Applied Intelligence (2025) 55:302
https://doi.org/10.1007/s10489-024-06173-8

From spatial to semantic: attribute-aware fashion similarity learning via iterative positioning and attribute diverging

Yongquan Wan[2,3] · Jianfei Zheng[1] · Cairong Yan[1] · Guobing Zou[2]

Accepted: 8 December 2024

Abstract

Fashion image retrieval emphasizes accurately perceiving the fine-grained features to meet users' precise needs. However, the existing global image-based retrieval methods encounter challenges such as imprecise positioning of attributes, difficulty in distinguishing visually similar but semantically different attribute values, and struggles in the learning of attribute features within specific regions and viewpoints. This paper proposes a two-stage hybrid framework called IPAD (Iterative Positioning and Attribute Diverging) for attribute-aware fashion similarity learning. In the initial stage, we present an iterative positioning strategy to precisely identify local attribute regions through an iterative attention mechanism with adaptive suppression. IPAD leverages the strengths of Convolutional Neural Networks and Vision Transformers. Subsequently, we design an attribute diverging strategy to optimize attribute value aggregation via online clustering using a momentum encoder, thereby enhancing model stability and representation. During inference, we further present a feature reasoning mechanism to refine retrieval results through subgraph similarity matrix generation and re-ranking to enhance accuracy and robustness. Extensive evaluations on three public datasets demonstrate IPAD's superior performance over state-of-the-art methods in retrieval accuracy, achieving an average improvement in MAP by +4.22%. The source code is available at https://github.com/h8e9r7/IPAD.

Keywords Image retrieval · Similarity learning · Iterative positioning · Attribute diverging · Feature reasoning

1 Introduction

The rise of e-commerce and online shopping [1–5] has significant advanced image retrieval technology [6–8]. This progress is driven by consumer's increasing demand for searching fashion items with specific details, promoting crucial research into identifying key visual attributes (such as patterns, materials, and styles) in images. This research focuses on learning to represent these attributes across various image feature regions [9–11]. Since attribute features can span large or small areas within an image, and there can be notable variations between different values of the same attribute, accurately understanding and representing these features remains a significant challenge.

With the continuous advancement of technology, researchers have increasingly employed advanced computer vision techniques for attribute-aware image retrieval [12–18]. This includes leveraging spatial attention and channel attention mechanisms to locate and identify specific attributes in fashion images [16]. These methods aim to enhance the system's understanding of subtle fashion attributes across the entire image, facilitating consumers in finding desired products by specifying details. However, the current positioning accuracy remains insufficient, leading to poor discriminability. For instance, existing approaches struggle to accurately identify visually similar yet semantically different attribute values, such as "half-high" and "ruffled collars", which posses distinct semantic nuances despite their spatial resemblance [19].

Yongquan Wan and Jianfei Zheng contributed equally to this paper

✉ Cairong Yan
cryan@dhu.edu.cn

✉ Guobing Zou
gbzou@shu.edu.cn

Yongquan Wan
wanyq@gench.edu.cn

Jianfei Zheng
1945306734@qq.com

[1] School of Computer Science and Technology, Donghua University, Shanghai 201620, China

[2] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China

[3] College of Information Technology, Shanghai Jian Qiao University, Shanghai 201306, China

Published online: 13 January 2025

Additionally, certain attributes present challenges to feature learning due to their unique components and viewing angles, consequently undermining retrieval accuracy. Furthermore, the absence of effective reasoning mechanisms further hampers the enhancement of retrieval-aware image retrieval [20].

To address these challenges, we propose an innovative two-stage framework for attribute-aware fashion similarity learning called Iterative Positioning and Attribute Diverging (IPAD), which combines Convolutional Neural Networks (CNNs) [21] and Vision Transformers (ViTs) [22] techniques for fashion image retrieval. Our approach begins with an iterative attention module featuring adaptive attention suppression. This module enhances the precision of attribute positioning through iterative spatial and channel attention while adaptively suppressing edge-irrelevant attention in spatial attention.

To tackle the consistency issue in real-time updating of cluster centers [23, 24], we employ a momentum encoder [25] and design a cluster center loss function. This function effectively constrain the feature representation of each instance to remain near the cluster center of its corresponding attribute label, thereby enhancing the model's ability to recognize the intrinsic structure of image data and improving network stability and representation capability.

Moreover, to further improve fine-grained fashion image retrieval, we propose a feature reasoning mechanism that generates a subgraph similarity matrix [26] for each query. During the reference stage, subgraph updates and feature optimization are performed via gradient descent. This approach aggregates updates to iteratively improve the overall similarity adjacency matrix, effectively overcoming re-ranking dependence and stability issues caused by direct learning through the similarity matrix. Ultimately, this boosts the model's accuracy and robustness through feature space optimization.

The main contributions of this paper are as follows:

- We propose a two-stage hybrid framework (IPAD) for attribute aware fashion similarity learning. IPAD leverages the strengths of convolutional neural networks and vision Transformers by adopting an attention-based iterative positioning strategy and an attribute diverging strategy based on online clustering using a momentum encoder. This approach addresses the problems of imprecise attribute positioning and semantic discrimination.
- We develop a feature reasoning mechanism for the inference stage that refines retrieval results through subgraph similarity matrix generation and re-ranking. This re-ranking process involves subgraph updates and feature optimization via gradient descent, enhancing both accuracy and robustness.
- Extensive evaluations on the FashionAI [27], DARN [28], and DeepFashion [29] datasets demonstrate IPAD's superior performance over state-of-the-art solutions in retrieval accuracy. Our approach yields MAP improvements of +4.11%, +7.51%, and +1.05%, respectively, showcasing significant advantages in fine-grained fashion image retrieval.

2 Related work

2.1 Traditional fashion image retrieval

Traditional fashion retrieval primarily emphasizes the overall similarity of fashion images [28, 30–32]. For example, in the cross-domain retrieval task, Huang et al. [28] proposed a Dual Attribute-aware Ranking Network (DARN) to solve the cross-domain image retrieval problem, significantly enhancing retrieval performance through semantic attribute learning and triplet similarity constraints. Ji et al. [31] utilized the rich label information on e-commerce websites to assist the attention mechanism in locating clothing in complex scenes, subsequently completing in-shop retrieval based on the extracted clothing features. Kang et al. [30] introduced the "Complete the Look" task, addressing the gap in existing research on predicting product image compatibility by recommending visually compatible products in complex real-world scene images. In the fashion compatibility task, Han et al. [32] trained a bidirectional LSTM model to sequentially predict the next item in an outfit based on previous items, leaning the compatibility relationship between them.

In contrast to these studies that primarily focus on whole-image similarity, fine-grained fashion image retrieval targets attribute-aware feature similarity, offering a more detailed and precise approach that global similarity measurements cannot adequately address.

2.2 Fine-grained fashion image retrieval

Fine-grained fashion image retrieval focuses on the features of local regions related to specified fashion attributes, rather than the features of the entire image [12–18]. Most methods adopt an end-to-end single-branch network to extract attribute-aware feature representations [12–16]. Veit et al. [12] proposed the CSN model, which learns embeddings in semantically different sub-spaces by mask selection and re-weighting relevant dimensions, addressing the issue that traditional similarity embeddings cannot capture multiple similarity concepts. Ma et al. [16] introduced the ASEN model, which first locates attribute-related regions through Attribute-aware Spatial Attention (ASA) and further extracts fine-grained features through Attribute-aware Channel Attention (ACA). Building on ASEN, Wan et al. [13] enhanced the relationship between attributes by adaptively fusing the features of ASA and ACA. Yan et al. [14] proposed

HAEN, which hierarchically extracts attribute embeddings according to the type and relationship of attributes. Following this, Yan et al. [15] further proposed ISLN, which iteratively extracts fine-grained features by repeatedly utilizing attention, presenting a general framework for iterative attention.

Different from the single-branch structure, Dong et al. [17] proposed a dual-branch network based on ASEN, where each branch has the same network architecture and extracts fine-grained features from global and local perspectives, respectively. Subsequently, Dong et al. [18] designed a Coarse-to-Fine dual-branch network, where each branch processes features at different granularities, demonstrating the advantage of a dual-branch network. They introduced the E-InfoNCE loss and utilized attribute-irrelevant background features to optimize model performance and amplify the distinguishing capacity of attribute-specific representations.

Unlike previous studies that focus on a dual-branch structure, we present a two-stage dual-branch network featuring a distinctive feature representation based on two strategies: iterative positioning and attribute divergence. Furthermore, rather than concentrating solely on model architecture, our approach enhances retrieval efficacy by incorporating a feature reasoning mechanism to optimize features during inference.

