Learning Attribute and Class-Specific Representation Duet for Fine-grained Fashion Analysis

Yang Jiao
Amazon
jaoyan@amazon.com

Yan Gao, Jingjing Meng, Jin Shang, Yi Sun
Amazon
{yanngao, ajmeng, imjshang, yisun}@amazon.com

Abstract

Fashion representation learning involves the analysis and understanding of various visual elements at different granularities and the interactions among them. Existing works often learn fine-grained fashion representations at the attribute level without considering their relationships and inter-dependencies across different classes. In this work, we propose to learn an attribute and class- specific fashion representation duet to better model such attribute relationships and inter-dependencies by leveraging prior knowledge about the taxonomy of fashion attributes and classes. Through two sub-networks for the attributes and classes, respectively, our proposed an embedding network progressively learns and refines the visual representation of a fashion image to improve its robustness for fashion retrieval. A multi-granularity loss consisting of attribute-level and class-level losses is proposed to introduce appropriate inductive bias to learn across different granularities of the fashion representations. Experimental results on three benchmark datasets demonstrate the effectiveness of our method, which outperforms the state-of-the-art methods by a large margin.

1. Introduction

Fashion products have become one of the most consumed products in online shopping. Unlike other types of products, fashion products are usually rich in visual elements at different levels of granularity. For instance, besides the overall visual appearance, a fashion product can be described by a set of *attributes*, such as "shape", "color" and "style", which focus on different aspects of the visual representation. Each attribute can be further categorized into various *classes*. For example, "fit", "flare" and "pencil" are different classes under attribute "shape" (Fig. 1). Therefore, modeling fashion representation in different granularities is essential for online shopping and other downstream applications, especially those that require analysis

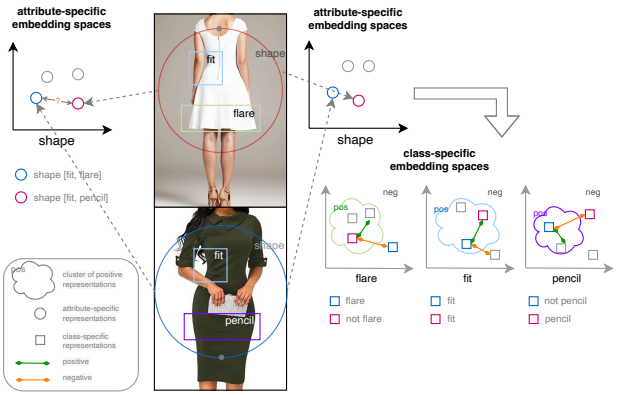

Figure 1. Left: existing fine-grained representation learning methods often learn *attribute*-specific representations for fashion products, thus may not be able to discern the two dresses that have different compositions of visual elements at the *class* level.Right: our proposed method (right) jointly learns attribute and class-specific representations. Therefore, it can discriminate between the two dresses by their class-specific representations.

of subtle or fine-grained details such as attribute-based fashion manipulation [1, 2, 27] and retrieval [6, 14, 19, 23, 24], fashion copyright [6, 19], and fashion compatibility analysis [11, 15, 21, 23].

Fine-grained fashion modeling and analysis in recent years explore the attribute-specific representation learning. The focus has recently shifted from earlier works that learn separate representations for each attribute independently [1, 2] to multi-task learning, which uses a common backbone for different attributes while tailoring the learning for each specific attribute via mechanisms such as attention masks [6, 14, 19, 24]. Success of these attribute-specific representation learning methods for fine-grained fashion analysis can be attributed to their capabilities to discriminate visual features associated with different aspects of fashion products, which learning an image-level global representation finds challenging.

However, when it comes to *classes*, such attribute-

specific representation methods face a similar challenge to the above. The reason is that due to the dynamic and aesthetic nature of fashion products, different visual elements are often composited together to achieve certain visual effects, making an attribute-level description insufficient to capture such interactions and granularity. For instance, under the same "shape" attribute, one may go for a dress design that combines *classes* "fit" and "flare" for a more casual look (top image, Fig. 1), but go for a different dress that combines "fit" and "pencil" for a more formal look while flattering one's natural curves (bottom image, Fig. 1). Therefore, an attribute-level representation is hard to differentiate the two dresses. Alternatively, one may directly learn a class-specific representation for each *class* under the "shape" attribute, which, however, faces the scalability issue. For instance, if a fashion image is associated with N attributes and M classes per attribute, one would need to learn $N \times M$ class-specific representations.

To better discriminate fashion products with distinct design considerations and model the interplay among various visual elements, we propose to leverage prior knowledge about fashion taxonomy to model fashion products. We jointly learn both attribute-specific and class-specific fashion representations through a multi-attribute multi-granularity multi-label embedding network (M3-Net). M3-Net consists of two sub-networks, for attributes and classes, respectively. Different attributes share the same backbone sub-network as well as two attribute-conditional attention modules, while different classes under a given attribute share two class-conditional attention modules.

The shared backbone and conditional attention modules allow the network to better capture the inter-dependencies and shared visual statistics among the attributes and classes. Through multi-label learning on attribute-specific representations, we also improve the scalability of the proposed network by focusing class-specific representation learning on high likelihood classes only. Finally, a multi-granularity loss consisting of attribute-level and class-level losses is designed to introduce appropriate inductive bias for learning across different granularities.

In summary, our contributions are:

- We propose to model fashion products at both attribute and class levels based on fashion taxonomy to better capture the inter-dependencies of various visual elements and improve the discriminative power of learned fashion representations.

