Fine-grained Fashion Representation Learning by Online Deep Clustering

Yang Jiao[1][0000−0002−6390−2517], Ning Xie[1][0000−0002−0116−1426], Yan Gao[1][0000−0002−8012−1392], Chien-chih Wang[1][0000−0001−5127−3939], and Yi Sun[1][0000−0001−6473−9777]

Amazon, Seattle, WA, U.S.A
{jaoyan, xining, yanngao, ccwang, yisun}@amazon.com

Abstract. Fashion designs are rich in visual details associated with various visual attributes at both global and local levels. As a result, effective modeling and analyzing fashion requires fine-grained representations for individual attributes. In this work, we present a deep learning based online clustering method to jointly learn fine-grained fashion representations for all attributes at both instance and cluster level, where the attribute-specific cluster centers are online estimated. Based on the similarity between fine-grained representations and cluster centers, attribute-specific embedding spaces are further segmented into class-specific embedding spaces for fine-grained fashion retrieval. To better regulate the learning process, we design a three-stage learning scheme, to progressively incorporate different supervisions at both instance and cluster level, from both original and augmented data, and with ground-truth and pseudo labels. Experiments on FashionAI and DARN datasets in the retrieval task demonstrated the efficacy of our method compared with competing baselines.

Keywords: fine-grained fashion representation learning, online deep clustering, image retrieval, semi-supervised learning

1 Introduction

The pursuit of fashion is one of the most prominent incentives for consumers. Therefore, modeling and analyzing fashion is an essential step to understand customer preferences and behaviors. In online shopping, it facilitates fashion trend prediction, fashion search, fashion recommendation, fashion compatibility analysis, etc. Many previous works attempt to learn a generic representation [29, 17, 9, 12, 21, 28] to establish a metric embedding for fashions. However, they often fail to capture the subtle details of different fashion styles and hence are not sufficient to support fine-grained downstream applications, such as attribute based fashion manipulation [33, 2, 1] and search [27, 26, 20].

Indeed, fashion designs are rich in visual details associated with a variety of fashion attributes at both global and local levels. Therefore, effective modeling and analyzing fashion necessitates fine-grained representations for individual attributes. A fashion attribute represents a specific aspect of fashion products. An example of a global fashion attribute is "*skirt length*", depicting the overall characteristic of the fashion product. "*Neckline style*", on the other hand, is a local attribute, which reflects the fashion design for a local product area. A naive way of learning such representations is to learn representations on each attribute independently. It is not ideal as it ignores the shared visual statistics among the attributes. A better way is to formulate it as a multi-task learning problem, such that the different fine-grained attribute-specific fashion representations may share a common backbone with a companion computing process to tailor to each specific attribute.

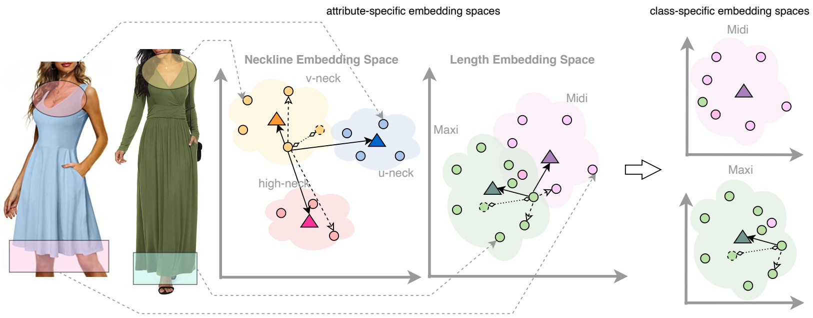

Fig. 1: Fine-grained representation learning by MODC. In each attribute-specific embedding space, a representation is learned by constraining it using the representations of the cluster centers (prototypes), other instances, and the augmented images. Based on the embedding similarity with cluster centers, class-specific embedding spaces for *Midi* and *Maxi* are segmented from *Length*. Solid arrow: the *cluster-level* constraint between instances and prototypes; dashed arrow: the *instance-level* constraint between instances; two-way arrow: the *augmentation constraint* between the two augmented images. More details are discussed in Section 3.

Attribute-Specific Embedding Network (ASEN) [20] and its extension ASEN ++ [8] are the most recent prior state-of-the-art research to learn attribute-specific representations in the fashion domain. They address the problem by two attention modules, *i.e.*, an attribute-aware spatial attention module and an attribute-aware channel attention module, and learn fine-grained representations by a triplet loss, which is widely adopted in fashion related problems [33, 13, 2, 1, 20]. Notwithstanding their demonstrated efficacy, these works regularize multiple attribute embedding spaces with the triplet loss only at the instance level. We speculate the instance level loss is insufficient to learn representations that can well capture a global view of the cluster structure, and the performance can be further improved by constructing an attribute-specific clustering method.

Therefore, we propose a Multi-task Online Deep Clustering (MODC) method to learn fine-grained fashion representations. MODC leverages the generic representation power via multi-task learning, while simultaneously integrating cluster-level constraints with the global structure. In addition to the instance-level triplet loss, we further introduce a cluster-level triplet loss between cluster centers and instances, which strives for an explicit optimization of the global structures of the clusters. We treat cluster centers to be class prototypes, akin to that of prototypical networks [25], and use a memory bank to compute the prototypes. The cluster centers can be further leveraged in the inference stage to segment the fine-grained fashion retrieval space. As shown in Figure 1, retrieval in class-specific embedding spaces prioritizes the positives (positives/negatives are assigned based on the embedding similarity to the given cluster center) compared with that in attribute-specific embedding spaces. Our proposed MODC is able to effectively leverage both labeled and unlabeled data for training, and we design a three-stage learning scheme to progressively guide the learning of the network parameters.