3 Proposed method

We propose IPAD, a two-stage CNNs & ViTs based hybrid framework for attribute embedding representation, as shown in Fig. 1. IPAD optimizes feature representations through two modules: Iterative Positioning (IP) and Attribute Diverging (AD). The IP module comprises two branches: IP-SP (in Space) and IP-SE (in Semantics), both of which extract attribute-level embeddings from fashion images. The IP-SP branch employs an iterative attention with adaptive suppression to accurately localize attribute regions by suppressing weaker attention activation regions in each iteration. The IP-SE branch refines attribute feature expressions using an attribute cross-attention network based on the vision Transformer [22], processing local images activated by the IP-SP branch's attention.

To enhance attribute-level clustering performance, we propose the AD module, which computes and forms clustering centers [33] using a momentum encoder. This promotes tight clustering of clothing features with the same attribute values, thereby constraining the network's clustering space representation.

During inference, we introduce the Feature Reasoning (FR) module, trained through supervised learning, to optimize the original similarity matrix and re-rank the image sequence for improved retrieval accuracy. Further details will be provided in subsequent sections.

3.1 Iterative positioning strategies in space and semantics

Consider a set of fashion images $\{I\}$, with attributes $\{a \in A\}$ and attribute value labels $\{v \in V_a\}$ associated with a specific attribute a. Given an image I and a specific attribute a,

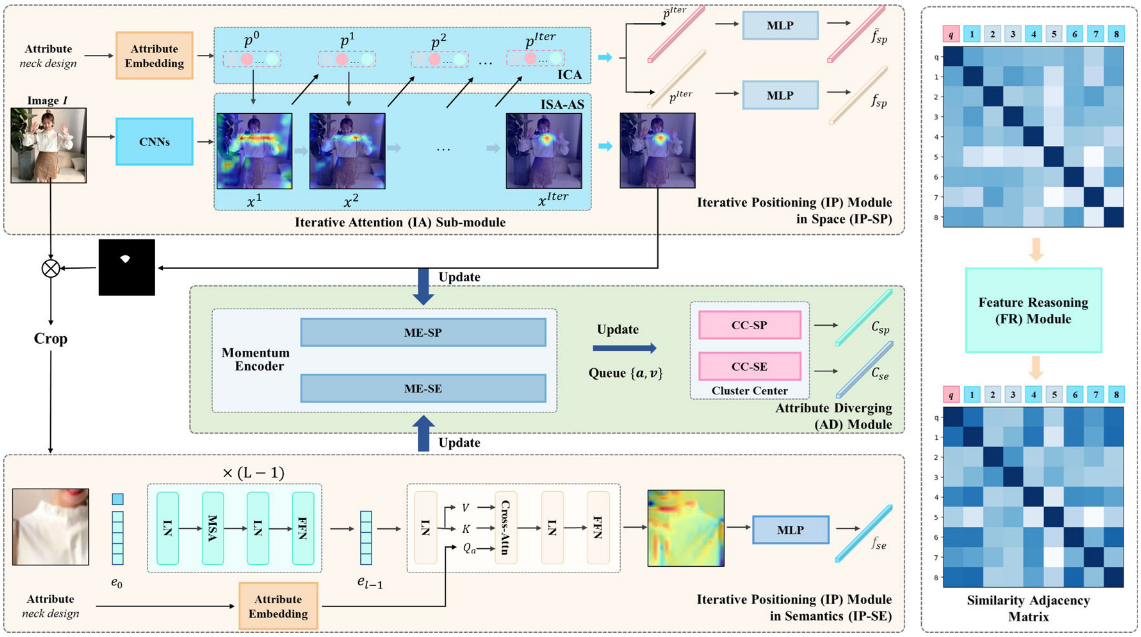


Fig. 1 The proposed IPAD framework. It consists of three key modules: (a) Iterative Positioning (IP) module, including two branches (IP-SP and IP-SE), (b) Attribute Diverging (AD) module, and (c) Feature Reasoning (FR) module

IPAD needs to learn the feature vector $f(I, a)$ for the specific attribute. We define the parameters of IPAD as θ, where the parameters of the IP-SP branch and IP-SE branch of IP module are represented by θ_{sp} and θ_{se}, respectively. The d-dimensional feature vectors learned by the two branches are denoted as $f_{sp}(I, a) \in \mathbb{R}^d$ and $f_{se}(I, a) \in \mathbb{R}^d$, respectively.

3.1.1 IP-SP: iterative positioning in space

Given an image and its attributes, we first utilize a CNN backbone, specifically ResNet50 [34], to obtain the features of the entire image $x \in \mathbb{R}^{c\times h\times w}$, where c represents the number of channels, and $h \times w$ corresponds to the image size. Meanwhile, the corresponding attribute a is encoded through an embedding network into $p^0 \in \mathbb{R}^c$. We design an attention iteration network with $Iter$ iterations, where each iteration produces two types of features: attribute features $p^i \in \mathbb{R}^c$ and image features $x^i \in \mathbb{R}^{c\times h\times w}$, where $i \in Iter$. Given the $i-1^{\text{th}}$ iteration process, the network generates feature representations p^{i-1} and x^{i-1}.

The main sub-module in IP-SP is the Iterative Attention (IA). Iterative Spatial Attention (ISA) and Iterative Channel Attention (ICA) components are used to obtain more precise feature representations in this sub-module. Specifically, the ISA component combines p^{i-1} and x^{i-1} using spatial attention [16] to generate the representation of x^i for the next iteration. Then, x^i and p^{i-1} are passed through the ICA component to obtain the representation of p^i for the next iteration using channel attention [16]. p^i and x^i are used as input for the next iteration. By iteratively stacking the ISA and ICA components, the Adaptive Suppression (AS) component learns to gradually extract fine-grained features based on ISA.

ISA: Iterative Spatial Attention In the $i-1^{\text{th}}$ iteration, due to the differing modalities between the image $x^{i-1} \in \mathbb{R}^{c\times h\times w}$ and the attribute vector $p^{i-1} \in \mathbb{R}^c$, we first project them into a shared latent space $P_s \in \mathbb{R}^{c'\times h\times w}$. Subsequently, we compute the attention weights $\alpha_s^{i-1} \in \mathbb{R}^{c'\times h\times w}$.

$$\alpha_s^{i-1} = P_s\left(x^{i-1}\right) \cdot P_s\left(p^{i-1}\right). \tag{1}$$

AS: Adaptive Suppression To improve attribute localization when iterative spatial attention encounters difficulty in further shrinking activation regions, we introduce an adaptive suppression component with an activation threshold. Initially, in the $i-1^{\text{th}}$ iteration, we aggregate the attention weights α_s^{i-1} across the channel dimension c' to produce an activation map $A^{i-1} \in \mathbb{R}^{h\times w}$. The intensity at each position in A^{i-1} reflects its discriminative ability, allowing us to identify the most relevant regions in the space. We introduce a threshold η and a ratio ζ of the maximum activation intensity. Thus, during each iteration, the threshold $\eta^i = \zeta \cdot \max_{h,w}\left(A^i\right)$ automatically adjusts based on A^i. Consequently, we compute the mask region $M^{i-1} \in \mathbb{R}^{h\times w}$ generated in the $i-1^{\text{th}}$ iteration.

$$M^{i-1}(h, w) = \chi(A^{i-1}(h, w) \geq \eta^{i-1}), \tag{2}$$

where χ is the indicator function.

Finally, we perform a spatial multiplication of M^{i-1} and α_s^{i-1} using the dot product operation to derive a refined attention representation $\alpha_{s'}^{i-1} = M^{i-1} \cdot \alpha_s^{i-1}$ at a reduced scale. Subsequently, we compute the spatial attention weights specific to the attribute by summing $\alpha_{s'}^{i-1}$ along the channel dimension c', followed by normalization adjustment by dividing by $\sqrt{c'}$. This normalization step helps to modulate the channel features and reduce dimensionality. The resultant weights $\alpha_s'^{i-1} \in \mathbb{R}^{h\times w}$ are then processed through a Software activation function for normalization across spatial dimensions. Finally, $\alpha_s'^{i-1}$ is applied to x^{i-1} via element-wise multiplication to compute the output $x^i \in \mathbb{R}^{c\times h\times w}$ of the ISA component in the $i-1^{\text{th}}$ iteration.

$$x^i = \alpha_s'^{i-1} \cdot x^{i-1}. \tag{3}$$

ICA: Iterative Channel Attention In the $i-1^{\text{th}}$ iteration, ICA takes the output x^i from ISA and p^{i-1}, which is mapped to a c'-dimensional vector $P_c\left(p^{i-1}\right) \in \mathbb{R}^{c'}$ through a fully connected layer followed by a ReLU activation function. Subsequently, this vector is concatenated with the result of summing x^i along the channel dimension c, and then passed through two fully connected layers with a reduction rate r for dimensionality reduction and expansion [35]. The Sigmoid activation function is applied to obtain the channel attention weights $\alpha_c^{i-1} \in \mathbb{R}^c$. By performing the dot product between α_c^{i-1} and p^{i-1}, we derive the output $p^i \in \mathbb{R}^c$ of ICA in the $i-1^{\text{th}}$ iteration. p^{Iter} serves as the feature representation learned by the IA sub-module.

$$p^i = \alpha_c^{i-1} \cdot p^{i-1}. \tag{4}$$

Foreground and Background Representation Contrasting with prior methods that prioritize foreground (attention-activated) features while disregarding background elements as irrelevant or noisy, we argue that these background regions, devoid of the attribute, are crucial negative counterparts [18] to foreground features, enriching learning via contrast. Within the ISA component, background spatial attention weights are derived via $\mathbf{1} - \alpha_s'^i$ (where $\mathbf{1} \in \mathbb{R}^{h\times w}$ symbolizes the spatial expansion matrix), enabling the extraction of background features $\tilde{p}^{Iter}$, parallel to the attribute-specific p^{Iter}, by applying identical iterative procedures.

