AdaptiveFashion: Improving Consumer-to-Shop Clothes Retrieval with Adaptive Margin

Pendar Alirezazadeh
University of the Basque Country UPV/EHU, Spain
Email: palirezazadeh@ehu.eus

Fadi Dornaika*
University of the Basque Country UPV/EHU, Spain
IKERBASQUE, Basque Foundation for Science, Spain
* *Corresponding author*:
Email: fadi.dornaika@ehu.eus

Abdelmalik Moujahid
University of the Basque Country UPV/EHU, Spain
Email: jibmomoa@gmail.com

Abstract*—Due to different camera angles and shooting conditions, different background environments, and different postures, fashion images captured by consumer cameras usually have limited resolution. Thus, in consumer-to-shop clothes retrieval, it is hard to retrieve high-resolution online shop clothes through low-resolution consumer images. Recently, Convolutional Neural Networks (CNNs)-based approaches have been employed to extract discriminative features and subsequently improve the accuracy of clothing retrieval. In order to enhance the discriminative power of the deeply learned features, margin-based softmax losses, such as CosFace, and Additive Margin-Softmax have been proposed, but since they consider the same margin for the positive and negative pairs, they are not suitable for fashion retrieval. In this paper, we proposed Fashion Adaptive Margin (FAM) method to learn two different margins for positive and negative pairs such that the negative margin is larger than the positive margin, which provides a stronger intra-class reduction for negative pairs compared to positive pairs. Considering the larger margin for negative pairs helps us to overcome the problem of negative pairs with small and large visual differences. Experimental results on publicly available fashion datasets DARN and two benchmarks of the DeepFashion dataset: 1)Consumer-to-Shop Clothes Retrieval and 2) InShop Clothes Retrieval demonstrate the effectiveness of our proposed approach. FAM achieved Top-50 retrieval performances of 0.759, 0.921, and 0.87 on the Consumer-to-Shop Clothes Retrieval benchmark, the InShop Clothes Retrieval benchmark, and DARN dataset, respectively.

I. Introduction

Garment search between consumers and stores was recently discovered by image processing scientists to find similar images of garments in stores based on a user photo. Due to the different shooting conditions, the main challenge is the discrepancy between the photos taken by customers and the fashion images taken by professional photographers. The same clothes can look different under different circumstances such as light, different situations or poses. In contrast, different clothes can be visually similar. Recently, consumer-to-shop clothes retrieval made progress using convolutional neural networks (CNNs) [1], [2], [3], [4], [5], [6], [3], [7]. Most of the existing methods focus on proposing stronger general and local feature descriptors to extract more complete characteristics.

The problem of consumer-to-shop clothes retrieval is particularly challenging due to the differences between the two domains. Another challenging aspect of clothing retrieval is the small visual differences between specific clothing items (e.g., jeans and pants). Small visual differences lead to finding hard examples. The hard examples have many similarities with the query image, but they do not match each other. The small superficial difference causes the images to be retrieved in the wrong way, which degrades system performance.

Loss functions play an important role in network convergence and discriminative feature extraction. Recently, much attention has been paid to softmax-based loss functions. Some researchers have introduced margin-based softmax loss functions for discriminative analysis. The basic idea behind these methods is to assign an equal decision margin to each class to help CNNs learn discriminative features. However, assigning an equal decision margin to positive and negative classes leads to poor performance of the system on negative pairs with small visual differences. These pairs require a larger decision margin to distinguish them from the positive pairs as much as possible. We propose a discriminative loss function called Fashion Adaptive Margin (FAM) to favour a larger margin for negative pairs to strongly squeeze the intraclass variations of negative pairs. Siamese networks using pre-trained VGG16 backbones are trained with FAM to learn discriminative deep features for finding similar clothing images. Experiments with DeepFashion [8] and DARN [9] databases show that our proposed method performs better than state-of-the-art approaches. The main contributions of the proposed work can be summarized as follows:

- A fashion adaptive margin loss function, called FAM, is proposed to learn deep discriminative features for consumer-to-shop fashion retrieval.
- FAM learns a larger margin for negative class compared to positive class which is caused to a adaptive decision margin, to extend inter-class variation and compact the negative intra-class.
- The proposed approach achieves state-of-the-art performance on consumer-to-shop fashion retrieval datasets, including DeepFashion [8] and DARN [9].

