Fashion Microscope: Pixel-Level Attribute Perception via Optimal Transport and Neural Semantic Aggregation

Shuili Zhang[12], Hongzhang Mu[1], Jiawei Sheng[12], Qianqian Tong[3], Wenyuan Zhang[12], Quangang Li[1], Tingwen Liu[12]*

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Department of Strategic and Advanced Interdisciplinary Research, Peng Cheng Laboratory, Shenzhen, China
{zhangshuili, muhongzhang, shengjiawei, zhangwenyuan, quangangli, liutingwen}@iie.ac.cn, [3]tongqq@pcl.ac.cn

Abstract

Attribute-specific fashion retrieval aims to enhance fine-grained image retrieval by emphasizing the similarity of specific attributes. Current methods primarily rely on attention mechanisms to extract attribute-related visual features but face two key challenges: the limitations of coarse-grained localization in achieving fine-grained accuracy, and an imbalance between global and local perception, where excessive focus on local features can undermine overall performance. To address these issues, we propose the fashion microscope ***Pro*Fashion**, which achieves pixel-level attribute awareness through optimal transport and neural semantic aggregation. The framework begins by employing optimal transport to align semantic attributes with visual patterns from a global perspective, generating an attribute-visual value map that highlights distinctive regions while reducing interference. This is followed by simulating the human brain's perception of attribute feature patterns through superpixel generation and aggregation, capturing attribute-related features at the pixel semantic level and forming key semantic clusters that preserve microstructures. Building on this, an attribute graph is constructed to facilitate feature clustering, significantly enhancing the framework's capability to handle overlapping features and cross-scale relationships. Comprehensive experiments on the *FashionAI*, *DeepFashion*, and *DARN* datasets demonstrate the framework's effectiveness, achieving overall MAP improvements of **3.11%**, **3.70%**, and **3.49%**, respectively. Additionally, the framework delivers relative average throughput gains of **26.94%**, **22.22%**, and **24.78%** on the FashionAI, DeepFashion, and DARN datasets, respectively.

Introduction

Unlike traditional fashion image retrieval (Tran et al. 2019; Liu et al. 2021a; Ma et al. 2022), the key difficulty of the **A**ttribute-**S**pecific **F**ashion **R**etrieval (ASFR) lies in accurately identifying attribute-relevant features or patterns in images based on the given attributes and retrieving similar images based on these features or patterns (Ma et al. 2020; Song and Han 2022; Dong et al. 2023). For instance, given a query image and an attribute (e.g., *skirt length*), ASFR leverages the visual features of the image along with the attribute to retrieve fashion images from the database that align with the desired attribute details (e.g., image featuring *ankle length*). This joint image-text retrieval approach allows for more accurate capture of fine-grained features, significantly improving the relevance and precision of search results (Veit, Belongie, and Karaletsos 2017; Dong et al. 2021; Jiao et al. 2022, 2023). This task is particularly challenging because spatially localized attribute patterns (e.g., *neckline design*, *sleeve length*) occupy only small regions of the image, whereas globally distributed ones (e.g., *fabric*, *texture*) exhibit heterogeneous and fragmented distributions. This necessitates both integrating multi-regional cues and disentangling overlapping characteristics to capture the discriminative representations, as illustrated in Figure 1.

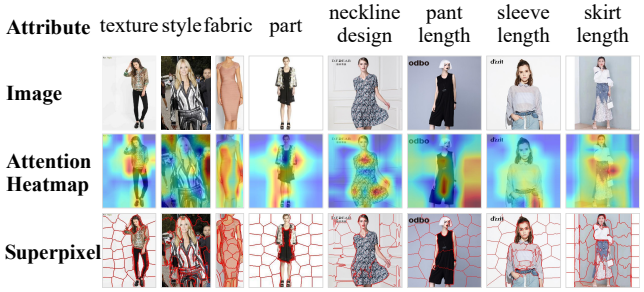

Figure 1: Examples of differences in the distribution of visual attribute values at different granularities: attribute, images, attention heatmap, and superpixel.

In recent years, researchers have primarily focused on attribute-guided attention mechanisms to capture attribute-aware features. The ASEN model (Ma et al. 2020) pioneered this direction by employing spatial and channel attention for fine-grained feature extraction. Subsequent works introduced ASEN variants (Yan et al. 2021; Wan et al. 2024; Yan et al. 2022a; Dong et al. 2021), further refining visual attribute representation. Advanced approaches enhanced fine-grained modeling through repeated attention mechanisms and attribute-aware transformers (Dong et al. 2023).

Despite their progress, existing studies still face critical limitations when addressing the inherent challenges of ASFR: *(1) Coarse-grained localization vs. fine-grained requirements.* Region-level attentions (Figure 1, middle row) are geometrically divided units based on a regular grid,

*is the corresponding author.