- We design a multi-attribute multi-granularity multi-label network (M3-Net) to jointly learn attribute-specific and class-specific representation duet for fine-grained fashion analysis. Through two sub-networks and conditional attention modules, M3-Net is able to progressively learn discriminative representations at different granularities, with appropriate inductive bias introduced by the attribute-level and class-level losses.

- Our model outperforms state-of-the-art methods in fine-grained fashion retrieval on three benchmark datasets. The experimental results demonstrate the efficacy of our proposed method.

2. Related Work

2.1. Generic Fashion Representation Learning

Earlier fashion representation learning works [7, 8, 16, 20, 25, 26] focus on the global representation of a fashion product by learning a generic metric embedding from the entire fashion image. The generic representations benefit tasks such as in-shop fashion retrieval [16, 22, 26], street-to-shop fashion retrieval [4, 7, 8, 13, 17, 18] and compatibility retrieval [11, 15, 20, 25]. For in-shop fashion retrieval, the images often have a consistent background and photo shooting angle. In comparison, street-to-shop retrieval is more challenging because the images are often taken in an uncontrolled environment with varying lighting conditions, scales, and viewing angles. Different from the above two tasks that focus on the overall similarity, compatibility retrieval focuses on a specific global attribute such as color, fabric, and style. Although effective in global representation learning, these works lack the capabilities to model fine-grain details and subtleties in fashion products.

2.2. Multi-attribute Representation Learning

Many works tackle the problem of fine-grained fashion representation learning by analyzing fashion attributes. We have seen great success of such approaches in tasks such as attribute-specific retrieval [2, 6, 19, 24] and retrieval with attribute manipulation [1, 12, 27]. One group of works [12, 27] utilizes fully connected layers to transform generic representations into attribute-specific representations. However, the linear transformation function of fully connected layers neglects the spatial relationship within attribute-specific representations. Another group of works learns attribute-specific representations by leveraging region proposal, either via a dedicated network [13] or via global pooling layers [1, 2, 9]. For example, some works [1, 2] localize the spatial area of each attribute using global pooling layers, and crop the spatial feature maps for further attribute-specific learning. Although cropping spatial feature maps allows representation learning to focus on a local region, it rigidly limits the visual representation to a specific area and ignores other correlated visual elements in a larger area. In contrast, the third group of works uses attention masks to incorporate a global view into representation learning with the flexibility to bring in contexts from other regions. For instance, the authors of [6, 14, 19, 24] utilize

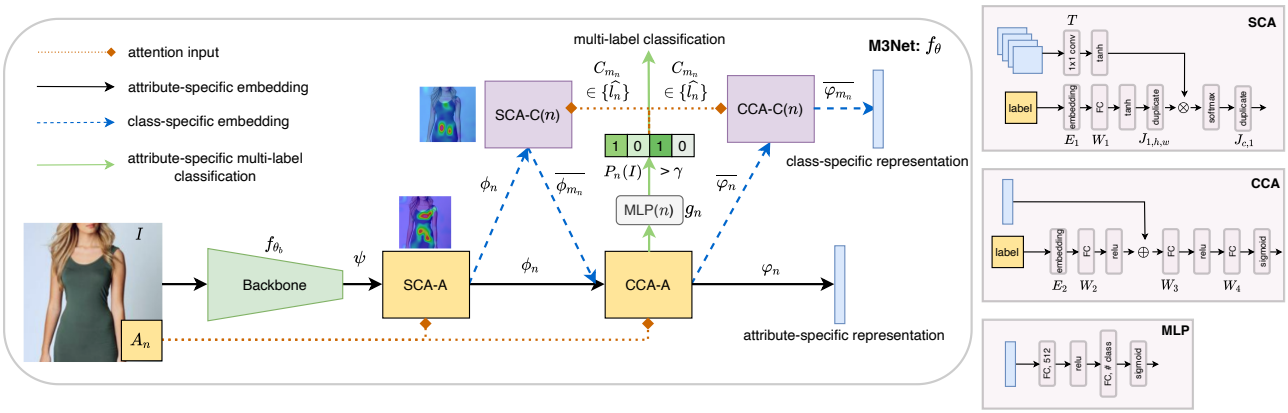

Figure 2. The architecture of the proposed M3-Net. M3-Net first obtains a generic representation using a backbone CNN. It learns attribute-specific representations through two attribute-conditional attention modules (SCA-A and CCA-A). Taking the high-likelihood classes from the multi-label classification, M3-Net employs two class-conditional attention modules (SCA-C and CCA-C) to learn the class-specific representations. SCA-A and SCA-C are extended from the SCA module (Alg. 1), and CCA-A and CCA-C are extended from CCA module (Alg. 2).

attention masks to dynamically assign weights to different dimensions of the global representation of an image for specific attributes. Veit, Belongie and Karaletsos [24] use attention masks to select and reweight relevant dimensions for each attribute to induce attribute-specific subspaces. Instead of learning attribute-specific weights, other works [6, 19] propose to learn attribute-aware spatial and channel attention modules. The attention modules are attached to the feature extraction network to enhance the participation of attributes in representation learning.