II. Related Work

A. Fashion Retrieval

In the last decade, consumer-to-shop image retrieval [8], [10], [11], [12], [13] has been studied comprehensively. [12] proposed the concept of cross-domain clothing retrieval. Using human body posture, they have estimated the human body area and implemented cross-domain through two-step sparse coding clothing search. [14] proposed a novel region representation method which uses a binary spatial appearance mask to constrain the human body posture for pose estimation. [10] proposed the concept of accurate retrieval across scenes, with the aim of online shopping to find the exact same item on the shopping website. Dual attribute perceptual ranking network based on two completely independent branches (DARN) [9] has used feature learning for different scene domains. FashionNet proposed by [8] learns clothes retrieval by jointly predicting clothing attributes and landmarks features and applies the network to cross-scenario services for DeepFashion dataset. YNET proposed by [13] builds different deep learning branches for each domain to model domain-specific characteristics, but the public branch at the bottom of the network has learned features without considering high-level semantic information. [2] proposed a Grid Search Network (GSN) for learning feature embedding for fashion retrieval. They also utilized a reinforcement learning-based strategy to learn a specialized transformation function over the feature embedding. [1] recommended the Siamese-based networks entitled as Graph Reasoning Network (GRNet) for similarity learning between a query and clothing gallery store using global and local representations. We perform the cross-domain consumer-to-shop clothes retrieval via the Siamese networks, which have the same weights for both sub-networks. To overcome the limitations of the data problem and avoid the complexity of the network structure to extract stronger features, a novel Fashion Adaptive Margin(FAM) is proposed which is suitable for apparel search. The network is optimized with FAM to learn discriminative features and achieve more accurate matching.

B. Loss Function

Loss functions have an important role in deep embedding learning. Deep embedding learning methods enhance discriminative power by improving loss functions. Contrastive loss [15], [16] and Discriminative loss [17] optimize input pairwise samples' Euclidean distance, within a margin for inter-class in a feature space. Triplet loss [18] constructs input triplet samples to separate the positive pair from the negative pair by a Euclidean distance margin for better inter-class feature embedding. Therefore, both contrastive loss and triplet loss impose Euclidean margin to learned features. These methods depend on the number of positive and negative input pairs or triplet images. Hence, the performance of these loss functions is sensitive to the introduction of pair or triplet mining procedures, which is time-consuming [19]. Recent approaches combine Euclidean margin-based losses with softmax loss. In [20], the authors proposed a center loss to learn centers for deep features so that each class minimizes the intra-class variations and the given centers are combined with softmax loss. The deep features learned by softmax loss have intrinsic angular distribution, and Euclidean margin-based losses are incompatible with softmax loss. Recently, researchers have optimized softmax loss for intra-class variation. [21] proposed a large margin softmax (i.e., L-softmax) by adding angular constraints to each identity in order to improve feature discrimination. Furthermore, [19] improved L-softmax by normalizing the weights and proposed Angular Softmax (A-Softmax). Due to the difficulty of angular constraints optimization, [22], [23] moved from angular space to cosine space and added a cosine margin to the cosine space and proposed CosFace loss function. In a similar way, ArcFace [24] moves the additive cosine margin into the angule space and uses an additive angular margin within the cosine space to enhance the intraclass compactness and inter-class discrepancy. CosFace puts more emphasis on increasing the distance between classes, while ArcFace improves the compactness within the class and the discrepancy between classes. When searching for fashion between consumers and stores, negative pairs with small visual differences can be classified as positive pairs, reducing retrieval performance.

In contrast to existing loss functions, we propose a novel cross-domain loss in order to import two different margins into the negative and positive intra-classes to extract discriminative features and improve fashion retrieval. Figure 1 depicts the margin-based softmax losses: CosFace, Arcface and our introduced FAM loss.

III. Proposed Approach

The main objective of the proposed approach is to learn a deep embedding using a training set of positive and negative pairs. At inference time, the deep features of a query image are extracted and compared with those of the images of a gallery in order to perform the retrieval using a simple similarity measure.