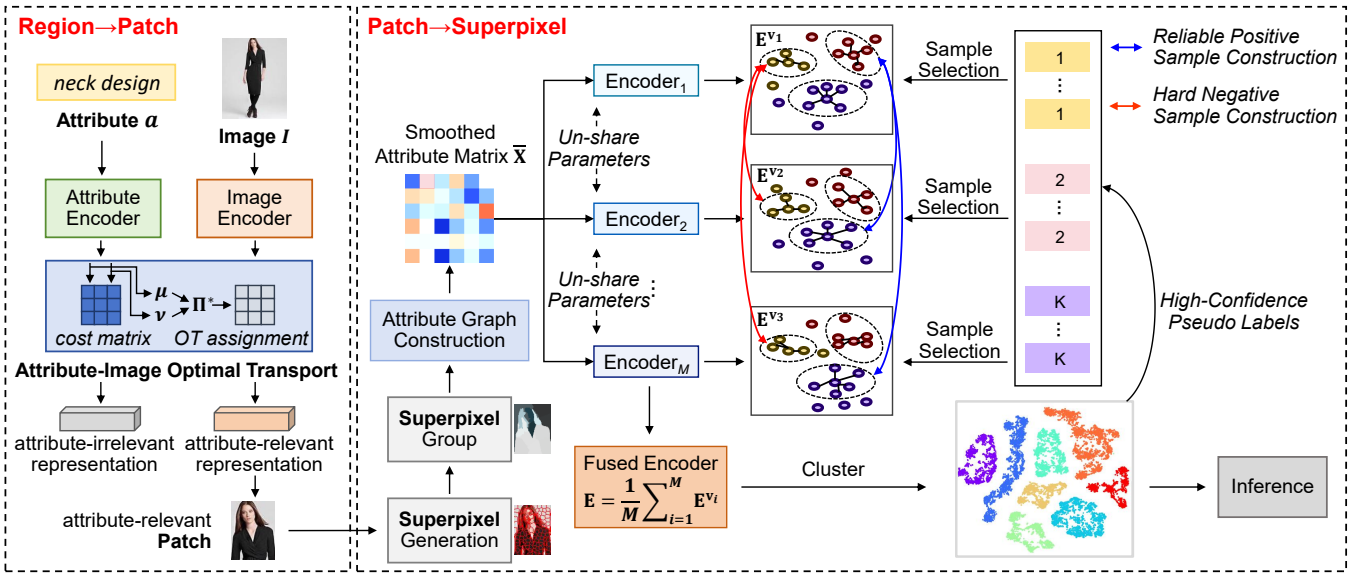

Figure 2: An overview of the proposed framework *Pro*Fashion.

which may fragment semantically coherent attributes across multiple units, disrupting attribute integrity. *(2) Imbalanced global-local perception.* The attention mechanism computes query–key correlations independently, which produces locally optimal weights and limits the ability to aggregate global information, thereby inducing local bias. As a result, existing attention mechanisms often emphasize salient but irrelevant regions while overlooking attributes distributed across the image. As shown in Figure 1 (*sleeve length*), such focus misalignment impairs the perception of globally coherent attributes, including *fabric* and *style*, which require the integration of both global context and local details.

Inspired by the hierarchical visual processing of the human perceptual system, we propose *Pro*Fashion, a progressive framework for pixel-level fashion attribute perception that integrates optimal transport, superpixel grouping, and neural semantic aggregation. Unlike conventional region-level attention mechanisms, our optimal transport module not only models query–key correlations but also imposes uniform distribution constraints on both queries and keys, enabling a comprehensive consideration of global relationships and effectively mitigating local bias. In parallel, the superpixel grouping and clustering modules capture localized attribute-level semantics to support fine-grained inference. Consequently, *Pro*Fashion achieves a balanced integration of global and local perception, effectively addressing the limitations of traditional attention mechanisms and leading to significant performance improvements in ASFR.

In summary, We make three key contributions:

- We propose a hierarchical localization framework (region → patch → superpixel) that aligns visual patterns and attribute semantics across multiple granularities. By modeling pixel-level microstructures, our method resolves feature-overlap ambiguity and achieves fine-grained accuracy. To the best of our knowledge, this is the first work addressing ASFR at superpixel granularity.
- We formulate attribute localization as a global optimization problem via optimal transport theory, minimizing Wasserstein distance to preserve key regions while suppressing interference. Further, pixel-level semantic clustering disentangles overlapping features by modeling local dependencies, ensuring precise attribute separation.
- Comprehensive experiments conducted on the FashionAI, DeepFashion and DARN datasets validate the efficacy of our framework, yielding overall MAP improvements of **3.11%**, **3.70%**, and **3.49%**, respectively. Otherwise, the framework achieves a **26.94%**, **22.22%** and **24.78%** relative average throughput on FashionAI, DeepFashion, and DARN datasets, respectively.

Related Work

Attribute-Specific Fashion Retrieval

In recent years, the ASFR task has garnered significant attention from academia and industry (Han et al. 2023; Yan et al. 2022a; Jiao et al. 2023). Early work focused on capturing attribute-aware visual features using attention mechanisms. As a seminal work in the ASFR field, CSNs (Veit, Belongie, and Karaletsos 2017) used fixed masks to select attribute-specific embedding dimensions for fine-grained similarity measurement. ASEN (Ma et al. 2020) advanced this by introducing Attribute-aware Spatial Attention (ASA) and Attribute-aware Channel Attention (ACA), jointly learning multiple attribute-specific embeddings end-to-end. Subsequent works built on ASA and ACA, including Yan et al.'s hierarchical attribute embedding (Yan et al. 2021) and Wan et al.'s parallel ASA and ACA application (Wan et al. 2024). ISLN (Yan et al. 2022b) enhanced attribute-guided image retrieval with iterative similarity learning, while AttnFashion (Wan et al. 2022) mapped attributes to image regions and semantics with adaptive feature fusion. MODC employs

a deep learning-based online clustering method that explicitly optimizes clustering by jointly leveraging instance-level and cluster-level triplet losses (Jiao et al. 2022). However, region-based methods often introduce irrelevant background noise due to coarse segmentation, limiting precise attribute feature capture. Patch-based methods, such as Dong et al.'s repeated ASA and ACA application (Dong et al. 2021) and RPF's transformer-based refinement (Dong et al. 2023), reduced noise but struggled with fixed patch segmentation, failing to adapt to varying attribute shapes and scales. Recent contrastive learning approaches with weak geometric distortion constraints (Xiao and Yamasaki 2025) or relational knowledge distillation (Xiao and Yamasaki 2024) have improved performance when integrated into existing models.

Despite these advances, two challenges persist: region-level features introduce significant noise, and fixed patch-level features cannot adapt to diverse attribute shapes and scales. To address these, we propose a superpixel-based method leveraging superpixel segmentation for semantically consistent regions. This approach reduces noise and captures attribute-relevant visual features with higher precision, offering a finer-grained solution for ASFR.