2.3. Multi-granularity Representation Learning

As attribute-level representations may still fall short of fashion tasks that require analysis of finer granular interactions, such as the one shown in Fig. 1, a few works propose to learn fashion representations in multiple granularities. Some works tackle the multi-granularity representation learning problem from the spatial domain, which essentially transforms multi-granularity learning into multi-scale learning. For instance, Dong, Ma and their co-authors [6, 19] propose to use a global branch and a local branch to learn two attribute-specific representations on two scales. Similarly, Bao, Zhang and their co-authors [3] propose a feature learning network that jointly learns the representation in two feature map scales and three image scales via a global, a part-base, and a local branch. Instead of multi-scale learning, Jiao, Xie and their co-authors [14] propose to segment the attribute-specific embedding spaces into class-specific embedding spaces using the cluster prototypes learned by online deep clustering. The proposed model achieves state-of-the-art performance on the fine-grained fashion retrieval task by prioritizing retrieval in class-specific embedding spaces. However, the representations are not optimized in class-specific embedding spaces because the segmentation happens in the inference stage. In our work, we propose to jointly learn both attribute-specific and class-specific representations through a multi-granularity embedding network.

3. Proposed Method

The proposed multi-attribute multi-granularity multi-label embedding network (M3-Net) is an end-to-end network that jointly learns the attribute and class level representations. As shown in Figure 2, M3-Net employs a backbone network, two attribute-conditional attention modules, two class-conditional attention modules, and a multi-label classification module. The backbone network shares learned weights across all attributes, which makes the embedding network scalable. It embeds an input image into a generic representation that represents the entire image. The generic representation is then fed through two attribute-conditional attention modules to focus learning on fine-grained attributes and obtain attribute-specific representations. Two class-conditional attention modules are further applied to learn more fine-grained class-specific representations. The shared backbone and conditional attention modules allow the network to better capture the inter-dependencies and shared visual statistics among the attributes and classes. The multi-label classification module serves to improve the scalability of the proposed network by focusing class-specific representation learning on high-likelihood classes.

3.1. M3-Net Architecture

Given a set of fashion product images $\{I\}$, we denote the set of attributes associated with these images as $\{A_n\}$, where $n \in [1, N]$. Similarly, we denote the set of classes

associated with a given attribute A_n as $\{C_{m_n}\}$, where $m_n \in [1, M_n]$ and M_n is the number of classes under attribute A_n. Given A_n, an image $I \in \{I\}$ can associate with a subset of labels in $\{1, 2, ..., M_n\}$, which is represented by a vector $[y_{1_n}, ...y_{m_n}, ..., y_{M_n}]$, where $y_{m_n} = 1$ if and only if image I is associated with class C_{m_n}, and 0 otherwise.

Generic representation. Denote the parameters of M3-Net as θ and the parameters of the backbone network as θ_b. Correspondingly, f_θ represents M3-Net and f_{θ_b} represents the backbone network. The generic visual representation of an image I is then denoted by $\psi = f_{\theta_b}(I)$, where $\psi \in \mathbb{R}^{c \times h \times w}$, c, h, w are the number of channels, height, and width, respectively.

From here, we will use ψ to denote a generic feature map, use ϕ to denote a spatially attended feature map and φ to denote a channel-wise attended feature map. Meanwhile, we use x to denote the attribute-specific representation, and $\overline{x}$ to denote the class-specific representation.

Attribute-specific representation. The attribute-specific representation learning acts as a connector between the generic representations and the class-specific representations. It refines the generic representation by focusing learning on the attribute-specific visual features, and the resulting attribute-specific representation will be further refined in subsequent modules to obtain more fine-grained class-specific visual representations. To reduce the complexity of the class-specific representation learning, we conduct the attribute-specific representation learning in a multi-label setting. Therefore, instead of learning $N + \sum_n^N M_n$ representations (N attribute-specific and $\sum_n^N M_n$ class-specific representations), we can exclude unlikely classes, thus improve scalability.

Motivated by the effectiveness of spatial and channel-wise attention for multi-attribute fashion representation learning in prior works, we utilize Spatial Conditional Attention module (SCA) and Channel-wise Conditional attention module (CCA) akin to the attribute-aware attention in [19]. Both SCA (Algorithm 1) and CCA (Algorithm 2) are applied to both attribute-specific and class-specific representation learning. The structures of SCA and CCA are shown in Figure 2.

Attribute-conditional attention modules help representation learning focus on spatial locations and dimensions relevant to a given attribute. The Spatial Conditional Attention on Attribute (SCA-A) applies SCA (Algorithm 1) at the attribute level, while the Channel-wise Conditional Attention on Attribute (CCA-A) applies CCA (Algorithm 2) at the attribute level. Given the generic representation ψ and an attribute A_n, SCA-A takes them as inputs and transforms them into feature maps with an identical size, $p_1(\psi), p_2(A_n) \in \mathbb{R}^{c' \times h \times w}$ (step 5, 6 in Algorithm 1). The attention map α_n is the Hadamard product between the image feature $p_1(\psi)$ and attribute feature $p_2(A_n)$ with a scaling factor $\sqrt{c'}$ (step 7). Then the spatially attended feature map is generated as

Algorithm 1 Spatial Conditional Attention (SCA)