Since the cross-domain fashion retrieval involves image analysis of two different domains, the use of Siamese networks in computing comparable output vectors has been preferred. Usually, Siamese networks are trained using distance metric losses such as contrastive and triplet losses for learning. Distance metric losses do not always optimize targets consistently across training, even if all possible distances within the mini-batch are considered [25]. Thus, researchers use the softmax cross-entropy loss because of its advantages. The extracted features by softmax are well-separated by the angle not by Euclidean distance and the optimization of targets by softmax converges fast and consistently. However, softmax loss is not able to extract the distinctive features optimally due to lack of normalization and limited decision boundary. This problem is especially evident when the negative pairs are very similar. Recent works [19], [21], [22], [23] have attempted to improve the performance of softmax by increasing the decision margin. By projecting Euclidean space into an angular

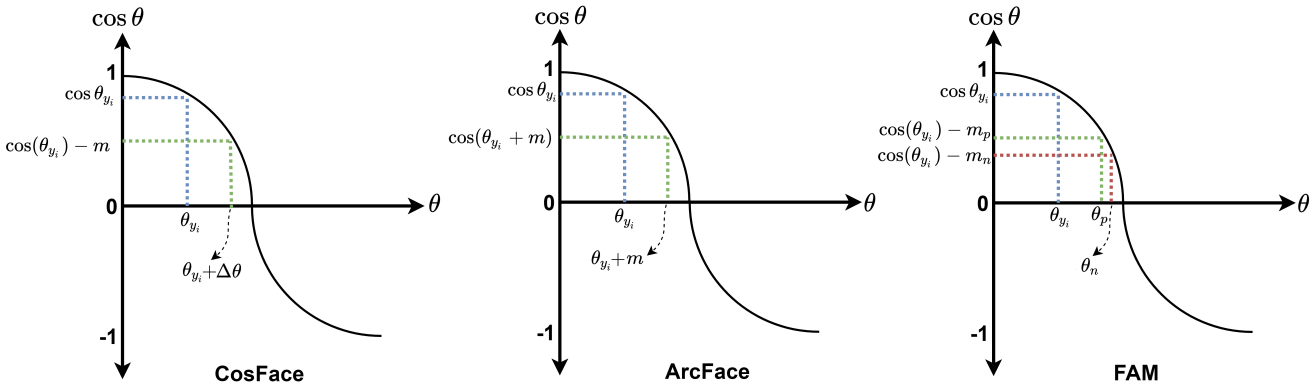

Fig. 1. The main idea of margin-based softmax losses. CosFace attempts to reduce the logit of the loss function by applying the margin m to the cos(theta) in cosine space. ArcFace, on the other hand, focuses directly on the angle and attempts to reduce the logit of the loss function and converge the network by increasing the angle between the normalized weight and the feature vector. Unlike CosFace and Arcface, which apply the same margin to positive and negative pairs, FAM learns two different margins for positive and negative pairs such that the negative margin is larger than the positive margin, allowing for intra-class reduction for negative pairs.

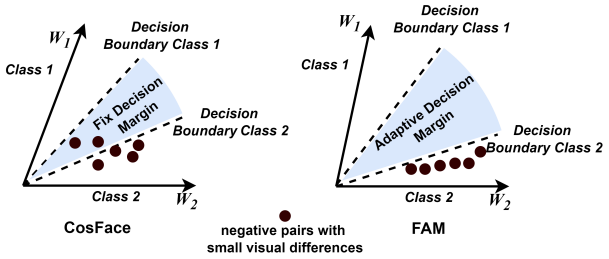

Fig. 2. Comparison between the CosFace loss and the Fashion Adaptive Margin (FAM) loss. Blue area represents the decision margin between the positive and negative classes. Extracted feature vectors of the positive or negative image pairs are fused at the feature level to form one single vector. CosFace [22] assigns the same margin m for positive and negative classes, resulting in a fixed decision margin, therefore the discrimination process cannot be strong enough. Compared to the margin for positive classes, FAM learns a larger margin for the negative classe, consequently expands the variations between classes and condenses the variations within classes, implicitly optimizing the discrimination space. Negative pairs with small visual differences move closer to negative pairs with large visual differences, resulting in hard examples being forced into the feature space of the negative class.