Finally, we employ a multilayer perceptron (MLP) with layer normalization (LN) and residual connections to obtain the final outputs $f_{sp}(I, a)$ and $\tilde{f}_{sp}(I, a)$ of the IP-SP branch.

$$f_{sp}(I, a) = LN\left(MLP_{\text{skip}}\left(p^{Iter}\right)\right), \tag{5}$$

$$\tilde{f}_{sp}(I, a) = LN\left(MLP_{\text{skip}}\left(\tilde{p}^{Iter}\right)\right), \tag{6}$$

where LN denotes layer normalization [36], and MLP_{skip} refers to a multilayer perceptron with skip (residual) connections.

3.1.2 IP-SE: iterative positioning in semantics

Within the IP-SE branch, we design an attribute-specific representation network based on the ViT architecture to capture subtle semantic differences in specific regions. This branch takes an image as input, initially cropped from the IP-SP branch's attention activation map and then resized back to its original size, along with a specific attribute input. The IP-SE branch conducts a cross-attention operation between the image's embeddings and those of the specified attribute, resulting in the branch's feature representation through a multilayer perceptron. The essence of our attribute-specific adaptation resides in the attribute-aware cross-attention mechanism, defined as follows:

$$l_i = \text{softmax}\left(\frac{Q_a K^T}{\sqrt{D}}\right)V, \tag{7}$$

where the query Q_a is directly derived from the attribute embedding, enabling a nuanced integration of attribute-specific insights into the visual representation. To synthesize the output $f_{se}(I, a)$ of the IP-SE branch, we concatenate and project the attention heads, followed by layer normalization and a feed-forward network, as outlined below:

$$l = FC_{cross}\left([l_1, l_2, \ldots, l_h]\right), \tag{8}$$

$$f_{se}(I, a) = l + FFN\left(LN(l)\right), \tag{9}$$

where FFN is a Feed-forward Network.

3.2 Attribute diverging strategy using momentum encoder

We dynamically update cluster centers using online clustering with image attribute labels a and value labels v, tightly clustering instance features around their $\{a, v\}$ centers. This approach enhances network stability and representational efficacy by avoiding the outdated centers issue of offline clustering, adapting to data shifts in real time [37]. A momentum encoder updates cluster centers every n_b mini-batches, incorporating both a main network and a momentum-updated auxiliary network. The auxiliary network's parameters are refined using those of the main network with a fixed momentum, ensuring smooth feature updates and consistency [25] across batches. We develop two sub-modules: Momentum Encoder (ME) and Cluster Center (CC).

ME: Momentum Encoder A dictionary queue, keyed by $\{a, v\}$ labels with feature queues as values, is maintained, with each queue length set to n_q. The ME sum-module enqueues features based on labels per batch, replacing the oldest batch with the current one when full, thereby preserving data consistency. The ME for IPAD's dual branches, $\hat{\theta} = \{\hat{\theta}_{sp}, \hat{\theta}_{se}\}$, replicates network parameters for both IP-SP and IP-SE branches, with $\hat{\theta}_{sp}$ and $\hat{\theta}_{se}$ denoting the respective ME parameters. Gradient updates for ME are disabled; instead, parameters are momentum-updated as follows:

$$\hat{\theta}_{\{sp,se\}} \leftarrow \mu_{\{sp,se\}}\hat{\theta}_{\{sp,se\}} + (1 - \mu_{\{sp,se\}})\theta_{\{sp,se\}}, \tag{10}$$

where $\mu_{\{sp,se\}} \in [0, 1)$ are momentum coefficients for IP-SP and IP-SE branches, with only θ_{sp} and θ_{se} undergoing back-propagation updates.

CC:Cluster Center The CC sub-module updates cluster centers after n_b mini-batches using queued features:

$$C_{\{sp,se\}}(a, v) = \frac{1}{n_q} \sum \{\hat{f}_{\{sp,se\}} | a, v\}, \tag{11}$$

with $C_{\{sp,se\}}(a, v) \in \mathbb{R}^d$ representing cluster center features for each branch.

3.3 Feature reasoning mechanism for re-ranking

In the inference phase, given a query set $\{query\}$ and a candidate set $\{candidate\}$ for attribute a, we generate a similarity matrix [38] from features derived by the original IPAD, refined by our Feature Reasoning (FR) module through subgraph updates and feature vector optimization via gradient descent [39]. After IPAD generates features and ranks them based on feature similarity, we focus on the top $n_r - 1$ candidate images most similar to the query. We merge a query image and its corresponding top $n_r - 1$ candidate images into a sequence of length n_r. For $i \in \{1, 2, \ldots, n_r\}$, f_i^0 is the initial feature of the i^{th} image, with its label value being y_i. The initial similarity adjacency matrix is $S^0 = \left[s_{ij}^0\right]_{i,j \in \{1,2,\ldots,n_r\}}$, obtained by calculating the cosine similarity between each pair of images: $s_{ij}^0 = \frac{\langle f_i^0 \cdot f_j^0 \rangle}{\|f_i^0\| \|f_j^0\|}$. The goal is to predict the target similarity adjacency matrix $\hat{S} = \left[\mathbb{1}_{y_i = y_j}\right]_{i,j \in \{1,2,\ldots,n_r\}}$, where $\mathbb{1}_{(y_i = y_j)}$ is the indicator function for $y_i = y_j$. However, the similarity adjacency matrix depends on the order of nodes, i.e., the arrangement of rows and columns, making it challenging to learn and predict directly from the similarity adjacency matrix because different permutations of rows and columns lead to difficulties and overfitting in the learning process [40]. To address this issue, we design an architecture that is invariant to node order, aiming to enable the model to effectively learn from the similarity adjacency matrix without being limited by the node ordering.

For each node i in S, a subgraph centered on i is created, with an associated similarity matrix δ_i. The nodes in δ_i are sorted based on their similarity to the center node i. Subsequently, we use a simple optimization network o to accept δ_i as input and update δ_i, i.e., $o(\delta_i)$. After updating each subgraph similarity matrix, we aggregate all updates by summing them at their corresponding positions in S to update the entire graph, defined as $O(S) = \left[o(\delta_1), o(\delta_2), \ldots, o(\delta_{n_r})\right]$. By processing subgraphs centered on each node and aggregating predictions, this method circumvents the complexity and permutation issues of learning from the full adjacency matrix, effectively leveraging subgraph structure while ensuring node order invariance. Subsequently, we can iteratively update the similarity adjacency matrix $\widetilde{S}^t = \left[\widetilde{s}^t_{ij}\right]_{i,j \in \{1,2,\ldots,n_r\}}$ through the FR module, where t represents the t^{th} iteration.

$$\hat{S}^t := S^{t-1} + O(S^{t-1}). \tag{12}$$

Discovering that direct updates to the similarity adjacency matrix introduce non-linearity and non-positive semidefiniteness [41], compromising its physical significance and computational stability, we shift to updating features f_i^t for the i^{th} image in the t^{th} round. This strategy, updating $S^t = \left[s^t_{ij}\right]_{i,j \in \{1,2,\ldots,n_r\}}$ indirectly via features using gradient descent on the Euclidean distance loss [42], denoted as $\Delta_f(f) := \frac{1}{2}\sum_{i,j}\left(\widetilde{s}^t_{ij} - \frac{\langle f_i, f_j\rangle}{\|f_i\|\|f_j\|}\right)^2$, addresses these challenges and targets optimization of an intermediate result $\hat{S}^t$, which is the optimization target.

$$\begin{aligned}
\forall i : f_i^t &= f_i^{t-1} - \rho \frac{\partial \Delta_f}{\partial f_i}(f^{t-1}),\\
\frac{\partial \Delta_f}{\partial f_i}(f^{t-1}) &= -\sum_j \left(\widetilde{s}^t_{ij} - s^{t-1}_{ij}\right) \frac{\partial s^{t-1}_{ij}}{\partial f_i^{t-1}}(f^{t-1}),
\end{aligned} \tag{13}$$

where ρ is the step size. Through this approach, the feature update in each iteration step aims to reduce the discrepancy between the updated similarity adjacency matrix and the target similarity adjacency matrix.

