matching. In contrast, our method introduces a local feature block matching strategy that enhances both fine-grained similarity and robustness to occlusions. Compared with αQE, DL-ENDR, SuperGlobal, and CAS, our model achieves Recall@1 improvements of 3.2%, 1.3%, 1.6%, and 1.1%, and mAP@100 improvements of 5.26%, 3.08%, 0.27%, and 0.65%, respectively. More detailed results are illustrated in Fig. 3.

4.4 Ablation Studies

Re-ranking Network Components. In order to test the effect of the local feature patch matching re-ranking model, three ablation experiments were carried out.

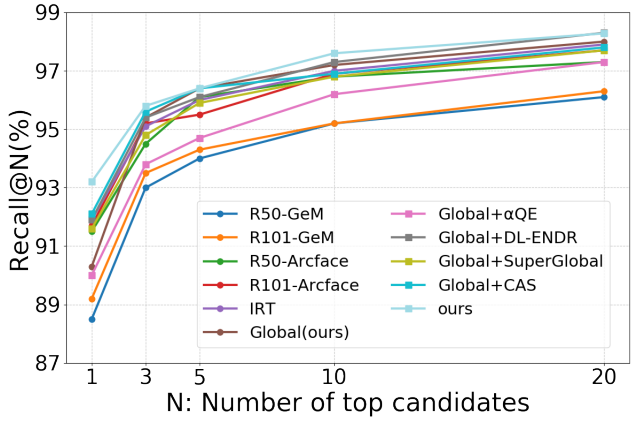

Fig. 3. Performance comparison of various methods. The recycle line corresponds to the retrieval results using global features, and the square line represents the results after applying a re-ranking method.

- w/o local: A network that relies solely on global features for retrieval without re-ranking.
- w/o PIW: Exclude the position importance weighting in the re-ranking network.
- w/o SRA: Remove the spatial relative attention from the re-ranking network.

Consistent with the results in Table. 2, after adding the reordering network, Recall@1, Recall@10, Recall@20 and mAP@100 increased by 2.9%, 0.4%, 0.2% and 1.07%, respectively. These results show that local feature patch matching enhances retrieval. Spatial relation attention strengthens spatial consistency, and positional importance weighting reduces the influence of irrelevant edge features.

Number of Re-ranking Images. To assess the impact of re-ranking quantity, experiments were conducted by re-ranking 100, 200, 300, and 400 images. As shown in Table. 3, Recall@1 improved by 2.9%, 2.8%, 3.1%, and 3.5%, respectively. Recall@20 increased by 0.1% in all settings. And mAP@100 improved by 1.07%, 0.94%, 0.92%, and 0.89%, respectively. These findings demonstrate that re-ranking enhances global image retrieval performance.

Impact of Local Features Extracted from Different Swin Transformer Stages. Features from the first to fourth stages of Swin Transformer were individually input into the re-ranking network for training. As shown in Fig. 4, the model achieved the highest accuracy when using local features extracted from the second stage. Specifically, the recall rates reached 93.20% for Recall@1, 96.40% for Recall@5, 97.60% for Recall@10, and 98.20% for Recall@20.

Furthermore, as illustrated in Fig. 5, Grad-CAM [22] was employed to generate heat maps for visualizing feature maps extracted from different stages. The first-stage features have high resolution but mainly capture edges. The third and fourth stages focus more on global semantics, discarding much of the local detail information. The second-stage features effectively capture clothing-specific details such as textures and collars while preserving semantic information, suiting

Table 1: Performance comparison of different methods.

Method	Recall (%)				mAP (%)
	@1	@5	@10	@20	@100
Global feature					
R50-GeM	88.50	94.00	95.20	96.10	51.03
R101-GeM	89.20	94.30	95.20	96.30	51.68
R50-Arcface	91.50	96.10	96.80	97.30	58.00
R101-Arcface	91.70	95.50	96.90	97.70	58.53
IRT	91.90	96.00	97.00	97.90	62.88
Global	90.30	**96.40**	97.20	98.00	65.39
Re-ranking					
Global+αQE	90.00	94.70	96.20	97.30	61.00
Global+DL-ENDR	91.90	96.10	97.30	**98.30**	63.38
Global+SuperGlobal	91.60	95.90	96.80	97.70	66.19
Global+CAS	92.10	**96.40**	96.90	97.80	65.81
ours	**93.20**	**96.40**	**97.60**	98.20	**66.46**

Table 2: Results of retrieval ablation experiment comparison.