In summary, our contributions are:

- We propose the Multi-task Online Deep Clustering (MODC) method for efficient attribute-specific fine-grained representation learning for fashion products. MODC combines the instant-level loss function and a cluster-level triplet loss function to explicitly optimize the local and global structure of the fine-grained representation clusters in individual attribute-specific embedding spaces, which leads to improved clustering results.
- Using the cluster centers learned via MODC, we further segment the attribute-specific embedding spaces to class-specific embedding spaces to boost the fine-grained fashion retrieval.
- Our experiments on attribute-specific fashion retrieval, including supervised and semi-supervised learning, achieve state-of-the-art results, on the FashionAI [34] and DARN [15] datasets, demonstrating the efficacy of our proposed method.

2 Related Work

2.1 General Fashion Representation Learning

Fashion representation learning is a popular task-driven problem. In recent years, many researchers propose to learn representations by deep convolutional neural networks used for multiple tasks including in-shop fashion retrieval [24, 29, 17], street-to-shop fashion retrieval [12, 15, 5, 19, 9, 18], compatibility search [28, 21, 14, 16], etc. One of the common approaches is to learn general representations [12, 21, 24]. The general fashion representations usually capture the patterns for the entire image, and are commonly used for general fashion image retrieval via conducting k nearest neighbor in the general representation space. Effective general fashion representations benefit tasks with similarity search for entire images.

2.2 Attribute-Specific Fashion Representation Learning

For fine-grained level tasks such as attribute-specific fashion retrieval, the fashion representations are asked to focus on specific attributes rather than the entire image. Therefore, for these tasks, instead of learning general representations, it is natural to involve attributes during the modeling stage. The attribute-specific fashion representation learning is usually formulated as a multi-task learning problem.

Attribute region proposal, classification, and ranking are popular components to involve attributes [15, 13, 33, 2, 1]. Huang *et al.* [15] proposes attribute region by Network-in-network, while [13, 2, 1] use global pooling layers to propose attribute activation maps. After extract attribute-specific representations, [33] further average the representations as the attribute prototypes to store in memory bank and achieve attribute manipulation. Although the aforementioned studies learn attribute-specific representations, region proposal with image cropping may result in losing global view. Furthermore, attribute-specific fully connected layer is the key component of these approaches, which is not scalable because it requires increased parameters with more attributes added.

To learn better representations and handle the scalability issue, another group of studies [27, 20, 8] apply attention masks to learn attribute-specific representations. Attention masks dynamically assign weights to different dimensions of the general representations for specific attributes. For instance, the state-of-the-art methods [20, 8] enhance the participation of attribute in representation learning by attaching attribute-aware spatial and channel attention modules to the feature extraction network. In aforementioned works, attribute classification and attribute-specific triplet ranking is commonly applied.

In fact, most of the existing works focus on handling the fashion representation learning by optimizing the instance representation relationships such as optimizing the relative distance between a triplet instances, while ignoring the global structure and patterns of the attribute-specific embedding spaces. In our work, we propose the Multi-task Online Deep Clustering (MODC) method that models the global structure in the representation learning procedure.

2.3 Unsupervised Fashion Representation Learning

Learning attribute-specific fashion representations demands attribute annotations that is expensive because a fashion item could associate with many attributes. Therefore, unsupervised learning becomes a potential solution to relief this demand. Deep clustering [3], online deep clustering [32, 4], and self-supervision [10, 31, 22, 30, 6, 11, 4, 7] are commonly adopted techniques for unsupervised representation learning. Particularly in the fashion domain, Kim *et al.* [16] define pretexts such as color histogram prediction, patch discrimination, and pattern discrimination to conduct self-supervised learning. However, the representations learned via task-specific pretexts may not be generalizable to other tasks. Revanur *et al.* [23] build a semi-supervised representation learning with item ranking and self-supervision for labeled and unlabeled data. Similar to previous works [27, 20, 8], the method proposed by [23] optimizes the instance representation distribution but ignores the global representation distribution. In our work, the Multi-task Online Deep Clustering can effectively exploits unlabeled data via constructing online clustering with attribute-specific memory bank.

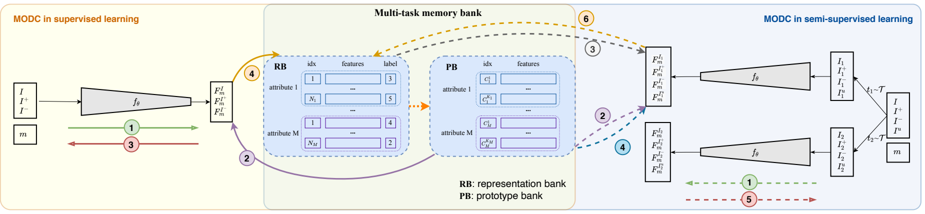

Fig. 2: The overall structure of the proposed *Multi-task Online Deep Clustering (MODC)*, where M tasks are involved, yielding M accounts in the representation bank (RB) and prototype bank (PB). MODC can adopt any multitask network that is denoted by f_θ. The proposed method works for both supervised learning and semi-supervised learning, which leverages two training schemes. The scheme details are shown in Algorithm 1 in Section 4, where the steps (for supervised learning: *step 1 to 4 on the left side*; for semi-supervised learning: *step 1 to 6 on the right side*) are consistent.