3.4 Objectives

3.4.1 Loss functions

Leveraging the IP module of IPAD, we acquire feature representations for both the overall image and specific attribute regions. Drawing from prior work [18], we employ intra- and inter-branch contrastive losses to boost attribute sensitivity and feature distinctiveness. To enhance network stability and representational quality by ensuring feature representations remain closely clustered, we introduce a cluster center loss. Additionally, we propose an FR contrastive loss for the inference phase.

IP-SP Branch and IP-SE Branch Contrastive Loss The IP-SP and IP-SE branches of IP target identifying key image attributes, aiming to align images sharing the same attribute value (e.g., sleeve length) closely in embedding spaces, while distancing those with differing values, ensuring that, for instance, images of short and long sleeves are distinctly separated [43]. Specifically, we construct triplets $T = \left(I_i, I_i^+, I_i^- | a\right)_{i=1}^{n_b}$ within a mini-batch, where the attribute value of I_i is the same as I_i^+ and different from I_i^-. They are all sampled under the same attribute a, and n_b is the batch size. Therefore, the contrastive loss for the IP-SP branch and IP-SE branch is defined as:

$$\begin{aligned}
\mathcal{L}_T^{\{sp,se\}} = & \frac{1}{n_b}\sum_{i=1}^{n_b} \max(0, m - d(f_{\{sp,se\}}(I_i, a), f_{\{sp,se\}}(I_i^+, a))\\
& + d(f_{\{sp,se\}}(I_i, a), f_{\{sp,se\}}(I_i^-, a))),
\end{aligned} \tag{14}$$

where m is the margin constraint, and $d(,)$ is the cosine distance function.

IP-SP to IP-SE Contrastive Loss To further improve the model's performance in distinguishing foreground and background features, following the previous work [18], we consider that for the same image, the foreground is a positive sample, while the background can be treated as harmful information and viewed as a negative sample. This approach enhances the model's ability to distinguish between foreground and background and improves the model's ability to identify attribute-related regions, thereby boosting the performance of the IP-SE branch. The loss function is defined as:

$$\begin{aligned}
\mathcal{L}_E &= -\frac{1}{n_b}\sum_{i=1}^{n_b} \log\left(\frac{\exp\left(f_{sp}\left(I_i, a\right) \cdot f_{se}\left(I_i, a\right)/\tau\right) + \Psi}{\Omega + \Psi + \varepsilon\Phi}\right),\\
\Psi &= \sum_{f_{se}\left(I_j^+, a\right) \in \psi} \exp\left(f_{sp}\left(I_i, a\right) \cdot f_{se}(I_j^+, a)/\tau\right),\\
\Phi &= \sum_{\widetilde{f}_{sp}\left(I_j, a\right) \in Neg_k} \exp\left(f_{sp}\left(I_i, a\right) \cdot \widetilde{f}_{sp}\left(I_j, a\right)/\tau\right),
\end{aligned} \tag{15}$$

where ψ represents the set of foreground positive samples with the same attribute value as I_i within the mini-batch, and Neg_k represents the set of backgrounds with the same attribute value as I_i within the mini-batch, serving as the negative sample set. ε is the scale factor, and a larger value indicates a stronger penalty on the similarity between the foreground and background representations of the same attribute.

Cluster Center Loss We define the loss function $\mathcal{L}_{CC}$, applicable to both IP-SP and IP-SE branches, for an image I with label $\{a, v\}$, comprising the attribute and its value. The features extracted by the two branches are $f_{sp}(I, a)$ and $f_{se}(I, a)$, respectively. For attribute a, the cluster center of positive samples is denoted by $C^+_{\{sp,se\}}(a) = C_{\{sp,se\}}(a, v)$, with the negative sample sets defined as $C^-_{\{sp,se\}}(a) = \{C_{\{sp,se\}}(a, 0), C_{\{sp,se\}}(a, 1), \ldots, C_{\{sp,se\}}(a, M_n)\} \backslash \{C_{\{sp,se\}}(a, v)\}$,
each excluding the positive center and comprising $M_n - 1$ elements. The loss function, $\mathcal{L}_{CC}^{\{sp,se\}}$, is expressed as:

$$\mathcal{L}_{CC}^{\{sp,se\}}(I, a|v) = \frac{1}{M_n - 1} \sum_{c}^{\{C^-_{\{sp,se\}}(a)\}} \max\left(0, m - d\left(f_{\{sp,se\}}(I, a), C^+_{\{sp,se\}}(a)\right) + d\left(f_{\{sp,se\}}(I, a), c\right)\right). \tag{16}$$

FR Contrastive Loss During inference, the FR module is employed to refine the similarity matrix for improved retrieval outcomes by re-ranking a query image against the top $n_r - 1$ candidates from initial similarity assessments. The objective is to amplify similarity with positive samples and diminish it with negatives, ensuring positive samples are ranked higher. Additionally, we introduce the FR contrastive loss, derived from the InfoNCE [44] loss function, to support this goal.

$$\mathcal{L}_O = \sum_{i,j: y_i = y_j = y_{query}} -\log\left(\frac{\exp\left(s_{ij}^{n_t}/\tau\right)}{\exp\left(s_{ij}^{n_t}/\tau\right) + \sum_{k: y_k \neq y_i} \exp\left(s_{ik}^{n_t}/\tau\right)}\right), \tag{17}$$

where $s_{ij}^{n_t}$ represents the similarity between the i^{th} sample image and the j^{th} sample image after the n_t^{th} round of updates.

3.4.2 Training and inference

The target loss during the training process is divided into two stages. When the queue of ME is fully filled, it transitions from the first stage to the second stage. The total loss is defined as:

$$\begin{aligned}
\mathcal{L}_{\text{warm}} &= \mathcal{L}_T^{\text{sp}} + \lambda_1 \mathcal{L}_T^{\text{se}} + \lambda_2 \mathcal{L}_E, && (18)\\
\mathcal{L}_{\{sp,se\}} &= \mathcal{L}_{\text{CC}}^{\{sp,se\}}(I, a|v) + \mathcal{L}_{\text{CC}}^{\{sp,se\}}(I^+, a|v^+) \\
&\quad + \mathcal{L}_{\text{CC}}^{\{sp,se\}}(I^-, a|v^-), && (19)\\
\mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{warm}} + \xi_1 \mathcal{L}_{sp} + \xi_2 \mathcal{L}_{se}, && (20)
\end{aligned}$$

where λ_1, λ_2, ξ_1, and ξ_2 are hyper-parameters. During the inference stage, we integrate the features from the IP module, the IP-SP branch, and the IP-SE branch according to predefined weights. These features have already been optimized through the AD process.

$$f = \sqrt{\beta} \cdot \text{norm}(f_{sp}) + \sqrt{1 - \beta} \cdot \text{norm}(f_{se}), \tag{21}$$

where β is the weight hyper-parameter and $norm()$ is the normalization.

For inference, we use the FR contrastive loss $\mathcal{L}_O$ directly.

4 Experiments

4.1 Experimental settings

4.1.1 Datasets

Three datasets were utilized to evaluate the model: FashionAI [27], DARN [28], and DeepFashion [29]. These datasets contain 180K, 254K, and 289K images respectively. Each dataset was split into training, validation, and test sets with a ratio of 8:1:1. We randomly selected a small number of samples from the training set for re-ranking training. During the training, 100K triplets were constructed by randomly selecting images for positive and negative samples. Data augmentation [45] was applied to the anchor samples. For testing, the images were divided into query and candidate images at a 1:4 ratio.

4.1.2 Metrics

Mean Average Precision (MAP) is used as evaluation metric, widely recognized in retrieval tasks [16], and reported as percentages (%).