Method	Recall (%)				mAP (%)
	@1	@5	@10	@20	@100
w/o local	90.30	96.40	97.20	98.00	65.39
w/o PIW	91.90	96.30	97.60	98.10	65.79
w/o SRA	92.70	96.20	97.60	98.10	66.23
full model	**93.20**	**96.40**	**97.60**	**98.20**	**66.46**

fine-grained matching. Therefore, local features from the second stage are adopted as input to the re-ranking network.

Extraction latency and Matching time. As illustrated in Table. 4, extraction latency and matching time are measured on NVIDIA GeForce RTX 3080 Ti graphics, for squared images of side 224. Our model has the matching time among the reproduced reranking methods. It has a extraction latency of 14.20 milliseconds and a matching time of 5.6 milliseconds in the reordering stage. And based on this, its mAP value is higher compared to other reranking models.

Visualization. As illustrated in Fig. 6, the proposed re-ranking network establishes dense local correspondences between image pairs. This allows it to focus on fine-grained and discriminative regions include patterns, collars, cuffs, and garment edges. These regions usually have strong discriminative properties and can effectively distinguish different styles or clothing with fine-grained differences. As shown in Fig. 7, compared to global retrieval alone, which struggles under complex visual variations, the proposed local feature matching strategy demonstrates improved robustness and retrieval accuracy. Furthermore, as shown in Fig. 8, in scenarios with limited positive sam-

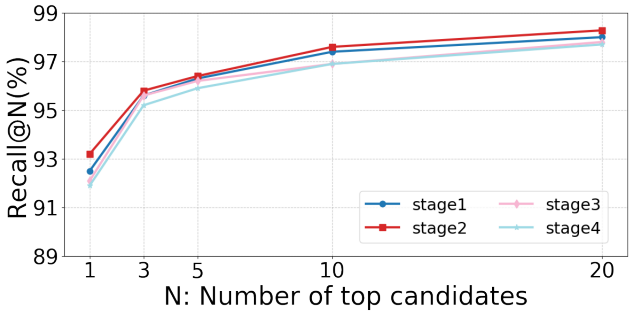

Fig. 4. Experimental results of local feature input reordering network extracted by swin transformer at different stages.

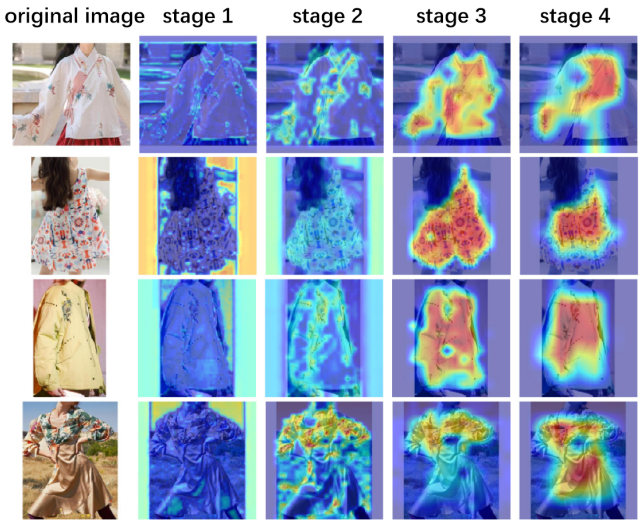

Fig. 5. Grad-CAM heat maps of the local feature input reordering network extracted by swin transformer at different stages.

Table 3: Results of different re-ranking numbers in retrieval.

Rerank Images	Recall (%)				mAP (%)
	@1	@5	@10	@20	@100
0	90.30	**96.40**	97.20	98.00	65.39
100	93.20	**96.40**	**97.60**	**98.20**	**66.46**
200	93.10	96.20	97.40	98.10	66.33
300	93.40	96.20	97.20	98.10	66.31
400	**93.80**	96.20	97.20	98.10	66.28

Table 4: Extraction Latency and Matching time

Method	Extraction Latency(ms)		Matching time(ms)
	Global	Rerank	
αQE	9.50	+55.38	11.60
DL-ENDR	9.50	+31.15	19.10
SuperGlobal	9.50	**+6.20**	7.40
CAS	9.50	+29.85	8.30
ours	9.50	+14.20	**5.40**

ples, the model retrieves visually similar results with consistent details, validating its scalability in large-scale clothing image retrieval.

5 Conclusions

We tackle the challenge of large-scale clothing image retrieval by proposing a novel re-ranking framework based on local feature block matching.Clothing images often exhibit subtle visual differences in patterns, colors, and textures, while also being affected by factors such as pose variation, occlusion, and background clutter. These challenges make it difficult for conventional global feature matching methods to accurately capture fine-grained visual similarities. To address this issue, the proposed framework integrates both global and local representations to enhance retrieval performance.