3 Multi-task Online Deep Clustering

To address the fine-grained fashion representation learning, each attribute is associated with its own embedding space to yield attribute-specific representations for a given input image. In order to provide an effective and scalable solution to capture visual details, we propose Multi-task Online Deep Clustering (MODC). MODC unifies classes and clusters for all attributes, and forms cluster centers as class prototypes. Furthermore, it learns not only the similarity between instances but also the similarity between instances and all class prototypes to constrain the fine-grained representations in a global manner. MODC is able to adopt unlabeled samples by assigning pseudo-labels and consequently can perform in both supervised and semi-supervised learning. Figure 2 illustrates the structure of MODC, and the training schemes for supervised and semi-supervised learning, which will be elaborated in Section 4.

3.1 Online Process with A Memory Bank

For a given attribute, a prototype is the "averaged" representation of samples belonging to the same class/cluster. Offline clustering methods like [3] commonly update prototypes after every training epoch, which have two drawbacks (1) the representation generated during training is not reused when updating prototypes; (2) prototypes cannot accommodate the latest training outcome so that the loss computation is always based on outdated prototypes. Online clustering, however, addresses the drawbacks by storing and reusing the representations to update cluster centers every n mini-batch. In this way, the training efficiency and computational cost are optimized. In this work, we create a multi-task memory bank in MODC to store the prototypes for each attribute separately.

The multi-task memory bank has two components: a representation bank (RB) and a prototype bank (PB). Each attribute owns an individual "bank account" in RB and PB respectively. The RB accounts store the attribute-specific representations corresponding to each *(sample ID, attribute, sample label)* tuple. The PB accounts store the class prototypes w.r.t. each attribute, which are generated from the RB accounts based on the sample labels on attributes. The structure of the multi-task memory bank is shown in Figure 2.

Let's use $\{I, \{l_1, l_2, ..., l_M\}\}$ to denote a set of images and the corresponding labels of each image for M different attributes. K_m is the total number of classes in attribute m, where $m \in [1, M]$. The attribute-specific representation for an arbitrary image I on a given attribute m is denoted as F_m^I, which is generated via $F_m^I = f_\theta(I, m)$.

During training, when a new representation F_m^I is generated, we retrieve the existing F_m^I from RB by *(sample id, attribute, sample label)* tuple. The F_m^I in RB is then updated by

$$F_m^I \leftarrow \lambda F_m^I + (1-\lambda) f_\theta(I, m), \tag{1}$$

where λ is a momentum coefficient for memory bank updating, and is set as 0.5 in our experiments. The storage limit of each attribute label is $1k$.

PB is updated along with RB. Suppose there are N_m^k representations belong to class k ($k \in K_m$) in attribute m, we update the prototypes by

$$C_m^k = \frac{1}{N_m^k} \sum \{F_m | l_m = k\}, \tag{2}$$

where C_m^k is the prototype of class k for attribute m, and $\{F_m\}$ is the representation set for a group of images w.r.t. attribute m. We compute Eq. 2 every n mini-batch, where n is set as 100 in our experiment. Only the representations of labeled samples are stored into RB and used to update PB in the experiment.

For unlabeled samples, MODC generates pseudo-labels by searching for the nearest prototype. For an unlabeled image sample I^u on attribute m, the representation is denoted as $F_m^{I^u}$, and the pseudo-labels $\hat{l_m^u}$ is calculated by

$$\hat{l_m^u} = \arg \min_{k \sim K_m} \{d(F_m^{I^u}, C_m^k)\}, \tag{3}$$

where d is the cosine similarity distance $d(a, b) = -\frac{a \cdot b}{\|a\| \|b\|}$.

3.2 MODC Learning Objectives

To learn the attribute-specific representations, we design three objective functions (defined by Eq. 4, 6, 7) covering both cluster-level and instance-level similarity learning, guiding the training of MODC in supervised and semi-supervised manners. If labels are involved in the objective function, MODC uses the ground truth label for labeled data, while using generated pseudo-labels for unlabeled data via Eq. 3.

Cluster-Level Similarity Objective: We define the cluster-level objective function in the form of a cluster-level triplet loss called *prototypical triplet loss*, which constructs triplet losses between a representation, the positive prototype, and the negative prototypes defined below. Let's use (I, l_m) to denote an image and label pair corresponding to attribute m, and the attribute-specific representation for this image I is F_m^I. In the embedding space of attribute m, there exist a positive prototype $C_m^+ = C_m^{l_m}$, and a negative prototype set with $K_m - 1$ prototypes $\{C_m^-\} = \{C_m^0, C_m^1, ..., C_m^{K_m}\} \backslash \{C_m^{l_m}\}$. We propose the prototypical triplet loss by averaging $K_m - 1$ triplet losses between the representation F_m^I, the positive prototype C_m^+, and the negative prototypes $\{C_m^-\}$,

$$\mathcal{L}_{\mathcal{C}}(I, m|l_m) = \frac{1}{K_m - 1} \sum_{c}^{\{C_m^-\}} max\{0, \alpha + d(F_m^I, C_m^+) - d(F_m^I, c)\}, \tag{4}$$

where α is a predefined margin, which is 0.2 in our experiment.

Given that a prototype is the representation of a class "center", learning a cluster-level similarity implies learning the similarity between an instance and the "center" of all instances in a class. The prototypical triplet loss learns the similarity between an instance and all class prototypes in an attribute, and consequently constrains the representation learning in a global manner.

The prototypical triplet loss has two major benefits. First, it considers the global distribution of all class prototypes and efficiently pushes a representation closer to its positive class prototype. Second, unlike the objective function in [25], it allows a margin rather than intensely forcing a representation to its positive prototype. As a result, when learning with a tiny labeled dataset, it reduces over-fitting in semi-supervised learning.

Instance-Level Similarity Objective: For fashion representation learning, learning instance-level similarity is a popular training approach. While the cluster-level objective function aids in the learning of similarities between instances and class abstracts, the instance-level objective function aids in the learning of subtle similarities between single instances.