4.1.3 Implementation details

The ResNet50 pretrained on ImageNet [46] serves as the backbone for the IP-SP branch, while the ViT-B/16 is used for the IP-SE branch. A two-stage training strategy was employed. Hardware-wise,model training utilized 32 RTX 3090 GPUs. Empirically, we set $\zeta = 0.95$ to control the atten-

Table 1 Performance comparison on the FashionAI dataset

Method	MAP for each attribute								MAP
	Skirt length	Sleeve length	Coat length	Pant length	Collar design	Lapel design	Neckline design	Neck design	
Baseline	17.20	12.50	13.35	17.45	22.36	21.63	11.09	21.19	15.79
Triplet	48.38	28.14	29.82	54.56	62.58	38.31	26.64	40.02	38.52
CSN [12]	61.97	45.06	47.30	62.85	69.83	54.14	46.56	54.47	53.52
ASEN [16]	64.44	54.63	51.27	63.53	70.79	65.36	59.5	58.67	61.02
HAEN [14]	64.13	55.52	56.41	72.31	73.32	69.22	62.41	59.80	64.13
AttnFashion [13]	65.70	56.46	54.64	71.12	74.45	69.36	65.69	65.54	65.37
ISLN [15]	65.91	58.83	56.45	71.22	74.53	70.55	65.71	65.61	66.10
ASEN^{++} [17]	66.34	57.53	55.51	68.77	72.94	66.65	66.81	67.01	64.31
RPF [18]	66.75	67.86	59.65	73.23	75.72	73.18	74.40	75.01	70.11
IPAD	68.93	68.32	64.85	76.88	79.37	**79.96**	79.44	75.66	73.69
IPAD+FR	**69.17**	**70.25**	**65.99**	**77.29**	**80.04**	79.69	**79.47**	**75.71**	**74.22**

The bold entries represent the best results in each column, indicating statistical significance

tion decay in the adaptive suppression module. The number of iterations for the IA sub-module was fixed at $Iter = 3$.

For the ME sub-module, momentum update coefficients were set to $\mu_{sp} = \mu_{se} = 0.995$, the queue length $n_q = 1000$, and the CC sub-module was updated every $n_b = 50$ iterations. In the FR module, the total sequence length was set to $n_r = 500$. During training, the margin value m was set to 0.2 [47], temperature factor $\tau = 0.07$, and penalty coefficient $\varepsilon = 12$.

For $\lambda_1, \lambda_2, \xi_1$, and ξ_2, values were set to 0.1, 0.04, 1.0, and 1.0, respectively, based on prior experience to ensure initial loss values for each component were relatively balanced.

4.2 Performance comparison

We evaluated our method against previous state-of-the-art methods on three datasets. Table 1 compares the performance on the FashionAI dataset. The competing methods including traditional methods (a random ranking baseline and the triplet network), single-branch models (CSN [12], ASEN [16], HAEN [14], AttnFashion [13], ISLN [15]), and dual-branch models (ASEN^{++} [17] and RPF [18]). Our IPAD method demonstrated superior performance, achieving an overall MAP of 73.69. Integrating the proposed FR mechanism further enhanced the performance to 74.22. Additionally, IPAD significantly outperformed the previous best model, RPF, on multiple specific attributes, such as lapel design, where IPAD's MAP surpassed RPF by 6.78%. These results highlight the effectiveness of our two-branch iterative solution and the re-ranking method.

We also evaluated IPAD on the DARN and DeepFashion datasets as shown in Tables 2 and 3. IPAD consistently outperformed all competing methods on every attribute, consolidating its leading position across multiple datasets and

Table 2 Performance comparison on the DARN dataset

Method	MAP for each attribute									MAP
	Category	Button	Color	Length	Pattern	Shape	Collar shape	Sleeve length	Sleeve shape	
Baseline	8.49	24.45	12.54	29.90	43.26	39.76	15.22	63.03	55.54	32.26
Triplet	23.59	38.07	16.83	39.77	49.56	47.00	23.43	68.49	56.48	40.14
CSN [12]	34.10	44.32	47.38	53.68	54.09	56.32	31.82	78.05	58.76	50.86
ASEN [16]	28.81	42.17	47.78	48.55	48.95	47.09	25.67	78.46	56.25	47.08
HAEN [14]	32.10	47.04	45.03	48.27	49.92	51.22	28.05	78.29	58.47	48.70
AttnFashion [13]	34.94	48.56	48.14	54.47	52.65	56.36	32.32	82.63	60.77	52.32
ISLN [15]	38.84	51.26	52.67	56.55	53.85	58.34	36.64	82.74	61.28	54.68
ASEN^{++} [17]	40.15	50.42	53.78	60.38	57.39	59.88	37.65	83.91	60.70	55.94
RPF [18]	45.18	54.92	55.08	63.51	57.04	63.54	41.20	86.95	62.43	58.80
IPAD	55.58	64.34	55.72	69.40	**64.53**	72.10	50.30	91.06	69.78	65.78
IPAD+FR	**56.18**	**64.68**	**56.01**	**70.53**	64.50	**72.34**	**50.66**	**91.55**	**69.98**	**66.31**

The bold entries represent the best results in each column, indicating statistical significance

Table 3 Performance comparison on the DeepFashion dataset

Method	MAP for each attribute					MAP
	Texture	Fabric	Shape	Part	Style	
Baseline	6.69	2.69	3.23	2.55	1.97	3.38
Triplet	13.26	6.28	9.49	4.43	3.33	7.36
CSN [12]	14.09	6.39	11.07	5.13	3.49	8.01
ASEN [16]	15.01	7.32	13.32	6.27	3.85	9.14
AttnFashion [13]	12.90	6.34	11.38	5.24	4.20	8.01
ASEN^{++} [17]	15.60	7.67	14.31	6.60	4.07	9.64
RPF [18]	15.62	8.30	15.02	7.38	4.77	10.22
IPAD	15.67	**8.57**	16.11	**8.77**	6.31	11.18
IPAD+FR	**15.88**	8.53	**16.46**	8.76	**6.73**	**11.27**

The bold entries represent the best results in each column, indicating statistical significance

verifying the universality and effectiveness of our approach. It is worth noting that some baseline methods did not provide results on the DeepFashion dataset.

4.3 Ablation studies

Ablation studies on the FashionAI dataset evaluated IPAD's modules and components.

4.3.1 Impact of IP-SP and IP-SE branches

We compared three variants to evaluate the impact of the IP module. Here, the AD module is not considered.

- w/o SE: Only the IP-SP branch in the IP module is kept.
- w/o SP: Only the IP-SE branch in the IP module is kept.
- w/ IP: The entire IP module is kept.

As shown in Fig. 2, the comparison between the IP-SP and IP-SE branches of the IP module in FashionAI reveals that w/ IP achieves a notably MAP of 73.17%, outperforming the individual branches which achieve 68.47% (w/o SE) and 70.45% (w/o SP). This superiority is particularly evident in attributes with diverse visual characteristics such as pant length and collar design, underscoring the effectiveness of integrating spatial and semantic insights.

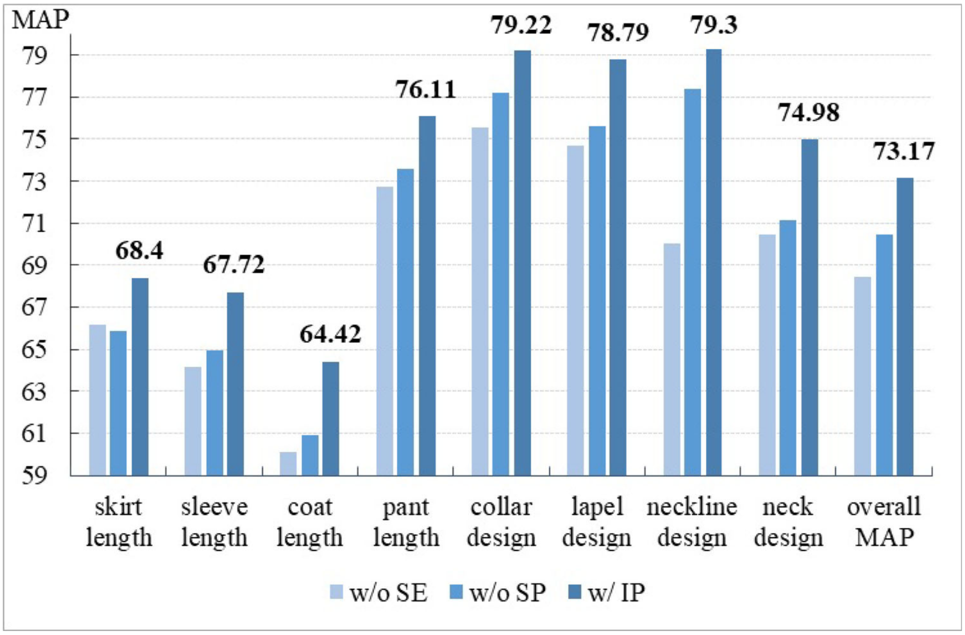
Additionally, the IP-SE branch demonstrates an advantage in collar design and neckline design, emphasizing the significance of semantic features in handling complex attributes. These findings strongly advocate for a comprehensive approach to clothing features extraction through the fusion of IP-SP and IP-SE, validating the importance of integrating multiple sources of information for sophisticated visual recognition tasks.