space, these works introduce an angular margin to extend inter-class variance that distinguishes different classes and extracts distinct features. Angular margin optimization is a difficult process and requires many assumptions [24]. Since angle cosine analysis is more compatible with softmax, the reasonable solution is to introduce margin based on the cosine of the angle between the deep features and the representation of the class in the last fully connected layer, which is also easier to optimize. Accordingly, CosFace and ArcFace losses [22], [24] are introduced by replacing the angular space with cosine space. These methods introduce a cosine margin m to maximize the decision margin in the angular space by formulating softmax as a cosine loss by L_2 normalization of the features and weights. The additive cosine margin assigns a margin m between the positive and negative classes. If a same margin m is set for the positive and negative classes, the feature distributions of the negative class may not be as compact as those of the positive class. The goal is to achieve a small intra-class for the negative pairs in addition to increasing the variation between classes. If the same margin is considered for the positive and negative classes, the negative pairs that are very similar can be considered as positive, which reduces the functionality of the system in the discrimination process. We further visualize the phenomenon through the process of distinguishing the positive pairs from the negative pairs as shown in Figure 2. Figure 2 demonstrates the process of distinguishing the positive pairs from the negative ones. Let us suppose that the normalized weight vectors $\mathbf{W_1}$ and $\mathbf{W_2}$ for the positive and negative pairs are given. In our work, the feature fusion of a pair of images is obtained by adding the deep feature vectors of the two images. Here the positive class is designated by $class1$, and the negative class by $class2$. The blue space represents the inter-class variation, and for CosFace, an equal value of m is dedicated for both positive and negative classes. This leads to compact intra-class variation of both classes equally.

In order to address this issue, we introduce a novel discriminative margin loss called FAM for cross-domain fashion retrieval. By assigning a larger margin to the negative class compared with the positive class, we attempt to simultaneously extend the inter-class variation and reduce the intra-class variation of the negative class, which further assures the absence of very similar negative pairs in the positive decision margin. Softmax separates features from different classes by maximizing posterior probability of the related class. Considering a deep feature vector $\mathbf{x}_i$ and its corresponding label y_i, softmax loss is defined as follows:

$$L_{Softmax} = \frac{1}{N}\sum_{i=1}^{N} -\log p_i = \frac{1}{N}\sum_{i=1}^{N} -\log \frac{e^{\mathbf{w}_{y_i}^T \mathbf{x}_i + b_{y_i}}}{\sum_{j=1}^{C} e^{\mathbf{w}_j^T \mathbf{x}_i + b_j}}, \tag{1}$$

where p_i indicates the posterior probability of feature vector $\mathbf{x}_i$ (one single vector which is formed by the fusion of the extracted feature vectors of the positive or negative image pairs

at the feature level) being correctly classified into related class y_i, $\mathbf{w}_j$ denotes the $j-th$ column of the weight matrix $\mathbf{W}$, b is the bias term, N is the number of training samples and C is the number of classes. By normalizing $\mathbf{x}_i$ and $\mathbf{w}_j$ using L_2 normalization, re-scaling $\mathbf{x}_i$ to s and fixing the bias $b = 0$ for simplicity [22], the feature distance is projected to feature angular as follows:

$$\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_i = \|\mathbf{w}_j\|\,\|\mathbf{x}_i\|\cos\theta_j = s(\cos\theta_j), \tag{2}$$

where θ_j is the angle between $\mathbf{w}_j$ and $\mathbf{x}_i$. Thus, both norm and angle of vectors contribute to the posterior probability. Based on this formulation, some methods have been proposed to optimize and extend the inter-class margin [22], [23] and intra-class margin [24]. Since optimization is much easier in cosine space compared to angular space, we further focus on the cosine margin analysis. By importing margin m into the cosine space of softmax and using $\cos(\theta_j) = \mathbf{w}_j^T\mathbf{x}_i$, the margin cosine loss (CosFace) [22] attempts to further distinguish it as follows:

$$L_{CosFace} = \frac{1}{N}\sum_{i=1}^{N} -\log\frac{e^{s\left(\cos\left(\theta_{y_i}\right)-m\right)}}{e^{s\left(\cos\left(\theta_{y_i}\right)-m\right)} + \sum_{j\neq y_i}^{C} e^{s\cos(\theta_j)}}, \tag{3}$$

where N is the number of training samples, $\mathbf{x}_i$ is the i-th feature vector corresponding to the ground-truth class y_i, $\mathbf{w}_j$ is the weight vector of the j-th class, and θ_j is the angle between $\mathbf{w}_j$ and $\mathbf{x}_i$.