1: **Input**: a general feature map $\psi \in \mathbb{R}^{c \times h \times w}$, targeted label l, intermediate channel number c'
2: **Output**: conditional spatial attention $\alpha_l \in \mathbb{R}^{c \times h \times w}$
3: **Define**: feature transform function $T, T(\psi) \in \mathbb{R}^{c' \times h \times w}$, label embedding function $E_1, E_1(l) \in \mathbb{R}^{c' \times 1}$, linear function $W_1, W_1 \in \mathbb{R}^{c' \times c'}$
4: **Define**: Hadamard product $\odot$, $J_{1,h,w} \in \mathbb{R}^{1 \times h \times w}$ and $J_{c,1} \in \mathbb{R}^{c \times 1}$ are all-ones matrix for spatial duplication
5: Transform the input feature: $p_1(\psi) = tanh(T(\psi)), p_1(\psi) \in \mathbb{R}^{c' \times h \times w}$
6: Embedding and transform the input label: $p_2(l) = tanh(W_1 E_1(l)) \cdot J_{1,h,w}$, $p_2(l) \in \mathbb{R}^{c' \times h \times w}$
7: Calculate spatial attention: $\alpha_l = J_{c,1} \cdot softmax \frac{\sum_i^{c'} [p_2(l) \odot p_1(\psi)]_i}{\sqrt{c'}}$

Algorithm 2 Channel-wise Conditional Attention (CCA)

1: **Input**: a general feature map $\psi \in \mathbb{R}^{c \times h \times w}$, targeted label l, intermediate channel number c'
2: **Output**: conditional channel attention $\beta_l \in \mathbb{R}^{c \times 1}$
3: **Define**: label embedding function $E_2, E_2(l) \in \mathbb{R}^{c' \times 1}$, embedding transform function $W_2, W_2 \in \mathbb{R}^{c' \times c'}$, linear functions $W_3, W_4, W_3 \in \mathbb{R}^{c' \times (c + c')}$, and $W_4 \in \mathbb{R}^{c \times c'}$
4: **Define**: feature vector concatenation $[a; b]$
5: Transform the input feature to vector: $q_1(\psi) = \sum_j^{h \times m} p(\psi), q_1(\psi) \in \mathbb{R}^{c \times 1}$
6: Embedding and transform the input label: $q_2(l) = relu(W_2 E_2(l)), q_2(l) \in \mathbb{R}^{c' \times 1}$
7: Calculate channel attention: $\beta_l = sigmoid(W_4(relu(W_3[q_1(\psi); q_2(l)])))$

$$\phi_n = \alpha_n \odot \psi, \phi_n \in \mathbb{R}^{c \times h \times w}, \tag{1}$$

where $\odot$ is Hadamard product. To further focus on the attribute-relevant dimensions, the Channel-wise Conditional Attention on Attribute, CCA-A (Algorithm 2), takes the spatially attended feature maps ϕ_n and attribute A_n as inputs, and transforms them into feature vectors $q_1(\phi_n)$ and $q_2(A_n)$ (step 5, 6 in Algorithm 2), where $q_1(\phi_n) \in \mathbb{R}^{c \times 1}, q_2(A_n) \in \mathbb{R}^{c' \times 1}$. The attention output β_n is obtained from the concatenation of $q_1(\phi_n)$ and $q_2(A_n)$ (step 7). $\beta_n \in \mathbb{R}^{c \times 1}$. Finally, the attribute-specific representation is generated as

$$\varphi_n = \beta_n \odot q_1(\phi_n), \varphi_n \in \mathbb{R}^{c \times 1}. \tag{2}$$

Subsequently, to reduce the number of class-specific representations in learning, we involve the multi-label classification to exclude low-likelihood classes. As Eq. 3 shows, given an attribute-specific MLP module g_n, the probability

of all classes for an attribute-specific representation on A_n is $P_n(I)$.

$$P_n(I) = [x_{1_n}, ..., x_{m_n}, ..., x_{M_n}] = g_n(\varphi_n). \quad (3)$$

By setting a threshold γ, we obtain the multi-label class prediction $\{\widehat{l_n}\} = \{C_{m_n} | \forall x_{m_n} > \gamma\}$. $\{\widehat{l_n}\}$ is the set of predicted high-likelihood classes. Empirically, with $\gamma = 0.8$, we can exclude on average 94% of classes for an image. Please see the sensitivity analysis of γ in Suppl.

Class-specific representation. As an attribute often involves multiple classes, the class-conditional attention modules help M3-Net focus on individual classes and learn class-specific representations.

Class-conditional attention modules consist of Spatial Conditional Attention on Class (SCA-C) and Channel-wise Conditional Attention on Class (CCA-C). SCA-C and CCA-C are attribute-specific. Different from the attribute-conditional attention modules, we expect them to focus on the relevant spatial locations and dimensions of corresponding classes.

Given an attribute A_n, its spatially attended feature map ϕ_n, and class $C_{m_n} \in \{\widehat{l_n}\}$, the SCA-C generates a class-conditional spatial attention map $\overline{\alpha_{m_n}}$. We obtain the spatially attended feature map $\overline{\phi_{m_n}}$ as

$$\overline{\phi_{m_n}} = \overline{\alpha_{m_n}} \odot \phi_n, \overline{\phi_{m_n}} \in \mathbb{R}^{c \times h \times w}. \quad (4)$$

$\overline{\phi_{m_n}}$ is subsequently fed through both attribute-conditional channel attention (CCA-A) and class-conditional channel attention (CCA-C). Especially, the CCA-A of A_n outputs the attribute-conditional channel attention map $\overline{\beta_n}$. The attended feature is

$$\overline{\varphi_n} = \overline{\beta_n} \odot q_1(\overline{\phi_{m_n}}), \overline{\varphi_n} \in \mathbb{R}^{c \times 1}. \quad (5)$$