For an image and a label pair (I, l_m) on attribute m, to construct a set of instance triplets, we select a set of (image, label) pairs,

$$\mathcal{T} = \{(I, l_m), (I^+, l_m^+), (I^-, l_m^-)|m\}, \tag{5}$$

where $l_m = l_m^+ \neq l_m^-$ on attribute m, indicating I and I^+ are more similar than I and I^- on attribute m. The instance-level similarity objective is defined as,

$$\mathcal{L}_{\mathcal{I}}(I, I^+, I^-, m) = max\{0, \alpha + d(F_m^I, F_m^{I^+}) - d(F_m^I, F_m^{I^-}\}, \tag{6}$$

where α is a predefined margin.
Self-supervised Objective: When only limited labeled samples or unlabeled samples are applied during training, regularizing the representations through image augmentations often helps the learning. We further explore some recent research on data augmentation and self-supervised learning. Similar to [7], we build a simple Siamese network to leverage the self-supervised learning constraint. For an image I, we employ a set of image augmentation methods $\{\mathcal{T}_{\mathcal{AUG}}\}$, from which we randomly select $t_1, t_2 \sim \{\mathcal{T}_{\mathcal{AUG}}\}$ to generate image augmentations $I_1 = t_1(I)$, and $I_2 = t_2(I)$. The self-supervised objective function is,

$$\mathcal{L}_{\mathcal{A}}(I_1, I_2, m) = \frac{1}{2}d(F_m^{I_1}, \oslash F_m^{I_2}) + \frac{1}{2}d(\oslash F_m^{I_1}, F_m^{I_2}), \tag{7}$$

where $\oslash$ is the stop-gradient operation , and $\oslash F_m^I$ is generated using the gradient-detached network in an iteration, as [7] defined.

3.3 Class-Specific Representation Space Segmentation

After the model is well trained via MODC, the cluster centers can be further leveraged during the inference stage to segment the attribute-specific embedding space into class-specific embedding spaces. Ideally, the class-specific embedding space only includes fine-grained representations that belong to this class, denoted as *space-positives*, while in practice, it may also include representations that do not belong to this class, denoted as *space-negatives*. Subsequently, the space construction finds the optimum trade-off between the accuracy of inclusion and coverage of space-positives.

We design a top_n segmentation strategy to allow elastic inclusion. Given an attribute m that has K_m classes, the prototypes (cluster centers) in the attribute-specific embedding space is denoted as $\{C_m^k\}$, which have been optimized by MODC. For a given image I_{new}, the attribute-specific image representation is $F_m^{I_{new}}$. Following the top_n segmentation strategy, $F_m^{I_{new}}$ is allowed to be assigned to n class-specific embedding spaces,

$$F_m^{I_{new}} \rightarrow S_m^{p_1}, ..., F_m^{I_{new}} \rightarrow S_m^{p_n}, \tag{8}$$

where $\{p_1, ..., p_n\}$ is the classes whose prototypes are the top n closest to $F_m^{I_{new}}$, and S_m^p is the class-specific embedding space of class p ($p \in K_m$) in the attribute-specific embedding space of m. Therefore, a small n leads to high inclusion accuracy but low space-positive coverage because representations are only assigned to high-confident spaces. On the other hand, a large n leads to lower accuracy but higher coverage.

4 The Training Scheme

In this section, we explain the training scheme which integrates the cluster-level and instance-level objective constraints for attribute-specific fine-grained representation learning. The training steps for supervised and semi-supervised learning are also illustrated in Figure 2.

The overall training scheme contains three stages: i) *a warm-up stage*, ii) *a supervised stage*, and iii) *a semi-supervised stage*. The warm-up stage is leveraged to form good initial representations for the prototypes, as the representations yielded based on a randomly initialized network may not be able to group images with the same labels closer in the representation space. In the warm-up stage, we only start with instance-level triplet loss on labeled samples to model convergence. The adopted network is updated using the loss function defined in Eq. 9.

After the warm-up, the supervised stage involves MODC with only labeled samples. To prepare for the MODC training, we initialize RB by pre-computing the representations for all training samples w.r.t. each attribute, and initialize PB by Eq. 2. To avoid immense memory bank but keep prototypes efficient, we limit the size of each RB account to 2,000 representations and refuse any extra representations. The cluster-level and instance-level objective functions are optimized together to learn the representations. An example of a supervised MODC iteration is shown in Algorithm 1 (*line 6-14*).

After the supervised stage, unlabeled samples are added to further guide the training in the semi-supervised stage. We incorporate the cluster-level and instance-level similarity learning for unlabeled data and also add the self-supervised learning constraints on augmented images. An example of a semi-supervised MODC iteration is shown in Algorithm 1 (*line 15-25*).