Since cross-domain fashion retrieval is a discriminative binary issue, we have only two classes (similar and dissimilar classes). Let θ_1 and θ_2 denote the angles between the embedding feature vector and the weight vectors of class C_1 and C_2 ($\mathbf{w}_1$ and $\mathbf{w}_2$), respectively. In CosFace method, the value of margin m for positive and negative classes are considered as a constant value, which causes pairs with small visual differences (hard examples) to be identified as positive pairs. This problem is most evident in cross-domain fashion retrieval, which has high similarity in design and appearance between different types of clothing. Our goal is to extend inter-class variation to prevent hard examples from entering to the positive feature space and enhances the discriminative power. For this purpose, we assign a larger m to negative class to further reduce the intra-class variation of negative class. The cross-domain loss is formulated as follows:

$$L_{Cross-Domain} = \frac{1}{N}\sum_{i=1}^{N} -\log\frac{e^{s\left(\cos\left(\theta_{y_i}\right)-m_{y_i}\right)}}{e^{s\left(\cos\left(\theta_{y_i}\right)-m_{y_i}\right)} + e^{s(\cos(\theta_j))}}, \tag{4}$$

where N is the number of training samples, m_{y_i} is the margin corresponding to the ground-truth class $y_i \in \{p, n\}$ of the i-th pair (where for positive class is m_p and for negative class is m_n), and $j \neq y_i$. m_n should be larger than m_p. Imposing $m_n > m_p$ aims to compact negative decision boundary and to expand inter-class and reduce negative intra-class, which further ensures the absence of the hard examples to the positive feature space.

To ensure the discriminative power of cross-domain loss and provide a crucial solution, we introduce the discriminative part as follows:

$$L_{Discriminative} = -(\lambda_1 * m_p + \lambda_2 * m_n), \tag{5}$$

where λ_1 and λ_2 ($\lambda_1 < \lambda_2$) are balance factors to control the magnitude of the positive and negative margins. By combining (4) and (5), Cross-Domain Fashion Adaptive Margin Loss (FAM) is proposed as:

$$L_{FAM} = L_{Cross-Domain} + L_{Discriminative} = \frac{1}{N}\sum_{i=1}^{N} -\log\frac{e^{s\left(\cos\left(\theta_{y_i}\right)-m_{y_i}\right)}}{e^{s\left(\cos\left(\theta_{y_i}\right)-m_{y_i}\right)} + e^{s(\cos(\theta_j))}} - (\lambda_1 * m_p + \lambda_2 * m_n), \tag{6}$$

where m_p, m_n are the margin for positive and negative classes, θ_{y_i} is the angel between $\mathbf{x}_i$ (the merged feature vector of the positive or negative pair) and the vector $\mathbf{w}_{y_i}$. The hyper-parameters λ_1 and λ_2 control the discriminative power of FAM.

IV. Experiments

A. Datasets

We evaluated our proposed method on the DARN dataset, and on two benchmarks of the DeepFashion dataset: 1) InShop Clothes Retrieval, and 2) Consumer-to-Shop Clothes Retrieval.

The DARN dataset was collected from street images taken by users and professional photos provided by online shopping sites. We followed the evaluation protocol provided in [26], which considers a subset of 62,812 street images and 238,499 shop images of 13598 distinct products. The dataset was partitioned into three subsets for training, validation and test, with no overlap of products. All the images were resized to 180×100 RGB images and randomly rescaled with a scaling factor between 0.5 and 1.5, random rotation between 0 and 90 degrees, and vertical and horizontal mirroring for data augmentation.

DeepFashion dataset [8] is one of the largest datasets for clothing image analysis, and contains more than 800k images. Each image of the dataset is accompanied by labels of categories, attributes, bounding boxes, and landmarks. The presence of occlusion, deformation, illumination variations, and large variations in pose and scale have made this a challenging dataset. Consumer-to-Shop Clothes Retrieval benchmark includes 239,557 consumer-to-shop clothing images across 33,881 clothing items. The InShop Clothes Retrieval benchmark contains 52,712 images across 7,982 clothing items. To ensure fairness in comparison, the train-val-test splits are provided. In accordance with the state-of-the-arts methods, we used these splits in all our experiments. Also, every image was cropped using provided bounding boxes.