The CCA-C of A_n takes the feature vector $\overline{\varphi_n}$ and the class C_{m_n} as inputs to generate the feature vector $q_1(\overline{\varphi_n})$ and gives the class-conditional channel-wise attention map $\overline{\beta_{m_n}}$. The class-specific representation is obtained as

$$\overline{\varphi_{m_n}} = \overline{\beta_{m_n}} \odot q_1(\overline{\varphi_n}), \overline{\varphi_{m_n}} \in \mathbb{R}^{c \times 1}. \quad (6)$$

M3-Net has three key outputs: the attribute-specific representation φ_n, class-specific-representation $\overline{\varphi_{m_n}}$, and multi-label probability vector $P_n(I)$. They are used to constrain the multi-granularity embedding spaces via a multi-granularity objective. In this part, we build our method upon the online clustering method and prototypical triplet loss in [14] and further extend them to learning two-granularity fine-grained representations.

3.2. Multi-granularity Objective

To learn across different granularities of the fashion representations in an end-to-end manner, we design two losses to introduce appropriate inductive bias: an attribute-level loss and a class-level loss.

Attribute-level loss. At the attribute level, we define the below multi-label classification loss to allow the representation learning on multi-label attributes. Given an image I with N attributes $\{A_n\}$, each with M_n classes $\{C_{m_n}\}$ and corresponding class labels $\{y_{m_n}\}$, the loss of multi-label classification is a binary cross-entropy loss,

$$\mathcal{L}_{\mathcal{M}}(I, A_n | y_{m_n}) = \frac{1}{M_n}(\sum_{m=1}^{M_n}[-w_p y_{m_n} \cdot \log x_{m_n} + (1 - y_{m_n}) \cdot \log(1 - x_{m_n})]), \forall x_{m_n} \in P_n(I), \quad (7)$$

where x_{m_n} is the predicted probability of I in class C_{m_n}, w_p is the weight on the positive samples to mitigate the class-imbalance problem.

Class-level loss. At the class level, we propose to regularize the class-specific representation learning on both global and local structures via two triplet losses. For the global structure, we construct a prototypical triplet loss between an image representation, the representation of the positive prototype of the class that the image belongs to, and a negative representation. The triplet loss associated with the local structure involves instance-level representations to refine the local distance.

A classic triplet loss between an anchor, a positive, and a negative representation is defined as

$$\mathcal{L}_{\triangle}(I, I^+, I^-) = max\{0, \zeta + d(I, I^+) - d(I, I^-)\}, \quad (8)$$

where $\zeta = 0.4$ is a predefined margin, and d is the cosine similarity.

Given a triplet of images $[I, I^+, I^-]$ associated with classes $[C_{m_n}, C^+_{m_n}, C^-_{m_n}]$ in attribute A_n, $C_{m_n} = C^+_{m_n} \neq C^-_{m_n}$. The prototypical triplet loss in the class-specific embedding spaces is defined as

$$\mathcal{L}_{\mathcal{CC}}(I, A_n, C_{m_n}) = \mathcal{L}_{\triangle}(\overline{\varphi_{m_n}(I)}, \overline{\varphi_{C^+}}, \overline{\varphi_{m_n}(I^-)}), \quad (9)$$

where $\overline{\varphi_{m_n}(I)}$ is the anchor representation, $\overline{\varphi_{m_n}(I^-)}$ is a random negative representation in the class-specific embedding space, and $\overline{\varphi_{C^+}}$ is the positive class prototype in the space. The computation of $\overline{\varphi_{C^+}}$ is akin to [14].

The instance triplet loss in the class-specific embedding spaces is defined as

$$\mathcal{L}_{\mathcal{CI}}(I, A_n, C_{m_n}) = \mathcal{L}_{\triangle}(\overline{\varphi_{m_n}(I)}, \overline{\varphi_{m_n}(I^+)}, \overline{\varphi_{m_n}(I^-)}), \quad (10)$$

where $\overline{\varphi_{m_n}(I)}$, $\overline{\varphi_{m_n}(I^+)}$, $\overline{\varphi_{m_n}(I^-)}$ are the class-specific representations for the anchor, positive, and negative sample, respectively.

Final objective function. The final objective function combines both attribute and class-level objectives and allows a simple end-to-end training, as shown in Eq.(11)

$$\min_{\theta}(\lambda_M \mathcal{L}_{\mathcal{M}} + \lambda_{CC} \mathcal{L}_{\mathcal{CC}} + \lambda_{CI} \mathcal{L}_{\mathcal{CI}}). \quad (11)$$

Dataset	Attr type	# Attr	# Class per attr	Train/val/test
DeepFashion [18]	multi-label	5	156-230	220K/28K/28K
FashionAI [28]	single-label	8	5-10	144k/18k/18k
DARN [13]	single-label	9	7-55	163k/20k/20k

Table 1. Summary of the datasets used in experimental validations. Attr: attribute.

where, λ_M, λ_{CC} and λ_{CI} are hyperparameters, which are set to 1 in our experiments.

4. Experiments

4.1. Datasets

In Table 1, we summarize the three benchmark datasets used in the experiments. DeepFashion is a large dataset containing image labels of fashion attributes, landmarks, etc. We use the coarsely-annotated attribute prediction subset, which is one of the most popular datasets in fashion retrieval, and is a multi-label dataset.

4.2. Experimental Settings

We compare our proposed method with the state-of-the-art solutions [6, 14, 19] on the aforementioned datasets.