4.3.2 Impact of the components in IP-SP

We investigated the IP-SP branch of the IP module for fine-grained image retrieval, focusing on the three components in the IA sub-module. The AS component aims to enhance recognition accuracy by iteratively emphasizing spatial aspects and optimizing attention distribution. Here, the AD module and the IP-SE branch are not considered.

- w/o ISA: Only the ICA component in the IA sub-module is kept.
- w/o ICA: Only the ISA and AS components in the IA sub-module are kept.
- w/o AS: Only the ISA and ICA components in the IA sub-module are kept.

Fig. 2 An ablation study examining the IP-SP and IP-SE branches of the IP module. It shows that IPAD excels in recognizing complex attributes, surpassing the individual IP-SP and IP-SE branches



- w/ IA: The entire ISA, ICA, and AS components in the IA sub-module are kept.

The results revealed that ICA and ISA significantly enhance model performance, as shown in Fig. 3. With ICA alone, the model achieved an MAP of 60.46%. This improved to 63.98% with ISA, highlighting ISA's capability in enhancing spatial feature detection. Utilizing both ICA and ISA without AS yielded an MAP of 67.42%, underlining their synergistic effect. The integration of AS with ISA and ICA further elevated the MAP to 67.9%, highlighting AS's role in refining attention by reducing focus on less relevant features. This improvement is particularly evident in the recognition of complex fashion attributes like collar design and sleeve length. These results underscore the critical contribution of ISA, ICA, and AS in advancing fashion attribute recognition and provide insights for enhancing similar models.

4.3.3 Impact of the ME sub-module in AD

We compared three variants to evaluate the impact of the ME sub-module in the AD module.

- w/o ME: The ME sub-module is removed from the AD module.
- w/o ME-SP: Only the ME-SP component is removed from the ME sub-module.
- w/o ME-SE: Only the ME-SE component is removed from the ME sub-module.

The results, as shown in Fig. 4, yield insightful conclusions. The integration of the ME sub-module notably enhances the model's MAP scores, particularly when fully utilized, demonstrating its capability to improve attribute recognition accuracy. Optimal performance is achieved with ME activated in both the ME-SP and ME-SE components, resulting in a MAP of 73.69%, a 1.63 percentage point increase over configurations without ME. This suggests a synergistic effect of ME.

The impact of ME varies across attributes, with significant improvements observed in detailed-feature attributes such as collar and neckline designs. A comparison between configurations without ME in the ME-SP versus ME-SE components indicates that the ME-SE branch, in particular, benefits more from ME for certain attributes, highlighting the encoder's differential contribution.

These findings underscore the importance of the momentum encoder in IPAD for enhancing attribute recognition. They suggest that ME's full potential is realized when employed across both components to efficiently integrate global and local information. Furthermore, the attribute-specific responses to ME highlight the necessity for tailored strategies to maximize performance improvements.

4.3.4 Hyper-parameter analysis

As shown in Fig. 5, we varied the parameter $Iter$ in the iterative attention module from 1 to 10 in intervals of 1. The performance of the IA sub-module reaches its peak when $Iter$ is set to 3. These results indicate that multiple iterations of attention are beneficial for improving the performance of IPAD, although the effectiveness slightly diminishes after more than 3 iterations. Therefore, it is recommended to carefully consider the number of attention iterations.

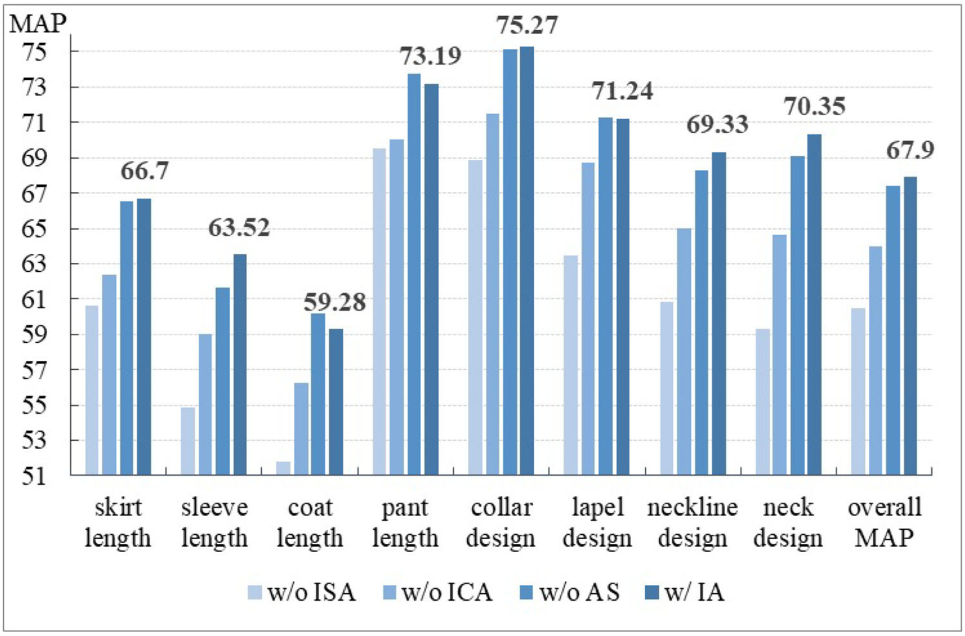


Fig. 3 An ablation study examining the three components in the IA sub-module. It shows that ISA, ICA, and AS each individually improve fashion attribute recognition accuracy

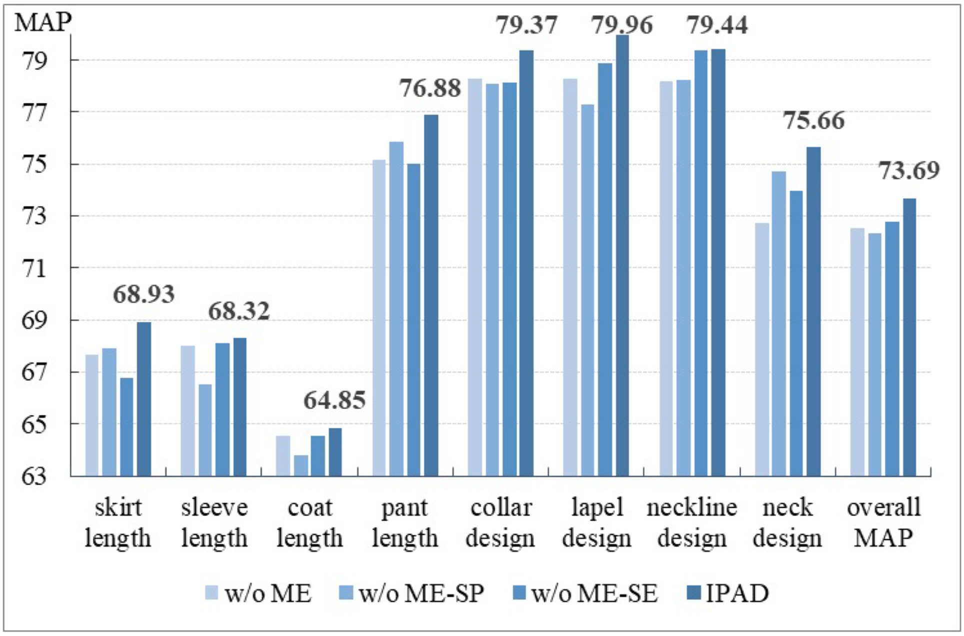


Fig. 4 An ablation study examining the momentum encoder. It shows that the ME enhances MAP, especially when utilizing both IP-SP and IP-SE components

Additionally, the influence of β in (21) is illustrated in Fig. 6. We adjusted the value of β from 0 to 1 in intervals of 0.1, and the performance of IPAD is optimal when β is 0.4. The finding suggests that IPAD tends to leverage the capabilities of IP-SE more effectively at this value.

4.4 Visualization analysis

4.4.1 Effect of iterative learning

To investigate IPAD's attribute localization capabilities, we performed a visualization analysis of the attribute activation regions [48], as shown in Fig. 7. The iterative attention sub-module demonstrates its dynamic nature: Initial attention distributions are dispersed, covering various potential attribute regions; As iterations progress, the focus sharpens on specific, relevant areas, illustrating the model's ability to refine attribute localization through iterative learning. For instance, in recognizing sleeve length, attention shifts from a broad focus on the upper body to precise areas such as cuffs and sleeves. Similarly, for collar design recognition, attention becomes finely tuned to collar shapes and details.