Superpixel Generation

Superpixels are clusters of homogeneous pixels defined by features like brightness, color, or texture (Barcelos et al. 2024). They serve as a key over-segmentation tool in computer vision, grouping pixels into perceptually meaningful regions (Kim, Park, and Shim 2023). Early methods like SLIC (Achanta et al. 2012) are widely used for their simplicity and efficiency. Recent advances leverage deep learning, with CNNs and transformer-based architectures enabling adaptive superpixel generation for applications like semantic segmentation and object detection (Zhu et al. 2023; Shen et al. 2016; Yu, Yang, and Liu 2021; Liu et al. 2021b). Unsupervised and weakly supervised approaches reduce reliance on labeled data (Kwak, Hong, and Han 2017), while lightweight models optimize efficiency for real-time processing on resource-constrained devices (Shang et al. 2020; Gendy, He, and Sabor 2023).

Superpixels may enhance fine-grained feature extraction by capturing localized semantic regions (e.g., *texture*), improving accuracy through context-aware segmentation while preserving structural details efficiently. To our knowledge, superpixels have not yet been applied to the ASFR task.

Methodology

Overview of *Pro*Fashion Framework

The *Pro*Fashion framework is as shown in Figure 2. Specifically, this framework consists of two modules: **R**egion→**P**atch (RP) and **P**atch→**S**uperpixel (PS). The RP module extracts attribute-relevant visual region features through attribute-image optimal transport, and then crops the image based on these features to obtain visual feature patches. The PS module processes these visual feature blocks through superpixel generation and grouping, obtaining attribute-aware superpixel features. These features are then utilized to construct attribute graphs and enhance clustering via contrastive learning, ultimately producing attribute centroids for subsequent retrieval tasks.

Region → Patch

We propose to formulate the visual attribute feature region localization as an **O**ptimal **T**ransport (OT) (Villani 2009; Sinkhorn and Knopp 1967) problem to align attribute distributions with visual patterns to extract attribute-relevant regions in images. This approach enables precise localization of attribute-specific features from a globally optimal perspective, avoiding the loss of key attribute features. More precisely, our objective is to effectively locate and learn attribute-relevant features from those image regions, rather than representing the entire image indiscriminately.

To localize attribute-relevant regions in an input attribute-image pair (a, I), we compute an assignment matrix encoding correlation scores between the attribute and visual features. The correlation quantifies their similarity. The image I is processed by a visual encoder to produce a feature map. Concurrently, the attribute a is converted into an embedding vector $\mathbf{f}_a \in \mathbb{R}^{l_m}$ via an attribute embedding module. The image's one-dimensional patch sequence feature is reshaped into a two-dimensional spatial feature map, $\mathbf{f}_I \in \mathbb{R}^{l_m \times h \times w}$, using a 1×1 convolutional layer to preserve spatial structure and semantic richness. A common strategy for computing the correlation map between attributes and images is to use cosine similarity as follows:

$$\mathbf{C}_{cor} = \frac{\mathbf{f}_a \mathbf{f}_I}{\|\mathbf{f}_a\| \|\mathbf{f}_I\|} \in \mathbb{R}^{h \times w}. \tag{1}$$

We adopt a global perspective to maximize total correlation, deriving the optimal assignment matrix $\mathbf{\Pi}^*$ rather than performing individual matchings. Our OT method computes query–key correlations while enforcing uniform distribution constraints across queries and keys, allowing comprehensive consideration of global weights and mitigating local bias. Superpixel grouping and clustering capture local attribute-level semantics for fine-grained inference. To avoid trivial solutions, we represent source and target feature maps with uniform empirical distributions $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, and constrain the marginals of $\mathbf{\Pi}$ to match them. The cost matrix is defined as $\mathbf{C} = \mathbf{1} - \mathbf{C}_{cor}$. We further employ entropy regularization (Cuturi 2013) to accelerate computation via smoothed optimization, formulated as:

$$\begin{aligned} \mathbf{\Pi}^* = \underset{\mathbf{\Pi} \in \mathbb{R}_+^{h \times w}}{\arg\min} \sum_{i=1}^{h} \sum_{j=1}^{w} \mathbf{C}_{ij} \mathbf{\Pi}_{ij} - \lambda H(\mathbf{\Pi}), \\ \text{s.t.} \quad \mathbf{\Pi} \mathbf{1}_w = \boldsymbol{\mu}, \quad \mathbf{\Pi}^\top \mathbf{1}_h = \boldsymbol{\nu}. \end{aligned} \tag{2}$$

In this way, our method can simultaneously achieve both global and local perception, overcoming the limitations of traditional attention mechanisms. After identifying attribute-relevant features in the image, we crop the features $\mathbf{f}_I$ by thresholding to retain high-weight regions, obtaining $\mathbf{f}_{\text{patch}}$ for the PS module.

Patch → Superpixel

Superpixel Generation Although $\mathbf{f}_{\text{patch}}$ captures attribute-relevant visual features, it includes some irrelevant elements.

Method	MAP for each attribute								Overall MAP
	skirt length	sleeve length	coat length	pant length	collar design	lapel design	neckline design	neck design	
CSN (Veit, Belongie, and Karaletsos 2017)	61.97	45.06	47.30	62.85	69.83	54.14	46.56	54.47	53.52
ASEN (Ma et al. 2020)	64.44	54.63	51.27	63.53	70.79	65.36	59.50	58.67	61.02
HAEN (Yan et al. 2021)	64.13	55.52	56.41	72.31	73.32	69.22	62.41	59.80	64.13
ASEN++ (Dong et al. 2021)	66.34	57.53	55.51	68.77	72.94	66.95	66.81	67.01	64.31
AttnFashion (Wan et al. 2022)	65.70	56.46	54.64	71.12	74.45	69.36	65.69	65.54	65.37
ISLN (Yan et al. 2022b)	65.91	58.83	56.45	71.22	74.53	70.55	65.71	65.61	66.10
RPF (Dong et al. 2023)	66.75	67.84	59.59	73.14	75.72	73.18	<u>74.40</u>	74.98	70.10
ASEN_V2+PKD (Xiao and Yamasaki 2024)	69.28	62.13	59.72	73.08	<u>80.11</u>	74.08	68.98	70.04	68.48
ASEN_V2+PT+PKD (Xiao and Yamasaki 2024)	68.94	62.13	60.88	73.56	78.20	<u>77.77</u>	69.94	69.32	69.14
ASEN+GeoDCL (Xiao and Yamasaki 2025)	65.20	53.95	50.42	67.10	76.32	70.47	64.60	67.55	62.81
ASEN_V2+GeoDCL(Xiao and Yamasaki 2025)	68.71	59.18	55.54	70.72	77.14	73.03	68.49	69.25	66.48
RPF+GeoDCL (Xiao and Yamasaki 2025)	<u>69.96</u>	<u>68.70</u>	<u>61.05</u>	<u>73.96</u>	78.34	77.19	70.72	**80.01**	<u>71.15</u>
***Pro*Fashion**	**70.73**	**70.01**	**64.45**	**76.69**	**80.86**	**78.57**	**78.01**	<u>78.40</u>	**74.26**