Baselines. ASEN networks are the state-of-the-art works on attribute-specific representation learning. ASEN [19] learns 1024 dimensional attribute-specific representations by using an attribute-aware spatial attention and an attribute-aware channel attention. ASEN$_{v2}$ [19] builds a new attention module structure and achieves better performance than ASEN. ASEN++ [6] further proposes a cascade network with a global branch and a local branch to learn multi-scale representations in 2,048 dimensions.

MODC [14] it is a fine-grained representation learning framework with online deep clustering to achieve retrieval in two granularities. It splits the attribute-specific embedding spaces into class-specific embedding spaces in the inference stage to enable retrieval at the finer granularity. MODC learns attribute-specific representations with 2048 dimensions and achieves the state-of-the-art results on multiple datasets.

M3-Net is our proposed method which jointly learns attribute-specific and class-specific representation duet for fine-grained fashion analysis. The attribute-specific and class-specific representations are 1024 dimensions. To isolate the effects of two-granularity attentions, we trained M3-Net with only the attribute-conditional attention modules, with only the class-conditional attention modules, and with both, called M3-Net$_a$, M3-Net$_c$, and M3-Net respectively.

Training details M3-Net employs a ResNet50 [10] pre-trained on ImageNet [5] as the shared backbone network, and removes the last residual block. It is identical to the backbone of the baselines for fair comparisons. To train M3-Net, we use a learning rate 1×10^{-4} with a 0.975 decay per epoch and a batch size of 16. We sample 40k triplet of images each epoch of training. In training, We set w_p to 100 for DeepFashion, and 1 for FashionAI and DARN. To train the baselines on DeepFashion, We follow the descriptions in [6, 14, 19].

Evaluation Tasks and Metrics. We evaluate our proposed method in comparison with the above baselines on the fine-grained fashion retrieval task. Following the existing protocol for multi-granularity retrieval as proposed in MODC [14], the retrieval is prioritized in the class-specific embedding space, followed by retrieval in the attribute-specific embedding space. Same as existing works [6, 14, 19], we employ Mean Average Precision (MAP) and Recall as the evaluation metrics.

4.3. Experimental Results

In this section, we discuss the experimental results and ablation study of M3-Net on multi-label attributes and single-label attributes. Table 2 presents the overall performance and the performance on each attribute of baselines and M3-Net on DeepFashion. Table 3 and Table 4 summarizes the performance on FashionAI and DARN. On all datasets, we show the ablation study of separately employing attribute-conditional attention modules (i.e., M3-Net$_a$) and class-conditional attention modules (i.e., M3-Net$_c$) to demonstrate the effectiveness of representation learning on attribute granularity and class granularity.

4.3.1 Quantitative evaluation on multi-label dataset

Table 2 summarizes the performance of fine-grained fashion retrieval on all attributes in DeepFashion. Note that ASEN$_{v2}$ [19] and ASEN++ [6] report performance on DeepFashion by treating it as a single-label dataset. To evaluate the performance in the multi-label setting, we split DeepFashion following the multi-class labels of each image in the dataset. The train/validation/test split is shown in Table 1. For the retrieval task, the validation set and test set are further split into the query and candidate sets by 1:4.

DeepFashion is an extremely challenging dataset for fine-grained fashion retrieval. Therefore, all methods perform much worse on DeepFashion than on other benchmark datasets Compared with the baselines, the proposed M3-Net consistently achieves the best performance with a large margin on all evaluation metrics on individual attributes and overall. Even when compared with the state-of-the-art multi-granularity method, MODC, the proposed M3-Net shows significant improvements over MODC on MAP@all (70.42%), MAP@100 (57.95%), and Recall@100 (112.5%). The results demonstrate the efficacy of the attribute and class-specific representations learned by M3-Net.

Model	DeepFashion							
	MAP@all					MAP@all	MAP@100	Recall@100
Attribute	texture	fabric	shape	part	style	overall	overall	overall
ASEN [19]	21.03	11.61	14.68	7.81	4.66	12.33	20.60	5.10
ASEN$_{v2}$ [19]	21.86	11.67	14.58	7.93	4.68	12.51	20.55	4.99
ASEN++ [6]	22.20	11.71	14.70	8.15	4.72	12.74	20.79	5.21
MODC [14]	22.26	11.98	14.68	7.96	5.23	12.78	22.21	6.06
M3-Net$_a$	27.37	18.34	22.96	14.66	9.74	19.03	28.14	10.59
M3-Net$_c$	30.09	19.60	25.24	16.43	11.32	20.92	30.58	12.10
M3-Net	**30.79**	**20.20**	**27.11**	**17.28**	**11.61**	**21.78**	**35.08**	**12.88**

Table 2. Performance comparison on all attributes of the multi-label dataset, DeepFashion.

Model	FashionAI										
	MAP@all								MAP@all	MAP@100	Recall@100
Attribute	skirt length	sleeve length	coat length	pant length	collar design	lapel design	neckline design	neck design	overall	overall	overall
ASEN [19]	64.61	49.98	49.75	65.76	70.30	62.86	52.14	56.42	57.37	64.70	22.77
ASEN$_{v2}$ [19]	65.58	54.42	52.03	67.41	71.36	66.76	60.91	59.58	61.13	67.85	24.14
ASEN++ [6]	66.31	57.51	55.43	68.83	72.79	66.85	66.78	67.02	64.27	70.62	25.30
MODC [14]	74.54	67.48	**68.25**	77.69	81.11	76.90	**77.46**	**77.10**	74.32	80.29	30.26
M3-Net$_a$	73.21	69.58	65.27	78.79	80.80	78.05	77.04	73.40	73.93	81.57	30.48
M3-Net$_c$	74.33	65.91	64.27	78.26	82.00	**78.85**	74.80	72.39	72.88	82.58	30.77
M3-Net	**75.27**	**70.04**	67.90	**79.31**	**82.82**	78.58	76.81	75.51	**75.04**	**87.04**	**32.01**

Table 3. Performance comparison on all attributes of the single-label dataset, FashionAI.