By binarizing the final iteration's attention weights, we created distinct attribute activation region masks, clearly

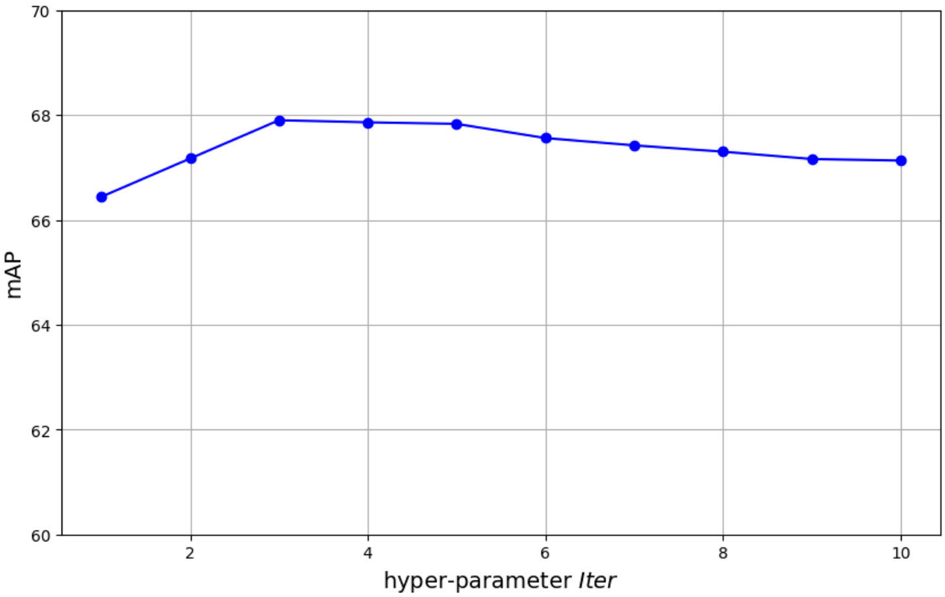


Fig. 5 The influence of hyper-parameter $Iter$

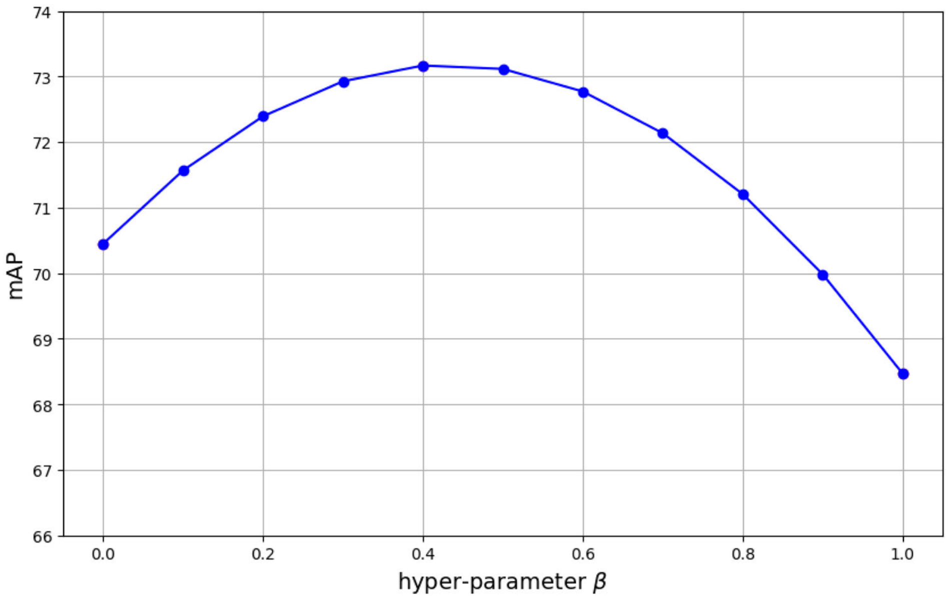


Fig. 6 The influence of hyper-parameter β

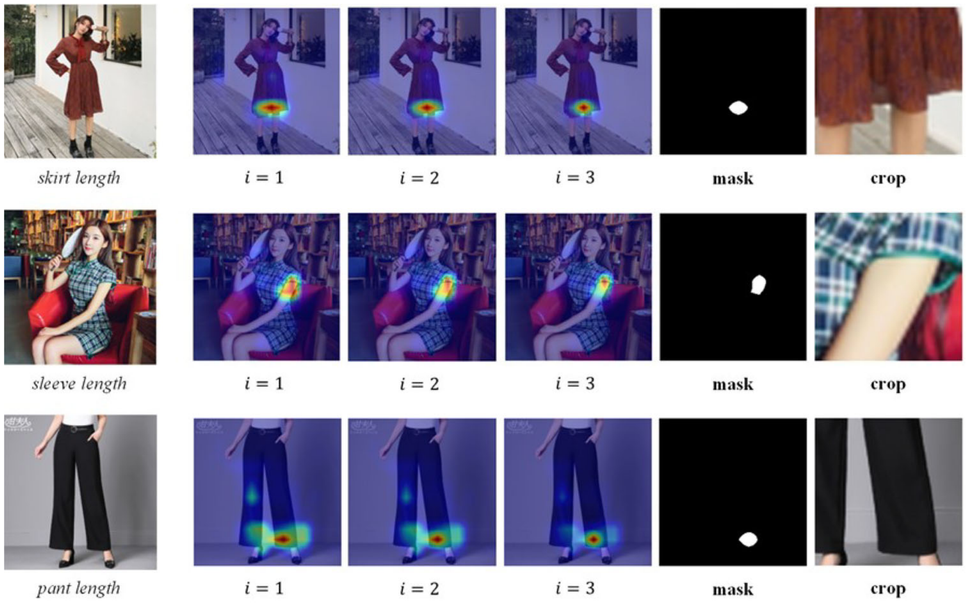


Fig. 7 Visualization of attribute activation regions through iterative learning. The initial attention distributions are dispersed, but as iterations progress, the focus sharpens on relevant areas

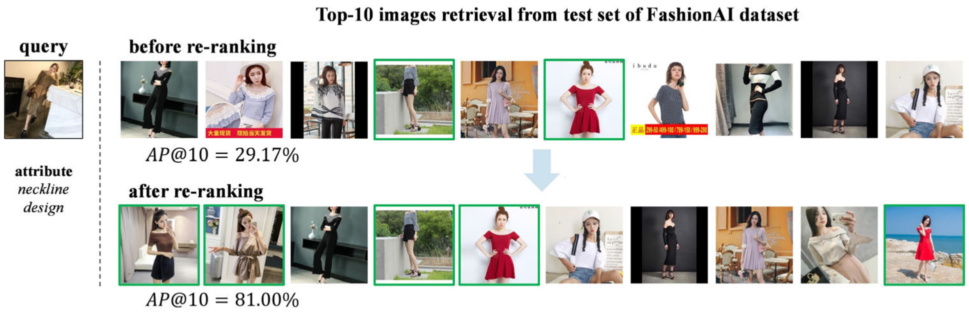


Fig. 8 Visualization of retrieval results before and after re-ranking. It illustrates the improvement in retrieval accuracy with the FR module, which prioritizes images with correct attributes

defining the areas considered most relevant by the model. Cropping these regions from the original images yields patches rich in attribute-specific information, highlighting the model's discriminative precision. These visualizations not only affirm the IPAD model's proficiency in progressively concentrating attention to enhance attribute recognition but also its capability in deciphering complex fashion details. They offer a clear, intuitive insight into the model's operational mechanics.

4.4.2 Effect of re-ranking

Figure 8 presents an example of retrieval results before and after re-ranking, visually illustrating the effectiveness of the FR module. It is observed that through re-ranking, more images with correct attributes are placed at the front of the sequence, leading to an improvement in MAP. However, the FR module can occasionally cause re-ranking failures when the images are influenced by irrelevant information, such as background elements.

4.5 Complexity analysis

This section presents a detailed complexity analysis of the proposed IPAD model, encompassing both model size and computational overhead during inference. We evaluate the model's trainable parameters and the required floating-point operations (FLOPs) for encoding a single image. The IPAD model contains approximately 142M parameters and requires 3.44G FLOPs. Additionally, the extraction of attribute-specific features for a single image takes an average of 11ms, achieving real-time response capability for system implementation. These benchmarks were obtained using a system equipped with 64GB of memory and a single NVIDIA RTX 3090 GPU.