Table 1: Comparative evaluation (%) on FashionAI dataset. *Top: Prior SOTA baselines; Middle: SOTA with knowledge distillation enhancements; Bottom: Ours. Best in bold, second underlined.*

To address this and enhance pixel-level microstructure analysis, inspired by the concept of pixel density (Ni et al. 2006; Nishiwaki et al. 2013; Zhang et al. 2024), the pixel density of pixel p is defined as:

$$\rho_p = \sum\nolimits_{q \in \mathcal{N}_p} d_s(p,q) d_c(p,q), \tag{3}$$

where $\mathcal{N}_p$ denotes the neighboring pixels around p, and $d(p,q)$ measures the similarity between pixels p and q. D_s and D_c represent the Euclidean distances in the spatial and color feature spaces, respectively, and are mapped to $[0,1]$ via a Gaussian kernel to generate d_s and d_c as follows:

$$d_s(p,q) = \exp\left(-\frac{D_s(p,q)^2}{2\delta_s^2}\right), \tag{4}$$

$$d_c(p,q) = \exp\left(-\frac{D_c(p,q)^2}{2\delta_c^2}\right), \tag{5}$$

$$d(p,q) = d_c(p,q) + \alpha d_s(p,q). \tag{6}$$

To enhance computational efficiency, we set $\mathcal{N}_p$ at a smaller size. Pixel density values are computed using Eq. (3). Initial candidate pixels are randomly sampled and iteratively moved to high-density positions within a search region (e.g., 3×3). Superpixels are generated based on Eq. (6) and cluster centers for semantic partitioning. Using the OT assignment matrix $\mathbf{\Pi}^*$ from Eq. (2), we calculate mean correlations between superpixels and attributes to select highly correlated superpixels for constructing the attribute graph. This integrates OT with superpixel generation, optimizing visual semantic partitioning and visual-textual semantic alignment for fine-grained attribute-relevant visual features $\mathbf{f}_{\text{su-pixel}}$.

Attribute-relevant Visual Feature Clustering Considering that the same attribute pattern can manifest in various forms, this paper constructs attribute graph and employs clustering to identify the core representations for subsequent prediction and retrieval tasks. We construct an attribute graph $\mathcal{G} = (\mathbf{X}, \mathbf{A})$ using the attributes of the fine-grained superpixel $\mathbf{f}_{\text{su-pixel}}$ in the image. Let $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ represent a set of N nodes, each associated with m attribute values, and let $\mathcal{E}$ denote the set of edges. Each node v_i corresponds to a fine-grained superpixel feature $I_{\text{su-pixel}}$, with its attributes reflecting the attribute values. The edges between nodes indicate whether two superpixel images share common attribute values, while the weight w_j represents the cosine similarity of their attribute-relevant features. The attribute matrix is denoted as $\mathbf{X} \in \mathbb{R}^{N \times D}$, and the original adjacency matrix is denoted as $\mathbf{A} \in \mathbb{R}^{N \times N}$, and the degree matrix of $\mathbf{A}$ is denoted as $\mathbf{D} \in \mathbb{R}^{N \times N}$. The non-symmetric normalized Laplacian matrix $\tilde{\mathbf{X}} = (\mathbf{I} + \mathbf{D}^{-1}\mathbf{A})^t \mathbf{X}$ is used to process the matrix $\mathbf{X}$, where t represents the number of layers of the Laplacian filter and $\widetilde{\mathbf{X}}$ is the smoothed attribute matrix. Inspired by (Liu et al. 2023), we embed nodes into a latent space through M parameter-untied MLP encoders $\mathbf{E}^{v_i}{}_{i=1}^{M}$, where each encoder processes the attribute matrix $\tilde{\mathbf{X}}$ and outputs ℓ^2-normalized representations $\mathbf{E}^{v_i}(\tilde{\mathbf{X}})$. This design constructs multiple semantic views without relying on data augmentation, effectively mitigating semantic drift in graph representation learning.

To enhance sample discriminative capability, we use high-confidence clustering in a contrastive learning. We average outputs from multiple encoders, denoted $\mathbf{E}(\tilde{\mathbf{X}})$, and apply K-means clustering to obtain K clusters, with centroids $\mathbf{C}_p$ used for prediction. High-confidence samples are selected based on confidence score $\text{conf}_j = \exp\left(-\|\mathbf{E}_j - \mathbf{C}_p\|^2\right)$, taking the top τ samples. These samples form K disjoint clusters across M views. Reliable positive samples are encodings from the same cluster across views, while different clusters provide hard negative samples. The training objective includes positive sample loss:

$$\mathcal{L}_p = \frac{\sum_{p=1}^{K} \sum_{i=1}^{M} \sum_{j=1, i \neq j}^{M} \left\| \mathbf{C}_p^i - \mathbf{C}_p^j \right\|_2^2}{2KM(M-1)}, \tag{7}$$

Method	MAP for each attribute					Overall MAP
	texture	fabric	shape	part	style	
CSN (Veit, Belongie, and Karaletsos 2017)	14.09	6.39	11.07	5.13	3.49	8.01
ASEN (Ma et al. 2020)	15.01	7.32	13.32	6.27	3.85	9.14
AttnFashion (Wan et al. 2022)	12.90	6.34	11.38	5.24	4.20	8.01
ASEN++ (Dong et al. 2021)	15.60	7.67	14.31	6.60	4.07	9.64
RPF (Dong et al. 2023)	15.62	8.30	15.02	7.38	4.77	10.22
ASEN+GeoDCL (Xiao and Yamasaki 2025)	16.09	7.84	12.80	6.27	5.25	9.41
ASEN_V2+GeoDCL (Xiao and Yamasaki 2025)	15.29	7.11	11.77	5.52	3.76	8.68
RPF+GeoDCL (Xiao and Yamasaki 2025)	16.69	8.95	15.47	8.02	5.19	10.80
***Pro*Fashion**	**20.51**	**11.98**	**19.45**	**9.06**	**8.12**	**14.50**

Table 2: Comparative evaluation (%) on DeepFashion dataset.