Ablation study on M3-Net$_a$, M3-Net$_c$ and M3-Net suggests that employing more fine-grained attention on classes (M3-Net$_c$) performs better than attribute-level attention alone (M3-Net$_a$). Yet combining attribute-conditional attentions and class-conditional attentions works the best (M3-Net). It shows that representations learned at the two granularities are complementary to each other.

Furthermore, we analyze the performance on co-occurring classes on DeepFashion to evaluate the effectiveness of M3-Net in capturing the inter-dependencies between different classes in representation learning. We first calculate the pairwise co-occurrence rates of all classes in the dataset. Two classes are considered co-occurring if both are associated with the same image. The range of co-occurrence rate is [0, 0.03], i.e., the most frequently co-occurring class pairs appear in 3% of the images. We set the cut-off at 0.002 (corresponding to 565 co-occurrences), resulting in 468 class pairs from the 1M class pairs in the dataset. This gives us 122 unique classes. M3-Net achieves a MAP@all at 35.61 on this set of classes, which is 63.50% higher than the average of all classes. We hypothesize that M3-Net is able to incorporate the inter-dependencies between these classes into representation learning, thus performing better on them.

4.3.2 Quantitative evaluation on single-label datasets

For the two single-label datasets, FashionAI and DARN, we follow the same split as in [19]. On both datasets, we observe similar competence of the proposed M3-Net in fine-grained fashion retrieval. On FashionAI, M3-Net achieve the best performance overall on all evaluation metrics (MAP@all, MAP@100, and Recall@100). It also outperforms the baselines on most individual attributes.

Figure 3. Two-granularity attentions learned by M3-Net on DeepFashion. The attribute-conditional attention tends to focus on a wider range of regions in an image, while the class-conditional attention focuses more narrowly on a region relevant to each class. For instance, row 1: class "slit" focuses on the dress slit, "elegant" focuses on the waist; row 2: "contrast trim" focuses on the part with two different colors, "stretch" focuses on the strap; row 3: "draped" focuses on the drape decoration, and "fit" focuses on the waist. Best viewed in color on a computer screen.

On DARN, we again observe a boost on all evaluation metrics. While baselines perform unsatisfactorily on "clothes category" and "collar shape", M3-Net improves them the most along with improvements on other attributes. Overall, M3-Net exceeds the best baseline, MODC, on MAP@all by 16.35%, MAP@100 by 14.59%, and Recall@100 by 18.24%. The results of M3-Net on single-label datasets again demonstrate the effectiveness of the attribute-specific and class-specific representations. Compared with MODC, which obtains class-specific representations by di-

Model	DARN											
	MAP@all									MAP@all	MAP@100	Recall@100
Attribute	clothes category	clothes button	clothes color	clothes length	clothes pattern	clothes shape	collar shape	sleeve length	sleeve shape	overall	overall	overall
ASEN [19]	36.62	46.01	52.76	56.85	54.89	56.85	34.40	79.95	58.08	52.75	58.72	20.26
ASEN$_{v2}$ [19]	37.97	49.24	52.26	59.13	55.32	59.06	36.86	81.54	58.82	54.29	59.66	20.88
ASEN++ [6]	40.21	50.04	53.14	59.83	57.41	59.70	37.45	83.70	60.41	55.78	61.09	21.51
MODC [14]	49.94	60.75	58.79	66.34	62.24	68.41	45.14	87.41	65.32	62.56	72.16	26.76
M3-Net$_a$	59.16	69.94	68.18	72.58	72.87	76.21	60.22	89.36	71.58	71.06	78.12	30.77
M3-Net$_c$	**60.98**	**70.91**	68.52	**74.95**	**74.48**	**79.15**	**63.09**	**90.31**	73.00	**72.79**	81.46	**31.64**
M3-Net	60.54	70.39	**69.52**	73.97	74.40	77.83	61.63	90.03	**73.94**	72.42	**82.69**	31.34

Table 4. Performance comparison on all attributes of the single-label dataset, DARN.

Figure 4. (a) single-label and (b) multi-label fashion retrieval by M3-Net and MODC on DeepFashion. Green: true positive retrieval; red: false positive retrieval. Best viewed in color on a computer screen.

rectly segmenting the attribute-specific representation into class-specific clusters in the same embedding space, M3-Net learns separate representation spaces for attributes and classes and performs better.

Ablation study on FashionAI shows that overall having attentions on two granularities (M3-Net) performs better than attentions on either granularity alone (M3-Net$_a$ and M3-Net$_c$). On DARN, M3-Net's performance is comparable to that of M3-Net$_c$, which only has class-conditional attention modules. We speculate that this is due to discriminating attributes and classes is less challenging on single-label datasets than on multi-label attribute datasets.