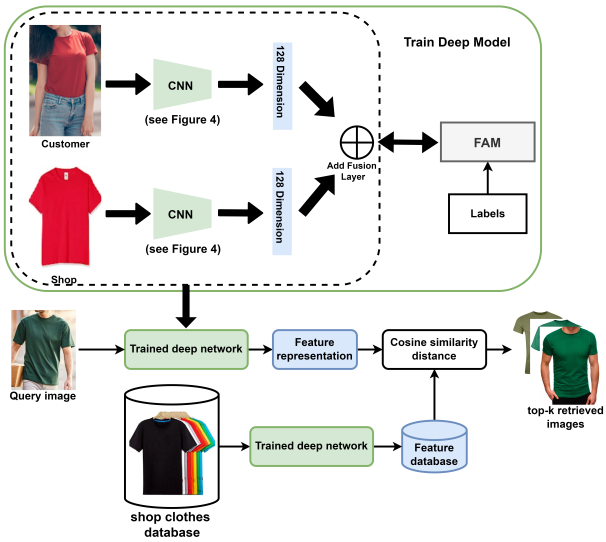

Fig. 3. The overview of our proposed cross-domain consumer-to-shop clothes retrieval system. The Siamese network consists of two subnetworks with the same architecture and weights. The extracted features of the two subnetworks are normalized by L_2 normalization. The two 128-dimensional embedding instances for customer and shop images are merged by Add fusion layer. The FAM loss drives the training of the network to learn features where the discriminative decision boundary increases and the negative margin becomes more compact. Then, the trained deep network is used to extract features from the image database and create a feature database. In the retrieval phase, features of the query image are extracted by the trained deep network and compared with the features of the feature database by the cosine similarity distance. Finally, top ranked results are displayed to the customer.

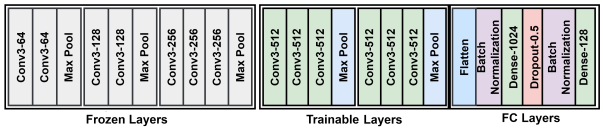

Fig. 4. The proposed CNN architecture is based on the VGG16 network. The weights of this net are pre-trained on ImageNet dataset. The fully connected layers are changed in our architecture. The weights of the first three groups are frozen and the weights of the last two groups are trained using the datasets.

B. Implementation Details

The proposed cross-domain fashion image retrieval Siamese network is shown in Fig. 3. A Siamese network includes two identical CNNs, one for shop images and another for customer images. We considered a modified VGG-16 architecture for each sub-network, pre-trained with the ImageNet dataset. The architecture of this model is given in Fig. 3. From each network, a 128-dimensional feature vector was extracted and normalized by l_2 norm, then these two features extracted from the two sub-networks are merged by Add Fusion Layer. The stochastic gradient descent (SGD) was adopted to optimize the network using the proposed L_{FAM} loss in (6). We used the initial learning rate 1×10^{-4} and weight decay with 5×10^{-4}. We follow [22] to set the feature scale s to 64 and momentum to 0.9. We choose an initial cosine margins m_n and m_p at 0.4 and 0.35, respectively. We empirically found that when $(\lambda_1, \lambda_2) = (70, 75)$, the system reaches its highest performance (see section IV-D). The model and loss layer were implemented in Python 3.6 using the Keras 2.2.4 deep learning library and trained with a batch size of 128 in a NVIDIA GeForce RTX 2080 Ti GPU. After training, the fashion retrieval problem between consumers and stores is formulated as an asymmetric (single-to-multiple) matching problem. These features are input to the similarity distance metric to perform pairwise matching between customer and store images. Then, the top ranked results are displayed to the customer. The retrieval performance of the proposed method is evaluated based on top-k accuracy as in [8], which is the ratio of correct matches within the top-k returned results.

TABLE I
TOP-K ACCURACY RATES COMPARISON ON CONSUMER-TO-SHOP CLOTHES RETRIEVAL BENCHMARK OF DEEPFASHION DATASET.

	Accuracy		
Method	Top 1	Top 20	Top 50
FashionNet	0.073	0.188	0.228
Triplet	0.109	0.378	0.499
VAM+ImgDrop	0.137	0.439	0.569
DREML	0.186	0.510	0.591
KPM	0.213	0.541	0.652
FAM	**0.236**	**0.624**	**0.759**

C. Experiments on Datasets

DeepFashion has introduced a standard protocol with training, validation and testing sets. [8] We followed this standard protocol and evaluated our approach on two benchmarks of DeepFashion: InShop Clothes Retrieval, and Consumer-to-Shop Clothes Retrieval. Table I compares the proposed FAM with state-of-the-art methods, including FashionNet [8], Siamese-Triplet [7], VAM+ImgDrop [7], DREML [27], and KPM [28] on Consumer-to-Shop Clothes Retrieval. We significantly outperform all previous works. KPM has achieved the best results among previous methods and it is based on the Siamese networks, which suggests that this type of network is the best architecture to address cross-domain image retrieval. Compared to KPM, FAM improves the retrieval performances for top-1, top-20, and top-50 by 2.3, 8.3, and 10.7 percentage points, respectively.