Specifically, the model employs the Swin Transformer as the backbone network to extract hierarchical visual features. To further enhance feature interaction, a BiAttention module is introduced to capture bidirectional fine-grained correlations between query images and candidate images, thereby improving the discriminative capability of the learned features. During the re-ranking stage, a SRA module is designed to model spatial relationships between matched local regions, effectively alleviating the misalignment problem caused by large relative displacements between clothing regions. In addition, a PIW strategy is proposed, which utilizes a two-dimensional Gaussian kernel to assign higher similarity weights to the central regions of feature maps. Extensive experiments conducted

Fig. 6. Local feature patch matching in the re-ranking network.

Fig. 7. Global retrieval and re-ranking results visualization: the first image is query image, green boxes indicate correct matches, red boxes indicate incorrect matches.

Fig. 8. Retrieval results visualization: the first image is query image, green boxes indicate correct matches, red boxes indicate incorrect matches.

on large-scale clothing datasets demonstrate that the proposed re-ranking framework consistently outperforms existing baseline methods across multiple evaluation metrics.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] S. Zhang, S. Liu, X. Cao, et al. Watch fashion shows to tell clothing attributes[J]. Neurocomputing, 2018, 282: 98-110.

[2] J. Chen, H. Yuan, Y. Zhang, et al. DCR-Net: Dilated convolutional residual network for fashion image retrieval[J]. Computer Animation and Virtual Worlds, 2023, 34(2): e2050.

[3] R. Hadsell, S. Chopra and Y. LeCun, Dimensionality Reduction by Learning an Invariant Mapping, 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 2006, pp. 1735-1742.

[4] K. Sohn, Improved deep metric learning with multi-class N-pair loss objective[C] // Neural Information Processing Systems. Curran Associates Inc. 2016.

[5] X. Wang, X. Han, W. Huang, D. Dong and M. R. Scott, Multi-Similarity Loss With General Pair Weighting for Deep Metric Learning, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 5017-5025.

[6] A. Gordo, F. Radenovic, T. Berg. Attention-based query expansion learning[C]//European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 172-188.

[7] S. Shao, K. Chen, A. Karpur, et al. Global features are all you need for image retrieval and reranking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 11036-11046.

[8] W. He, Z. Lu, X. Liu, et al. A real-time and high precision hardware implementation of RANSAC algorithm for visual SLAM achieving mismatched feature point pair elimination[J]. IEEE Transactions on Circuits and Systems I: Regular Papers, 2024.

[9] A. García-Hernández, R. Giubilato, H. StroblK, et al. Unifying local and global multimodal features for place recognition in aliased and low-texture environments[C]//2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024: 3991-3998.

[10] H. Xu. Encompass obstacle image detection method based on UV disparity map and RANSAC algorithm[J]. Scientific Reports, 2025, 15(1): 6164.

[11] S. Lee, H. Seong, S. Lee, et al. Correlation verification for image retrieval[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5374-5384.

[12] O. Jianbo, Z. Wengang, W. Min, T. Qi, L. Houqiang. Collaborative image relevance learning for visual re-ranking. IEEE Transactions on Multimedia, 2020.

[13] K. Wang, Y. Wang, L. Xue, et al. INRNet: Neighborhood Re-ranking Based Method for Pedestrian Text-Image Retrieval[J]. IEEE Access, 2024.

[14] J. Luo, H. Yao, C. Xu. Cluster-aware similarity diffusion for instance retrieval[J]. arXiv preprint arXiv:2406.02343, 2024.

[15] Z. Liu, Y. Lin, Y. Cao, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[16] J. Sun, Z. Shen, Y. Wang, et al. LoFTR: Detector-free local feature matching with transformers[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 8922-8931.

[17] J. Fu, J. Liu, H. Tian, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146-3154.

[18] F. Radenovic, G. Tolias and O. Chum, Fine-Tuning CNN Image Retrieval with No Human Annotation, in IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 41, no. 07, pp. 1655-1668, July 2019.

[19] J. Deng, J. Guo, N. Xue, et al. Arcface: Additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4690-4699.

[20] A. El-Nouby, N. Neverova, I. Laptev, et al. Training vision transformers for image retrieval[J]. arXiv preprint arXiv:2102.05644, 2021.

[21] W. Wanyin et al. Deep features for person re-identification on metric learning. Pattern Recognit. 110 (2021): 107424.

[22] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization[C]//Proceedings of the IEEE international conference on computer vision. 2017: 618-626.