Algorithm 1 An MODC iteration

```
1:  A multi-task embedding network $Net$.
2:  A targeted attribute $m$, labeled image set $S_l$, and unlabeled image set $S_u$
3:  A batch of image (with label) triplets $\{(I, l_m), (I^+, l_m^+), (I^-, l_m^-)|m\} \sim S_l$
4:  A batch of unlabeled images $\{I^u|m\} \sim S_u$
5:  Image augmentation methods $t_1, t_2$
6:  if supervised stage then
7:      for $I, I^+, I^-$ do
8:          ① Obtain representations ← $f_\theta$
9:      end for
10:     for $I, I^+, I^-$ do
11:         ② Obtain positive and negative prototypes
12:     end for
13:     ③ Update $Net$ by Eq. 10
14:     ④ Update RB by Eq. 1, update PB by Eq. 2 every n mini-batch
15: else if semi-supervised stage then
16:     for $I, I^+, I^-, I^u$ do
17:         ① Obtain augmentations ← $t_1, t_2$, obtain representations ← $f_\theta$
18:     end for
19:     ② Assign pseudo-label $l_m^{\hat{u}}$ to $I_1^u$ by Eq. 3
20:     ③ Based on $l_m^{\hat{u}}$, randomly index a pseudo-positive and a pseudo-negative image $(I^{\hat{u}+}|l_m^{\hat{u}+} = l_m^{\hat{u}}), (I^{\hat{u}-}|l_m^{\hat{u}-} \neq l_m^{\hat{u}}) \sim S_l$
21:     for $I_1, I_1^+, I_1^-, I_1^u$ do
22:         ④ Obtain positive and negative prototypes
23:     end for
24:     ⑤ Update $Net$ by Eq. 11
25:     ⑥ Update RB by Eq. 1, update PB by Eq. 2 every n mini-batch
26: end if
```

Supervised learning involves the first two stages, while semi-supervised learning involves all three stages. The next stage starts when the previous one converges. Eq. 9, 10, 11 are the full objective functions for the warm-up, supervised, and semi-supervised stage, respectively,

$$\mathcal{L}_{warm} = \lambda_1 \mathcal{L}_{\mathcal{I}}(I, I^+, I^-, m), \tag{9}$$

$$\begin{aligned} \mathcal{L}_{SL} = \lambda_1 \mathcal{L}_{\mathcal{I}}(I, I^+, I^-, m) + \lambda_1 (\mathcal{L}_{\mathcal{C}}(I, m|l_m) \\ + \mathcal{L}_{\mathcal{C}}(I^+, m|l_m^+) + \mathcal{L}_{\mathcal{C}}(I^-, m|l_m^-))/3, \end{aligned} \tag{10}$$

$$\begin{aligned} \mathcal{L}_{SSL} = \mathcal{L}_{SL} + \lambda_1 \mathcal{L}_{\mathcal{I}}(I_1^u, I^{\hat{u}+}, I^{\hat{u}-}, m) + \lambda_2 \mathcal{L}_{\mathcal{C}}(I_1^u, m|l_m^{\hat{u}}) \\ + \lambda_1 (\mathcal{L}_{\mathcal{A}}(I_1, I_2, m) + \mathcal{L}_{\mathcal{A}}(I_1^+, I_2^+, m) \\ + \mathcal{L}_{\mathcal{A}}(I_1^-, I_2^-, m))/3 + \lambda_1 \mathcal{L}_{\mathcal{A}}(I_1^u, I_2^u, m), \end{aligned} \tag{11}$$

where λ_1 is set to 10^0 and λ_2 is set to 10^{-1}.

5 Experiments

5.1 Datasets

FashionAI [34] is a public dataset for fashion challenges. It contains 180,335 fashion images and 8 fine-grained attribute annotations on coat length, dress length, collar design, and so on, with 5-10 classes. We adopt the labeled train/val/test split of [20], which contains 144k/18k/18k samples.

DARN [15] is a fashion attribute prediction and retrieval dataset with 253,983 images. DARN has 9 attributes with class numbers varying from 7 to 55. We follow the labeled train/val/test split of [20], which contains 163k/20k/20k images after excluding unavailable ones[1].

For semi-supervised learning, we further partition the full set of training split into "labeled"/"unlabeled" subsets by ratio 10%/90% and the ground truth labels for the "unlabeled" subsets is not used during model training, even though they are available in the original dataset.

5.2 Experimental Settings

We compare our proposed method with the state-of-the-art solutions [20, 8] on the aforementioned two datasets.

Baselines. ASEN [20] attaches an attribute-aware spatial attention and an attribute-aware channel attention to a backbone network, and learns 1024 dimensional attribute-specific representations Compared with ASEN, ASEN$_{v2}$ [20]

[1] Note: Only 203,990 images are available due to broken URLs.

Model	100% labeled			10% labeled		
	MAP@100	MAP@all	Recall@100	MAP@100	MAP@all	Recall@100
ASEN	64.70	57.37	22.77	49.68	41.35	16.81
MODC(ASEN)$_{top1}$	**77.10**	**70.02**	**28.89**	**65.29**	**56.64**	**24.32**
MODC(ASEN)$_{top2}$	68.91	64.30	24.95	57.33	51.99	20.46
ASEN$_{v2}$	67.85	61.13	24.14	50.20	41.89	17.06
MODC(ASEN$_{v2}$*)*$_{top1}$	**79.29**	**72.51**	**29.78**	**64.78**	**56.36**	**24.13**
MODC(ASEN$_{v2}$*)*$_{top2}$	72.00	67.77	26.34	57.61	52.27	20.72
ASEN++	70.62	64.27	25.30	48.37	39.51	16.16
MODC(ASEN++)$_{top1}$	**80.29**	**74.32**	**30.26**	**61.61**	**52.00**	**22.73**
MODC(ASEN++)$_{top2}$	72.99	68.75	26.75	53.98	47.96	19.05

Table 1: Overall performance comparison on all attributes of FashionAI.

updates the attention module structures and achieves similar performance as ASEN with fewer training iterations. ASEN++ [8] is an extension of ASEN$_{v2}$ that further utilize the multi-scale information with a global branch and a local branch. The final representation is the composition of global and local representations with 2,048 dimensions.

Evaluation Tasks and Metrics. Mean Average Precision (MAP) and Recall are commonly used performance metrics for retrieval-related tasks [20, 8]. We further utilize these at different scales, including MAP@100, MAP@all, and Recall@100, to comprehensively evaluate the performance. MAP@100 and Recall@100 are the evaluations for top 100 retrieval results. In the e-commerce fashion retrieval domain, customer satisfaction is usually influenced by the quality of top retrieval results. MAP@all further reports the evaluation considering all retrieval results.