Method	MAP for each attribute									Overall MAP
	clothes category	clothes button	clothes color	clothes length	clothes pattern	clothes shape	collar shape	sleeve length	sleeve shape	
CSN (Veit, Belongie, and Karaletsos 2017)	34.10	44.32	47.38	53.68	54.09	56.32	31.82	78.05	58.76	50.86
ASEN (Ma et al. 2020)	36.69	46.96	51.35	56.47	54.49	60.02	34.18	80.11	60.04	53.31
HAEN (Yan et al. 2021)	32.10	47.04	45.03	48.27	49.92	51.22	28.05	78.29	58.47	48.70
AttnFashion (Wan et al. 2022)	34.94	48.56	48.14	54.47	52.65	56.36	32.32	82.63	60.77	52.32
ISLN (Yan et al. 2022b)	38.84	51.26	52.67	56.55	53.85	58.34	36.64	82.74	61.28	54.68
ASEN++ (Dong et al. 2021)	40.15	50.42	53.78	60.38	57.39	59.88	37.65	83.91	60.70	55.94
RPF (Dong et al. 2023)	44.60	55.30	54.02	63.85	56.91	60.15	38.70	84.57	59.35	56.88
***Pro*Fashion**	**49.61**	**59.20**	**58.75**	**69.58**	**59.16**	**66.08**	**40.16**	**87.90**	**63.28**	**60.37**

Table 3: Comparative evaluation (%) on DARN dataset.

and negative sample loss:

$$\mathcal{L}_n = \frac{\sum_{p=1}^{K}\sum_{q=1,p\neq q}^{K}\sum_{i=1}^{M}\sum_{j=1}^{M}\left\|\mathbf{C}_p^i - \mathbf{C}_q^j\right\|_2^2}{2KM(K-1)(M-1)}. \quad (8)$$

The total loss is:

$$\mathcal{L} = \mathcal{L}_p + \beta\mathcal{L}_n, \quad (9)$$

where minimizing $\mathcal{L}_p$ enhances intra-cluster similarity and $\mathcal{L}_n$ ensures inter-cluster separation, reducing false negatives.

Experiment

Experimental Setup

Datasets Following established practices in the field (Ma et al. 2020; Dong et al. 2021, 2023; Jiao et al. 2023), we evaluate *Pro*Fashion's performance on three standard benchmark datasets: *FashionAI* (Zou et al. 2019), *DeepFashion* (Liu et al. 2016), and *DARN* (Huang et al. 2015). The dataset partitioning and preprocessing procedures remain consistent with prior works. Notably, while images in *DeepFashion* may have multiple attribute annotations, *DARN* and *FashionAI* label only a single attribute per image.

Baseline Models We select several representative baseline methods for comparison: CSN (Veit, Belongie, and Karaletsos 2017), ASEN (Ma et al. 2020), HAEN (Yan et al. 2021), ISLN (Yan et al. 2022b), ASEN++ (Dong et al. 2021), AttnFashion (Wan et al. 2022), and RPF (Dong et al. 2023). Additionally, we include knowledge distillation-enhanced approaches (GeoDCL (Xiao and Yamasaki 2025) and PKD (Xiao and Yamasaki 2024)), all of which represent the current State-Of-The-Art (SOTA) in this domain.

Implementation Details To ensure fair evaluation and compatibility with baseline models, we report the mean average precision (MAP) for each attribute and the overall MAP across all datasets. For our model architecture, we use a CLIP (ViT-B/32) as the image encoder for the RP module and a ViT-B/16 network pre-trained on ImageNet for the PS module. Our training strategy follows a two-stage procedure, as outlined in (Dong et al. 2021): (1) Initial learning rate of $1e-4$, decaying by a factor of 0.3 every 3 epochs, for a total of 50 epochs. (2) Learning rate reduced to $1e-5$, decaying by a factor of 0.95 per epoch, for a total of 50 epochs. The optimal model is selected according to the average overall MAP of three datasets on their validation sets by grid-search with three trials. Hyperparameters are set as follows: $\mathcal{N}_p = (7, 7)$ (Eq. (3)), $\delta_s = 60$ (Eq. (4)), $\delta_c = 150$ (Eq. (5)), $\alpha = 0.1$ (Eq. (6)), $\beta = 0.5$ (Eq. (9)), $M = 3$ (Eqs. (7)-(8)), and $\tau = 0.5$.

Performance Comparison

Comparative Analysis of Basic Performance Tables 1-3 show *Pro*Fashion outperforms all baselines, achieving Overall MAP improvements of **3.11%**, **3.70%**, and **3.49%** over prior SOTA (RPF+GeoDCL/RPF) on FashionAI, DeepFashion, and DARN datasets, respectively. We observe that *Pro*Fashion achieves consistent performance improvements across all three datasets, demonstrating the effectiveness of the proposed progressive fine-grained feature penetration framework in handling both global fine-grained attributes and scenarios with overlapping attribute characteristics. To further illustrate the strengths of *Pro*Fashion, we

Method	DARN → FashionAI			Overall MAP
	sleeve length	coat length	neckline design	
ASEN (Ma et al. 2020)	29.36	25.08	16.86	23.35
ASEN++ (Dong et al. 2021)	30.56	26.08	17.26	24.31
RPF (Dong et al. 2023)	34.93	27.96	20.89	26.09
***Pro*Fashion**	<u>70.01</u>	<u>64.45</u>	<u>78.01</u>	<u>70.05</u>
	39.62	**31.06**	**24.92**	**30.09**
Method	**FashionAI → DARN**			**Overall MAP**
	sleeve length	clothes length	collar shape	
ASEN (Ma et al. 2020)	65.63	43.67	24.08	37.46
ASEN++ (Dong et al. 2021)	65.68	44.35	24.08	38.05
RPF (Dong et al. 2023)	66.14	44.87	23.62	38.81
***Pro*Fashion**	<u>87.90</u>	<u>69.58</u>	<u>40.16</u>	<u>66.29</u>
	68.78	**48.35**	**28.21**	**41.95**

Table 4: Cross-dataset evaluation performance. S → T denotes the setting of training on S dataset and evaluation on T dataset. Within-dataset test results are underscored.