5 Conclusion

This paper introduces the IPAD framework, which integrates iterative positioning strategy, attribute diverging strategy, and feature reasoning mechanism to significantly enhance the precision and robustness of fine-grained fashion image retrieval. By leveraging innovative iterative attention and online clustering methods, we effectively address challenges related to inaccurate attribute localization and semantic differentiation. Experimental evaluations on diverse datasets, including FashionAI, DARN, and DeepFashion, validate IPAD's effectiveness, showcasing substantial improvements in attribute recognition and image retrieval accuracy.

This research contributes novel techniques and insights to the filed of fine-grained fashion image retrieval. Future work will focus on enhancing the model's generalization and real-time performance, as well as optimizing similarity computation to achieve a more accurate and efficient retrieval system.

Acknowledgements This work was partly supported by the National Natural Science Foundation of China (No. 62477006). I would also like to thank the reviewers for their insightful suggestions, which have helped improve the quality of this research.

Author Contributions Cairong Yan: Conceptualization, Methodology, Commenting on the proposed idea, Writing-reviewing & editing. Jianfei Zheng: Software, Data curation, Investigation. Yongquan Wan: Conceptualization, Methodology, Formal analysis. Guobing Zou: Writing reviewing & editing.

Data availability and access The authors declare that the used three datasets are open online.

Declarations

Ethical and informed consent for data used The authors declare that they comply with ethical and informed consent for data used.

Competing interests The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Ma J, Sun H, Yang D, Zhang H (2024) Personalized fashion recommendations for diverse body shapes and local preferences with contrastive multimodal cross-attention network. ACM Transactions on Intelligent Systems and Technology
2. De Divitiis L, Becattini F, Baecchi C, Del Bimbo A (2023) Disentangling features for fashion recommendation. ACM Trans Multimed Comput Commun Appl 19(1):1–21
3. Ma Y, Ding Y, Yang X, Liao L, Wong WK, Chua T-S (2020) Knowledge enhanced neural fashion trend forecasting. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, pp 82–90
4. Mall U, Matzen K, Hariharan B, Snavely N, Bala K (2019) Geostyle: Discovering fashion trends and events. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 411–420
5. Al-Halah Z, Stiefelhagen R, Grauman K (2017) Fashion forward: Forecasting visual style in fashion. In: Proceedings of the IEEE International Conference on Computer Vision, pp 388–397
6. Cheng W-H, Song S, Chen C-Y, Hidayati SC, Liu J (2021) Fashion meets computer vision: A survey. ACM Comput Surv 54(4):1–41
7. Dubey SR (2021) A decade survey of content based image retrieval using deep learning. IEEE Trans Circuits Syst Video Technol 32(5):2687–2704
8. Deldjoo Y, Nazary F, Ramisa A, Mcauley J, Pellegrini G, Bellogin A, Noia TD (2023) A review of modern fashion recommender systems. ACM Comput Surv 56(4):1–37
9. Duan K, Parikh D, Crandall D, Grauman K (2012) Discovering localized attributes for fine-grained recognition. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp 3474–3481. IEEE

10. Huynh D, Elhamifar E (2020) Fine-grained generalized zero-shot learning via dense attribute-based attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4483–4493
11. Seo Y, Shin K-s (2018) Image classification of fine-grained fashion image based on style using pre-trained convolutional neural network. In: Proceedings of the 3rd International Conference on Big Data Analysis, pp 387–390. IEEE
12. Veit A, Belongie S, Karaletsos T (2017) Conditional similarity networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 830–838
13. Wan Y, Yan K, Yan C, Zhang B (2022) Learning attribute-guided fashion similarity with spatial and channel attention. J Exp Theor Artif Intell 1–17
14. Yan C, Ding A, Zhang Y, Wang Z (2021) Learning fashion similarity based on hierarchical attribute embedding. In: Proceedings of the 8th International Conference on Data Science and Advanced Analytics, pp 1–8. IEEE
15. Yan C, Yan K, Zhang Y, Wan Y, Zhu D (2022) Attribute-guided fashion image retrieval by iterative similarity learning. In: Proceedings of the 2022 IEEE International Conference on Multimedia and Expo, pp 1–6. IEEE
16. Ma Z, Dong J, Long Z, Zhang Y, He Y, Xue H, Ji S (2020) Fine-grained fashion similarity learning by attribute-specific embedding network. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 11741–11748
17. Dong J, Ma Z, Mao X, Yang X, He Y, Hong R, Ji S (2021) Fine-grained fashion similarity prediction by attribute-specific embedding learning. IEEE Trans Image Process 30:8410–8425
18. Dong J, Peng X, Ma Z, Liu D, Qu X, Yang X, Zhu J, Liu B (2023) From region to patch: Attribute-aware foreground-background contrastive learning for fine-grained fashion retrieval. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 1273–1282
19. Deldjoo Y, Nazary F, Ramisa A, Mcauley J, Pellegrini G, Bellogin A, Noia TD (2023) A review of modern fashion recommender systems. ACM Comput Surv 56(4):1–37
20. Tan F, Yuan J, Ordonez V (2021) Instance-level image retrieval using reranking transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 12105–12115
21. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp 1097–1105
22. 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
23. Yin J, Zhang X, Ma Z, Guo J, Liu Y (2023) A real-time memory updating strategy for unsupervised person re-identification. IEEE Trans Image Process 32:2309–2321
24. Dai Z, Wang G, Yuan W, Zhu S, Tan P (2022) Cluster contrast for unsupervised person re-identification. In: Proceedings of the Asian Conference on Computer Vision, pp 1142–1160
25. He K, Fan H, Wu Y, Xie S, Girshick R (2020) Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9729–9738
26. Shen X, Xiao Y, Hu SX, Sbai O, Aubry M (2021) Re-ranking for image retrieval and transductive few-shot classification. In: Proceedings of the Advances in Neural Information Processing Systems, pp 25932–25943
27. Zou X, Kong X, Wong W, Wang C, Liu Y, Cao Y (2019) Fashionai: A hierarchical dataset for fashion understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 296–304
28. Huang J, Feris RS, Chen Q, Yan S (2015) Cross-domain image retrieval with a dual attribute-aware ranking network. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1062–1070
29. Liu Z, Luo P, Qiu S, Wang X, Tang X (2016) Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 1096–1104
30. Kang W-C, Kim E, Leskovec J, Rosenberg C, McAuley J (2019) Complete the look: Scene-based complementary product recommendation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 10532–10541
31. Ji X, Wang W, Zhang M, Yang Y (2017) Cross-domain image retrieval with attention modeling. In: Proceedings of the 25th ACM International Conference on Multimedia, pp 1654–1662
32. Han X, Wu Z, Jiang Y-G, Davis LS (2017) Learning fashion compatibility with bidirectional lstms. In: Proceedings of the 25th ACM International Conference on Multimedia, pp 1078–1086
33. Li Y, Hu P, Liu Z, Peng D, Zhou JT, Peng X (2021) Contrastive clustering. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 8547–8555
34. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778
35. Hu J, Shen L, Sun G (2018) Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7132–7141
36. Ba JL, Kiros JR, Hinton GE (2016) Layer normalization. arXiv preprint arXiv:1607.06450
37. Zhan X, Xie J, Liu Z, Ong Y-S, Loy CC (2020) Online deep clustering for unsupervised representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6688–6697
38. Yuan Z, Zhang W, Tian C, Rong X, Zhang Z, Wang H, Fu K, Sun X (2022) Remote sensing cross-modal text-image retrieval based on global and local information. IEEE Trans Geosci Remote Sens 60:1–16
39. Gong D, Zhang Z, Shi Q, Hengel A, Shen C, Zhang Y (2020) Learning deep gradient descent optimization for image deconvolution. IEEE Trans Neural Netw Learn Syst 31(12):5468–5482
40. Huang Z, Wang Y, Li C, He H (2022) Going deeper into permutation-sensitive graph neural networks. In: Proceedings of the International Conference on Machine Learning, pp 9377–9409
41. Kim S, Kojima M, Mevissen M, Yamashita M (2011) Exploiting sparsity in linear and nonlinear matrix inequalities via positive semidefinite matrix completion. Math Program 129(1):33–68
42. Pavllo D, Grangier D, Auli M (2018) Quaternet: A quaternion-based recurrent model for human motion. arXiv preprint arXiv:1805.06485
43. Dong X, Shen J (2018) Triplet loss in siamese network for object tracking. In: Proceedings of the European Conference on Computer Vision, pp 459–474
44. Oord Avd, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748
45. Xiao L, Zhang X, Yamasaki T (2023) Toward a more robust fine-grained fashion retrieval. In: Proceedings of the 6th International Conference on Multimedia Information Processing and Retrieval, pp 1–4

46. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp 248–255
47. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 815–823
48. Jalwana MA, Akhtar N, Bennamoun M, Mian A (2021) Cameras: Enhanced resolution and sanity preserving class activation mapping for image saliency. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 16327–16336

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.