5.3 Experimental Results

In this section, we discuss the experimental results of different models for two datasets[2]. Table 1 and Table 2 summarize the overall performance of baselines and MODC on supervised and semi-supervised learning. Table 3 and Table 4 show the detailed performance on each attribute of FashionAI and DARN. The best performers on each network are in **bold**. The global best performers are underlined.

In the experiment, MODC allows us to leverage the query image labels that are usually available in e-commerce fashion retrieval domain. With MODC, we prioritize the retrieval in class-specific embedding spaces to retrieve space-positives, and subsequently, process the retrieval in attribute-specific embedding spaces to retrieve all the rest candidates. If the query image labels are unknown, the prioritized retrieval strategy can still be applied by assigning the pseudo-label to the query image[3].

[2] *MODC(Net)* means *MODC* build upon a specific multi-task network *Net*. The subscript top_n means using top n similarity to segment class-specific embedding spaces.

[3] More experimental result of leveraging query image with pseudo labels is included in the Supplementary.

Model	100% labeled			10% labeled		
	MAP@100	MAP@all	Recall@100	MAP@100	MAP@all	Recall@100
ASEN	58.72	52.75	20.26	51.35	45.10	16.03
MODC(ASEN)$_{top1}$	**69.45**	**59.61**	**25.36**	**61.67**	53.21	21.53
MODC(ASEN)$_{top2}$	65.85	58.67	25.04	59.17	**53.55**	**21.72**
ASEN$_{v2}$	59.66	54.29	20.88	55.34	50.02	18.00
MODC(ASEN$_{v2}$*)*$_{top1}$	**69.25**	**60.43**	**25.65**	**64.94**	**57.60**	**23.86**
MODC(ASEN$_{v2}$*)*$_{top2}$	65.80	59.14	24.95	60.91	56.25	23.15
ASEN++	61.09	55.78	21.51	54.83	49.85	17.63
MODC(ASEN++)$_{top1}$	**72.16**	**62.56**	**26.76**	**65.35**	**57.07**	**23.05**
MODC(ASEN++)$_{top2}$	67.76	61.37	26.01	61.26	55.52	22.47

Table 2: Overall performance comparison on all attributes of DARN.

Model	MAP@all for each attribute								Overall
	skirt length	sleeve length	coat length	path length	collar design	lapel design	neckline design	neck design	
ASEN	64.61	49.98	49.75	65.76	70.30	62.86	52.14	63.73	57.37
MODC(ASEN)$_{top1}$	**74.77**	**62.79**	**64.70**	**76.62**	**80.50**	**74.33**	**66.99**	**69.52**	**70.02**
ASEN$_{v2}$	65.58	54.42	52.03	67.41	71.36	66.76	60.91	59.58	61.13
MODC(ASEN$_{v2}$*)*$_{top1}$	**73.56**	**66.95**	**64.76**	**76.09**	**81.63**	**76.56**	**73.88**	**74.01**	**72.51**
ASEN++	66.31	57.51	55.43	68.83	72.79	66.85	66.78	67.02	64.27
MODC(ASEN++)$_{top1}$	**74.54**	**67.48**	**68.25**	**77.69**	**81.11**	**76.90**	**77.46**	**77.10**	**74.32**

Table 3: Performance comparison on MAP@all of each attribute on FashionAI.

Model	MAP@all for each attribute									Overall
	clothes category	clothes button	clothes color	clothes length	clothes pattern	clothes shape	collar shape	sleeve length	sleeve shape	
ASEN	36.62	46.01	52.76	56.85	54.89	56.85	34.40	79.95	58.08	52.75
MODC(ASEN)$_{top1}$	**47.45**	**54.52**	**59.37**	**63.30**	**58.95**	**64.77**	**41.24**	**86.54**	**61.53**	**59.61**
ASEN$_{v2}$	37.97	49.24	52.26	59.13	55.32	59.06	36.86	81.54	58.82	54.29
MODC(ASEN$_{v2}$*)*$_{top1}$	**46.67**	**59.13**	**58.94**	**64.60**	**61.49**	**65.95**	**42.25**	**84.61**	**61.58**	**60.43**
ASEN++	40.21	50.04	53.14	59.83	57.41	59.70	37.45	83.70	60.41	55.78
MODC(ASEN++)$_{top1}$	**49.94**	**60.75**	**58.79**	**66.34**	**62.24**	**68.41**	**45.14**	**87.41**	**65.32**	**62.56**

Table 4: Performance comparison on MAP@all of each attribute on DARN.

Performance of MODC on Supervised Learning: Table 1 shows the experimental results on FashionAI dataset. When trained on 100% labeled FashionAI data, MODC with top_1 segmentation strategy shows significant improvement on all evaluation metrics (MAP@100, MAP@all, Recall@100) consistently for all adopted baseline networks (ASEN, ASEN$_{v2}$, ASEN++). Specifically, the best MAP@all is 74.32, which is achieved by MODC(ASEN++)$_{top1}$, exceeding the corresponding baseline model ASEN++ by 15.64%. We observe the top_1 segmentation results consistently outperform those of top_2 segmentation. The potential reason is, MODC with top_1 segmentation achieves 75% inclusion accuracy and 75% space-positive coverage, which is a better trade-off compared with the 46% inclusion accuracy and 91% space-positive coverage of top_2 segmentation. On DARN, we observe a similar improvement, as shown in Table 2. In supervised learning, the best MAP@all reaches to 62.56. Table 3 and Table 4 demonstrates the MAP@all is improved for each attribute on FashionAI and DARN, respectively.