Method	MAP for each attribute					Overall MAP
	texture	fabric	shape	part	style	
w/o RP	15.17	9.92	15.66	7.08	4.96	11.05
w/o PS	14.02	8.68	14.86	6.40	4.45	10.44
w/o $\mathcal{L}_p$	18.46	10.64	17.08	8.91	7.49	12.55
w/o $\mathcal{L}_n$	17.80	10.10	16.76	8.86	6.08	12.97
w/o $\mathcal{L}_p, \mathcal{L}_n$	17.26	9.71	16.52	8.50	6.76	12.60
***Pro*Fashion**	**20.51**	**11.98**	**19.45**	**9.06**	**8.12**	**14.50**

Table 5: Ablation studies on DeepFashion dataset for the contribution of *Pro*Fashion's different components.

Method	FashionAI	DeepFashion	DARN
OT → Attention	72.24	10.36	57.42
CLIP → ResNet	72.75	12.58	59.16
***Pro*Fashion (CLIP and OT)**	**74.26**	**14.50**	**60.37**

Table 6: Ablation studies on three datasets for different RP key component choices (overall MAP).

analyze its performance on individual attributes. Notably, *Pro*Fashion substantial relative improvements of **34.26%** compared with RPF+GeoDCL on DeepFashion dataset, particularly in attributes such as *texture* (**+22.89%**), *fabric* (**+33.85%**), and *style* (**+54.67%**). This strong performance, especially in granular attributes like *texture*, *fabric*, and *style*, suggests that *Pro*Fashion's underlying architecture and learning mechanisms are particularly adept at capturing fine-grained visual cues and complex semantic relationships crucial for distinguishing these detailed fashion characteristics. These results highlight its superior global fine-grained attribute pattern extraction capability. More importantly, our framework also demonstrates relative improvements in distinguishing attributes that primarily focus on local features (e.g., *collar shape* 9.85% on Deepfashion, *neck design* 9.95% on FashionAI), indicating that it can effectively handle both global and local fine-grained attributes.

Generalization on Out-of-domain Data We investigate *Pro*Fashion's generalization ability through cross-dataset evaluation. As shown in Table 4, *Pro*Fashion achieves consistent improvements over baseline models across all attributes in bidirectional cross-dataset evaluations (DARN ↔ FashionAI). Notably, it surpasses RPF by **4.00%** in overall MAP for DARN → FashionAI and **3.14%** for the reverse transfer (FashionAI → DARN), demonstrating robust knowledge transfer capabilities. However, label granularity differences may lead to performance degradation: Overall MAP drops by **39.96%** (DARN→FashionAI) and **24.34%** (FashionAI→DARN) over to within-dataset evaluation.

Ablation Study

Effect of Key Component Table 5 shows the performance variation when different framework components are systematically removed. Removing any module leads to significant performance degradation, with the PS module having the most substantial impact. Without the PS module, the model cannot capture microstructural attribute patterns or identify overlapping features, resulting in the largest performance drop. This confirms that both RP and PS modules are essential, with their synergy driving the framework's effectiveness. The loss function analysis in Table 5 reveals that removing both $\mathcal{L}_p$ and $\mathcal{L}_n$ from Eq. (9) yields the lowest MAP scores. In contrast, combining $\mathcal{L}_n$ and $\mathcal{L}_p$ achieves optimal performance. This demonstrates that cluster refinement using high-confidence samples consistently improves model effectiveness.

Choice of Key Component Table 6 presents an ablation study examining the core components of the RP modules. We systematically substitute CLIP (ViT-B/32) and OT with ResNet and Attention mechanisms respectively, while maintaining identical experimental conditions. The results demonstrate that backbone encoder selection significantly impacts performance. **The larger encoders yield performance gains.** CLIP-based encoding consistently outperforms ResNet across all three datasets by 1.51%, 1.92%, and 1.21% respectively. More critically, replacing OT with attention mechanisms causes substantial performance drops of 2.02%, **4.14%**, and 2.95%. This degradation occurs because attention mechanisms primarily capture localized correlations while neglecting the global feature interactions essential for modeling complex attribute relationships. In contrast, OT's global perspective preserves semantically important features that attention-based approaches would otherwise lose.

Effect of Training Data Size We evaluate *Pro*Fashion's performance across varying proportions of training data from the DeepFashion dataset. Specifically, we randomly selected 20%, 50%, and 80% of the training set for model training and then assessed its performance on the test set. As shown in the Figure 3 (a), the model's performance improves as the size of the annotated dataset increases. Moreover, compared to prior SOTA RPF, *Pro*Fashion exhibits lower sensitivity to the annotated dataset.

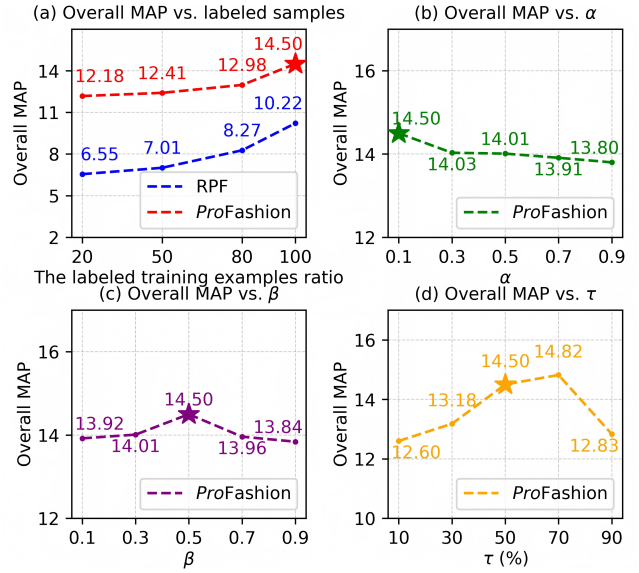

Figure 3: Performance under varying training sizes and hyperparameter sensitivity analysis on DeepFashion dataset.