4.3.3 Qualitative evaluations

Figure 3 visualizes the two-granularity attentions learned by M3-Net. The attribute-conditional attention tends to focus on all regions that are likely to be associated with the attribute. For example, in the first row of Figure 3, the attention of attribute "part" involves areas covering various *parts* of the dress such as collar, belt area, and hemline (second image in the first row). On the contrary, the class-conditional attention tends to focus on class-specific regions in the image. For instance, class "slit" highlights the dress slit region (third image in the first row). The visualized attention is consistent with our hypothesis.

Figure 4 presents examples of fine-grained fashion retrieval results on single-label and multi-label datasets. In the figure, we compared the results from our proposed M3-Net with those from MODC (the best-performing baseline). Figure 4 (a) shows that on single-label retrieval, M3-Net can better discriminate fine-grained visual features than MODC, leading to more accurate retrieval results on individual classes. Moreover, on multi-label retrieval, M3-Net is able to retrieve images containing multiple class labels. For example, when searching an image on attribute "texture" with both "print" and "tribal" class labels (last example in Figure 4), M3-Net accurately retrieves related images, while MODC retrieves many other printed textures that is not "tribal".

5. Conclusion

We have proposed a multi-attribute multi-granularity multi-label network (M3-Net) for fine-grained fashion analysis. Our proposed architecture learns both attribute and class-level representations for a fashion image through a shared backbone and two sub-networks with attribute and class conditional attention modules. This design, together with a multi-granularity loss, allows the network to effectively learn discriminative representations while capturing the inter-dependencies among various visual elements in different granularities. Our experiments show that the proposed M3-Net sets new state-of-the-art performance on both single-label and multi-label benchmark datasets in the fine-grained fashion retrieval task.

References

[1] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7708–7717, 2018.

[2] Kenan E Ak, Joo Hwee Lim, Jo Yew Tham, and Ashraf A Kassim. Efficient multi-attribute similarity learning towards attribute-based fashion search. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1671–1679. IEEE, 2018.

[3] Chen Bao, Xudong Zhang, Jiazhou Chen, and Yongwei Miao. Mmfl-net: multi-scale and multi-granularity feature learning for cross-domain fashion retrieval. *Multimedia Tools and Applications*, pages 1–33, 2022.

[4] Qiang Chen, Junshi Huang, Rogerio Feris, Lisa M Brown, Jian Dong, and Shuicheng Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5315–5324, 2015.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Jianfeng Dong, Zhe Ma, Xiaofeng Mao, Xun Yang, Yuan He, Richang Hong, and Shouling Ji. Fine-grained fashion similarity prediction by attribute-specific embedding learning. *arXiv preprint arXiv:2104.02429*, 2021.

[7] Bojana Gajic and Ramon Baldrich. Cross-domain fashion image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1869–1871, 2018.

[8] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE international conference on computer vision*, pages 3343–3351, 2015.

[9] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Ruining He, Charles Packer, and Julian McAuley. Learning compatibility across categories for heterogeneous item recommendation. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 937–942. IEEE, 2016.

[12] Yuxin Hou, Eleonora Vig, Michael Donoser, and Loris Bazzani. Learning attribute-driven disentangled representations for interactive fashion retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12147–12157, 2021.

[13] Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE international conference on computer vision*, pages 1062–1070, 2015.

[14] Yang Jiao, Ning Xie, Yan Gao, Chien-Chih Wang, and Yi Sun. Fine-grained fashion representation learning by online deep clustering. In *European Conference on Computer Vision*, pages 19–35. Springer, 2022.

[15] Donghyun Kim, Kuniaki Saito, Samarth Mishra, Stan Sclaroff, Kate Saenko, and Bryan A Plummer. Self-supervised visual attribute learning for fashion compatibility. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1057–1066, 2021.

[16] Furkan Kinli, Baris Ozcan, and Furkan Kirac. Fashion image retrieval with capsule networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[17] Zhanghui Kuang, Yiming Gao, Guanbin Li, Ping Luo, Yimin Chen, Liang Lin, and Wayne Zhang. Fashion retrieval via graph reasoning networks on a similarity pyramid. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3066–3075, 2019.

[18] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[19] Zhe Ma, Jianfeng Dong, Zhongzi Long, Yao Zhang, Yuan He, Hui Xue, and Shouling Ji. Fine-grained fashion similarity learning by attribute-specific embedding network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11741–11748, 2020.

[20] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52, 2015.

[21] Ambareesh Revanur, Vijay Kumar, and Deepthi Sharma. Semi-supervised visual representation learning for fashion compatibility. In *Fifteenth ACM Conference on Recommender Systems*, pages 463–472, 2021.

[22] Devashish Shankar, Sujay Narumanchi, HA Ananya, Pramod Kompalli, and Krishnendu Chaudhury. Deep learning based large scale visual recommendation and search for e-commerce. *arXiv preprint arXiv:1703.02344*, 2017.

[23] Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 390–405, 2018.

[24] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 830–838, 2017.

[25] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4642–4650, 2015.

[26] Zhonghao Wang, Yujun Gu, Ya Zhang, Jun Zhou, and Xiao Gu. Clothing retrieval with visual attention model. In *2017 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–4. IEEE, 2017.

[27] Bo Zhao, Jiashi Feng, Xiao Wu, and Shuicheng Yan. Memory-augmented attribute manipulation networks for interactive fashion search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1520–1528, 2017.

[28] Xingxing Zou, Xiangheng Kong, Waikeung Wong, Congde Wang, Yuguang Liu, and Yang Cao. Fashionai: A hierarchical dataset for fashion understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.