Performance of MODC on Semi-supervised Learning: Similar performance improvement trend of adopting our method is also observed for semi-

Component	10% labeled w. *MODC(ASEN)*			100% labeled w. *MODC(ASEN++)*		
	MAP@100	MAP@all	Recall@100	MAP@100	MAP@all	Recall@100
1 baseline network	49.68	41.35	16.81	70.62	64.27	25.30
2 +cluster-level loss	51.17	42.34	17.19	71.64	65.07	25.63
3 +augmentation	53.07	44.85	18.05	-	-	-
4 +pseudo label	54.19	46.25	18.66	-	-	-
5 +top_1 segmentation	65.29	56.64	24.32	80.29	74.32	30.26
6 +top_2 segmentation	57.33	51.99	20.46	72.99	68.75	26.75

Table 5: Ablation study on MODC components. Study on the overall performance on semi-supervised/supervised learning on FashionAI. Models are selected based on the best performers on these cases as we show in Table 1.

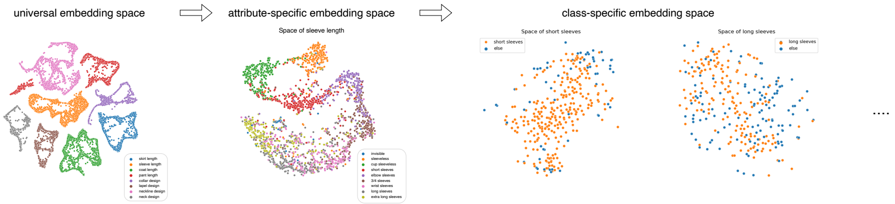

Fig. 3: Fine-grained representation distribution in the universal embedding space, attribute-specific embedding space, and class-specific embedding space. Model: MODC(ASEN$_{v2}$)$_{top1}$.

supervised settings[4], with an even larger margin. The potential reason of large margin performance improvement is, MODC is able to effectively leverage the global view and fully utilize the rich amount of unlabeled data via the designed objectives. On FashionAI, as shown in Table 1, MODC with *top*1 segmentation improves the most, compared with baseline approaches. MODC(ASEN)$_{top1}$ performs the best and reaches to 56.64 on MAP@all. On DARN, as shown in Table 2, the best MAP@all reaches to 57.60, which is 15.15% higher than the best baseline performance. Particularly, MODC(ASEN)$_{top2}$ on MAP@all and Recall@100 surpasses the top_1 MODC. With a deep-diving study, in this particular case, we find MODC with *top*2 segmentation has a better trade-off. MODC with top_2 segmentation includes 57% space-positives with 28% accuracy, while the top_1 segmentation only includes 35% space-positives with 35% accuracy.

Ablation Study of MODC: To show the effectiveness of each component of MODC, we conduct an ablation study for the case of semi-supervised/supervised learning on FashionAI, and the results are shown in Table 5. This case leverages all the components of MODC. We observe that in the supervised learning stage (row 1 and 2), the cluster-level triplet loss helps improve the representation learning, as we hypothesized. Row 3 and 4 show the data augmentation and pseudo-label assignment in semi-supervised learning is able to effectively utilize the rich amount of unlabeled data, leading to further performance improvement.

[4] More results on various labeled and unlabeled data ratio are in Supplementary.

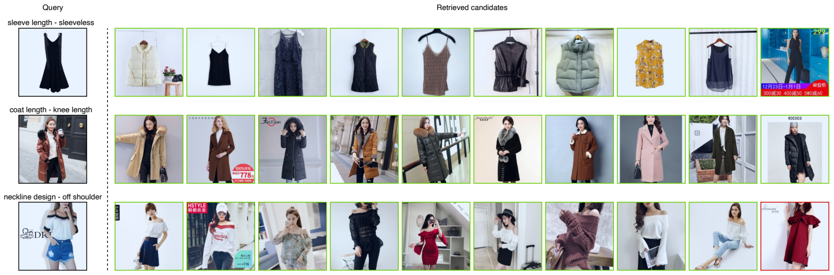
Query
Retrieved candidates
sleeve length - sleeveless
coat length - knee length
neckline design - off shoulder

Fig. 4: Fine-grained fashion retrieval examples by MODC(ASEN_{v2})$_{top1}$ on FashionAI. Green shows positive retrieves, while red shows negatives.

Furthermore, row 5 and 6 illustrate the performance boost introduced by the class-specific embedding space segmentation, which is benefit from better representation learning in rows 1-4.

Class-Specific Embedding Space and Retrieval Examples: We perform the t-SNE algorithm for the three-scale embedding spaces and generate 2D visualizations as shown in Figure 3. The left part is the universal multi-task embedding space that contains the representations of all attributes. The middle one shows the attribute-specific embedding space for a given attribute, where the instance-level and cluster level losses constrain the representation distribution. The right part is the class-specific embedding space for a given class belong to a specific attribute, which is generated based on the segmentation strategy introduced in Section 3.3. More detailed three-scale embedding spaces illustration is included in the supplementary. Figure 4 shows the fashion retrieval results in class-specific embedding spaces. We observe that most retrieved results share the same attribute class as the query image.

6 Conclusion

In this paper, we introduce Multi-task Online Deep Clustering (MODC), which learns at cluster-level and instance-level to optimize the representation distribution comprehensively. We design a three-stage training scheme to guide fine-grained representation learning in both supervised and semi-supervised fashion. MODC is able to fully utilize the rich amount of unlabeled data for performance boosting. By leveraging the cluster centers learned via MODC, the attribute-specific embedding spaces can be segmented into class-specific embedding spaces, enabling the prioritized retrieval strategy for fashion retrievals. We conduct experiments on FashionAI and DARN datasets, using evaluation metrics of MAP@100, MAP@all, and Recall@100. According to the experimental results, our proposed MODC is able to exceed the state-of-the-art solutions by a large margin, demonstrating the effectiveness of our method on fashion retrieval tasks.