Time Efficiency Analysis

Compared to RPF, *Pro*Fashion (ResNet) achieves a **26.94%**, **22.22%** and **24.78%** relative average throughput on FashionAI, DeepFashion, and DARN datasets, *Pro*Fashion (CLIP) shows a relative decrease in throughput of 0.56%, 8.36%, and 6.92%, respectively. This is due to the increased inference time caused by the CLIP encoder, but it also brings a greater improvement. CLIP-based encoding consistently outperforms ResNet across all three datasets by 1.51%, 1.92%, and 1.21% respectively.

Hyperparameter Sensitivity Analysis

We assess the effects of hyperparameters α, β, and τ on the DeepFashion dataset, with results shown in Figure 3 (b)-(d). Our analysis shows that α and β are largely insensitive within the 0.1–0.9 range, while τ is highly sensitive. The MAP reaches its peak at $\tau = 50\%$ and drops when τ deviates from this value, likely due to overly confident pseudo-labels causing confirmation bias or an insufficient number of positive samples limiting the network's discriminative ability. Overall, most hyperparameters exhibit low sensitivity and do not constrain the method's applicability.

Visualization Analysis

Figure 4 shows visualization examples in the context of fashion attribute analysis. The first row of images displays the original clothing images, each representing a different fashion attribute. The second row shows the attention heatmaps corresponding to the attributes displayed in the first row. These heatmaps provide insights into the image regions the model focuses on when making attribute predictions, although some attributes may exhibit bias or introduce irrelevant noise. The third row displays the superpixel semantic segmentation results, which align more closely with attribute-relevant regions.

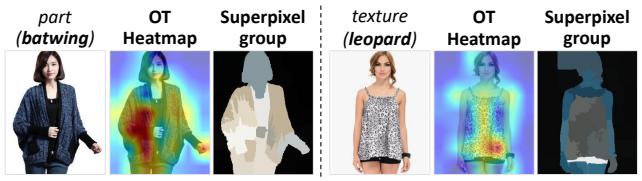

Figure 4: Visualization examples of attribute-relevant visual features at different granularities.

Figure 5: Case examples of retrieval results on the FashionAI dataset. Incorrectly retrieved results are indicated by red borders. Blue text shows the attribute values of the input image, while red text denotes the actual attribute values of the incorrectly retrieved results.

Case Study

We present example cases of retrieval given a query image and attribute, showing the results of *Pro*Fashion and the baseline model RPF (Dong et al. 2023), respectively. The top 5 images retrieved from the dataset that match the query attribute are displayed. Figure 5 demonstrates that *Pro*Fashion can distinguish fine-grained differences in local attributes, such as accurately differentiating between *deep V-neck* and *V-neck* in the neckline design row. In contrast, RPF returns incorrect results. These results highlight the superior capability of *Pro*Fashion in handling complex attribute-specific fashion retrieval tasks compared to the baseline model.

Conclusion

In ASFR tasks, patch-level mechanisms introduce coarse-grained localization and local bias, motivating our design. To address these challenges, we propose *Pro*Fashion, a progressive framework inspired by the human brain's processing of attribute patterns. Our key innovation lies in extracting attribute-related features at the superpixel level, representing the first approach of its kind. Acting like a microscopic lens, *Pro*Fashion employs optimal transport and neural semantic aggregation to achieve pixel-level attribute awareness, enabling precise modeling of fine-grained image structures. Comprehensive experiments on FashionAI, DeepFashion, and DARN highlight the framework's accuracy and computational efficiency, achieving MAP improvements of 3.11%, 3.70%, and 3.49%, with relative average throughput of 26.94%, 22.22%, and 24.78%, respectively.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62406319, 62572465) and the Youth Innovation Promotion Association of CAS (No.2021153).

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11): 2274–2282.

Barcelos, I. B.; Belém, F. D. C.; João, L. D. M.; Patrocínio Jr, Z. K. D.; Falcão, A. X.; and Guimarães, S. J. F. 2024. A Comprehensive Review and New Taxonomy on Superpixel Segmentation. *ACM Computing Surveys*, 56(8): 1–39.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. *Advances in Neural Information Processing Systems*, 26.

Dong, J.; Ma, Z.; Mao, X.; Yang, X.; He, Y.; Hong, R.; and Ji, S. 2021. Fine-Grained Fashion Similarity Prediction by Attribute-Specific Embedding Learning. *IEEE Transactions on Image Processing*, 30: 8410–8425.

Dong, J.; Peng, X.; Ma, Z.; Liu, D.; Qu, X.; Yang, X.; Zhu, J.; and 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*, 1273–1282.

Gendy, G.; He, G.; and Sabor, N. 2023. Lightweight Image Super-Resolution Based on Deep Learning: State-of-the-Art and Future Directions. *Information Fusion*, 94: 284–310.

Han, Y.; Zhang, L.; Chen, Q.; Chen, Z.; Li, Z.; Yang, J.; and Cao, Z. 2023. FashionSAP: Symbols and Attributes Prompt for Fine-Grained Fashion Vision-Language Pre-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15028–15038.

Huang, J.; Feris, R. S.; Chen, Q.; and Yan, S. 2015. Cross-Domain Image Retrieval with a Dual Attribute-Aware Ranking Network. In *Proceedings of the IEEE International Conference on Computer Vision*, 1062–1070.

Jiao, Y.; Gao, Y.; Meng, J.; Shang, J.; and Sun, Y. 2023. Learning Attribute and Class-Specific Representation Duet for Fine-Grained Fashion Analysis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11050–11059.

Jiao, Y.; Xie, N.; Gao, Y.; Wang, C.-c.; and Sun, Y. 2022. Fine-Grained Fashion Representation Learning by Online Deep Clustering. In *Uropean Conference on Computer Vision*, 19–35.

Kim, S.; Park, D.; and Shim, B. 2023. Semantic-Aware Superpixel for Weakly Supervised Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1142–1150. AAAI Press.

Kwak, S.; Hong, S.; and Han, B. 2017. Weakly Supervised Semantic Segmentation Using Superpixel Pooling Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 4111–4117.

Liu, A.-A.; Zhang, T.; Song, D.; Li, W.; and Zhou, M. 2021a. FRSFN: A Semantic Fusion Network for Practical Fashion Retrieval. *Multimedia Tools and Applications*, 80: 17169–17181.

Liu, Y.; Yang, X.; Zhou, S.; Liu, X.; Wang, S.; Liang, K.; Tu, W.; and Li, L. 2023. Simple Contrastive Graph Clustering. *IEEE Transactions on Neural Networks and Learning Systems*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021b. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and 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*, 1096–1104.

Ma, H.; Zhao, H.; Lin, Z.; Kale, A.; Wang, Z.; Yu, T.; Gu, J.; Choudhary, S.; and Xie, X. 2022. EI-CLIP: Entity-Aware Interventional Contrastive Learning for E-Commerce Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18051–18061.

Ma, Z.; Dong, J.; Long, Z.; Zhang, Y.; He, Y.; Xue, H.; and Ji, S. 2020. Fine-Grained Fashion Similarity Learning by Attribute-Specific Embedding Network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11741–11748.

Ni, T.; Schmidt, G. S.; Staadt, O. G.; Livingston, M. A.; Ball, R.; and May, R. 2006. A Survey of Large High-Resolution Display Technologies, Techniques, and Applications. In *Proceedings of the IEEE Virtual Reality Conference*, 223–236. IEEE.

Nishiwaki, S.; Nakamura, T.; Hiramoto, M.; Fujii, T.; and Suzuki, M.-a. 2013. Efficient colour splitters for high-pixel-density image sensors. *Nature Photonics*, 7(3): 240–246.

Shang, R.; Zhang, J.; Jiao, L.; Li, Y.; Marturi, N.; and Stolkin, R. 2020. Multi-Scale Adaptive Feature Fusion Network for Semantic Segmentation in Remote Sensing Images. *Remote Sensing*, 12(5): 872.

Shen, J.; Hao, X.; Liang, Z.; Liu, Y.; Wang, W.; and Shao, L. 2016. Real-Time Superpixel Segmentation by DBSCAN Clustering Algorithm. *IEEE Transactions on Image Processing*, 25(12): 5933–5942.

Sinkhorn, R.; and Knopp, P. 1967. Concerning Nonnegative Matrices and Doubly Stochastic Matrices. *Pacific Journal of Mathematics*, 21(2): 343–348.

Song, C. H.; and Han, H. J. 2022. Convolutional Attribute Mask with Two-Step Attention for Fashion Image Retrieval. In *Proceedings of the 2022 26th International Conference on Pattern Recognition*, 2093–2099. IEEE.

Tran, S.; Du, M.; Chanda, S.; Manmatha, R.; and Taylor, C. J. 2019. Searching for Apparel Products from Images in the Wild. In *Proceedings of the KDD 2019 Workshop on AI for Fashion*.

Veit, A.; Belongie, S.; and Karaletsos, T. 2017. Conditional Similarity Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 1781–1789.

Villani, C. 2009. *Optimal Transport: Old and New*, volume 338. Springer.

Wan, Y.; Yan, K.; Yan, C.; and Zhang, B. 2022. Learning Attribute-guided Fashion Similarity with Spatial and Channel Attention. *J. Exp. Theor. Artif. Intell.*, 36(5): 703–719.

Wan, Y.; Yan, K.; Yan, C.; and Zhang, B. 2024. Learning Attribute-Guided Fashion Similarity with Spatial and Channel Attention. *Journal of Experimental & Theoretical Artificial Intelligence*, 36(5): 703–719.

Xiao, L.; and Yamasaki, T. 2024. Boosting Fine-grained Fashion Retrieval with Relational Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8229–8234. IEEE.

Xiao, L.; and Yamasaki, T. 2025. GeoDCL: Weak Geometrical Distortion Based Contrastive Learning for Fine-Grained Fashion Image Retrieval. *IEEE Transactions on Artificial Intelligence*, 6(3): 1234–1245.

Yan, C.; Ding, A.; Zhang, Y.; and Wang, Z. 2021. Learning Fashion Similarity Based on Hierarchical Attribute Embedding. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics*, 1–8. IEEE.

Yan, C.; Yan, K.; Zhang, Y.; Wan, Y.; and Zhu, D. 2022a. Attribute-Guided Fashion Image Retrieval by Iterative Similarity Learning. In *2022 IEEE International Conference on Multimedia and Expo*, 1–6.

Yan, C.; Yan, K.; Zhang, Y.; Wan, Y.; and Zhu, D. 2022b. Attribute-Guided Fashion Image Retrieval by Iterative Similarity Learning. In *2022 IEEE International Conference on Multimedia and Expo*, 1–6.

Yu, Y.; Yang, Y.; and Liu, K. 2021. Edge-Aware Superpixel Segmentation with Unsupervised Convolutional Neural Networks. In *Proceedings of the 2021 IEEE International Conference on Image Processing*, 1504–1508. IEEE.

Zhang, Z.; Hu, W.; Lao, Y.; He, T.; and Zhao, H. 2024. Pixel-GS: Density Control with Pixel-Aware Gradient for 3D Gaussian Splatting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 326–342.

Zhu, A. Z.; Mei, J.; Qiao, S.; Yan, H.; Zhu, Y.; Chen, L.-C.; and Kretzschmar, H. 2023. Superpixel Transformers for Efficient Semantic Segmentation. In *Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 7651–7658.

Zou, X.; Kong, X.; Wong, W.; Wang, C.; Liu, Y.; and Cao, Y. 2019. FashionAI: A Hierarchical Dataset for Fashion Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.