References

1. Ak, K.E., Kassim, A.A., Lim, J.H., Tham, J.Y.: Learning attribute representations with localization for flexible fashion search. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7708–7717 (2018) 1, 2, 4
2. Ak, K.E., Lim, J.H., Tham, J.Y., Kassim, A.A.: Efficient multi-attribute similarity learning towards attribute-based fashion search. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1671–1679. IEEE (2018) 1, 2, 4
3. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 132–149 (2018) 4, 5
4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020) 4
5. Chen, Q., Huang, J., Feris, R., Brown, L.M., Dong, J., Yan, S.: Deep domain adaptation for describing people based on fine-grained clothing attributes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5315–5324 (2015) 3
6. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020) 4
7. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021) 4, 8
8. Dong, J., Ma, Z., Mao, X., Yang, X., He, Y., Hong, R., Ji, S.: Fine-grained fashion similarity prediction by attribute-specific embedding learning. arXiv preprint arXiv:2104.02429 (2021) 2, 4, 5, 10, 11
9. Gajic, B., Baldrich, R.: Cross-domain fashion image retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1869–1871 (2018) 1, 3
10. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: International Conference on Learning Representations (2018) 4
11. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33**, 21271–21284 (2020) 4
12. Hadi Kiapour, M., Han, X., Lazebnik, S., Berg, A.C., Berg, T.L.: Where to buy it: Matching street clothing photos in online shops. In: Proceedings of the IEEE international conference on computer vision. pp. 3343–3351 (2015) 1, 3
13. Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S.: Automatic spatially-aware fashion concept discovery. In: Proceedings of the IEEE international conference on computer vision. pp. 1463–1471 (2017) 2, 4
14. He, R., Packer, C., McAuley, J.: Learning compatibility across categories for heterogeneous item recommendation. In: 2016 IEEE 16th International Conference on Data Mining (ICDM). pp. 937–942. IEEE (2016) 3
15. Huang, J., Feris, R.S., Chen, Q., Yan, S.: Cross-domain image retrieval with a dual attribute-aware ranking network. In: Proceedings of the IEEE international conference on computer vision. pp. 1062–1070 (2015) 3, 4, 10

16. Kim, D., Saito, K., Mishra, S., Sclaroff, S., Saenko, K., Plummer, B.A.: Self-supervised visual attribute learning for fashion compatibility. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1057–1066 (2021) 3, 4
17. Kinli, F., Ozcan, B., Kirac, F.: Fashion image retrieval with capsule networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 1, 3
18. Kuang, Z., Gao, Y., Li, G., Luo, P., Chen, Y., Lin, L., Zhang, W.: Fashion retrieval via graph reasoning networks on a similarity pyramid. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3066–3075 (2019) 3
19. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: 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 (2016) 3
20. Ma, Z., Dong, J., Long, Z., Zhang, Y., He, Y., Xue, H., Ji, S.: Fine-grained fashion similarity learning by attribute-specific embedding network. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 11741–11748 (2020) 1, 2, 4, 5, 10, 11
21. McAuley, J., Targett, C., Shi, Q., Van Den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. pp. 43–52 (2015) 1, 3
22. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: European conference on computer vision. pp. 69–84. Springer (2016) 4
23. Revanur, A., Kumar, V., Sharma, D.: Semi-supervised visual representation learning for fashion compatibility. In: Fifteenth ACM Conference on Recommender Systems. pp. 463–472 (2021) 4, 5
24. Shankar, D., Narumanchi, S., Ananya, H., Kompalli, P., Chaudhury, K.: Deep learning based large scale visual recommendation and search for e-commerce. arXiv preprint arXiv:1703.02344 (2017) 3
25. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in Neural Information Processing Systems **30**, 4077–4087 (2017) 3, 7
26. Vasileva, M.I., Plummer, B.A., Dusad, K., Rajpal, S., Kumar, R., Forsyth, D.: Learning type-aware embeddings for fashion compatibility. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 390–405 (2018) 1
27. Veit, A., Belongie, S., Karaletsos, T.: Conditional similarity networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 830–838 (2017) 1, 4, 5
28. Veit, A., Kovacs, B., Bell, S., McAuley, J., Bala, K., Belongie, S.: Learning visual clothing style with heterogeneous dyadic co-occurrences. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4642–4650 (2015) 1, 3
29. Wang, Z., Gu, Y., Zhang, Y., Zhou, J., Gu, X.: Clothing retrieval with visual attention model. In: 2017 IEEE Visual Communications and Image Processing (VCIP). pp. 1–4. IEEE (2017) 1, 3
30. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018) 4
31. Zhai, X., Oliver, A., Kolesnikov, A., Beyer, L.: S4l: Self-supervised semi-supervised learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1476–1485 (2019) 4

32. Zhan, X., Xie, J., Liu, Z., Ong, Y.S., Loy, C.C.: Online deep clustering for unsupervised representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6688–6697 (2020) 4
33. Zhao, B., Feng, J., Wu, X., Yan, S.: Memory-augmented attribute manipulation networks for interactive fashion search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1520–1528 (2017) 1, 2, 4
34. Zou, X., Kong, X., Wong, W., Wang, C., Liu, Y., Cao, Y.: Fashionai: A hierarchical dataset for fashion understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019) 3, 10