TABLE II
TOP-K ACCURACY RATES COMPARISON ON INSHOP CLOTHES RETRIEVAL BENCHMARK OF DEEPFASHION DATASET.

	Accuracy		
Method	Top 1	Top 20	Top 50
FashionNet	0.529	0.764	0.796
VAM	0.669	**0.892**	**0.945**
DARN	0.382	0.675	0.717
GoogleNet	0.554	0.823	0.877
FAM	**0.712**	0.875	0.921

To evaluate and demonstrate the effectiveness of the proposed method for images from the same domains, we evaluated InShop Clothes Retrieval (Table II). As can be seen in Table II, our approach achieves the best top-1 accuracy

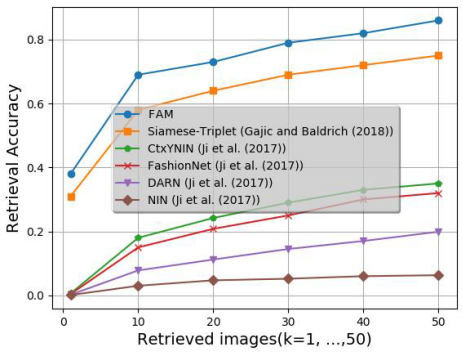

Fig. 5. Top-k accuracy rates for different methods under comparison on DARN Consumer-to-shop retrieval dataset.

of 0.712. For top-20 and top-50, our approach achieves an accuracy slightly lower than the performance of VAM. It is worth noting that VAM uses an attention subnetwork that requires a clothing segmentation dataset for training, while FAM is trained using only image pairs from queries and galleries, which is more practical. We further evaluate our method on the DARN dataset. The DARN dataset is specifically collected for street-to-shop retrieval. Because no standard protocol for the DARN dataset has been provided by its collectors, for a fair comparison, we follow the evaluation protocol provided by [26], [5]. The results are shown in Figure 5. Again, FAM outperforms the state-of-art, which is based on a Siamese-Triplet (a Siamese architecture coupled with a triplet loss function to optimize the network).

In order to show the benefit of our approach in cross-domain problems, we compare the performance of the proposed method with the state-of-the-art softmax-based loss functions. Following the implementation details in Section IV-B, we report results obtained with our Siamese network on the DeepFashion and DARN datasets with the same CNN architecture and different loss functions. Table III shows the retrieval performance (top-20) of different loss functions on DeepFashion and DARN. As can be seen, FAM achieves competitive results when compared to the other softmax-based losses across the two datasets. In particular, our loss method significantly outperforms the margin loss functions such as CosFace, which tries to extend the decision boundary and distinguish positive and negative pairs.

D. Effects of λ_1 and λ_2 on Discriminative Margin Loss

Discriminative Margin Loss consists of two parts, the cross-domain loss, and the discriminative margin average loss. The discriminative part of FAM plays an important role in preventing the positive margin m_p from becoming equal to the negative margin m_n during the training process. In this part, we conduct an experiment to investigate the effects of the different combinations of λ_1 and λ_2. By varying the value of λ_1 from 0 to 100 and λ_2 from 5 to 105, we obtain different combinations of λ_1 and λ_2. Then, we train our model on DeepFashion and DARN training subsets and validate it on the test subsets. Since our ultimate goal is to make m_n larger than m_p, we set the value of λ_2 above λ_1. As shown in Table IV, the retrieval performances on Consumer-to-Shop Clothes Retrieval benchmark of DeepFshion and DARN improves with the increase of λ_1 and λ_2 from 0 to 70 and from 5 to 75, respectively. When $(\lambda_1, \lambda_2) = (70, 75)$, the system appears to reach its highest performance and enters saturation, after which system performance begins to decline.

TABLE III
COMPARISON OF THE PROPOSED FAM WITH STATE-OF-THE-ART SOFTMAX-BASED LOSS FUNCTIONS IN CONSUMER-TO-SHOP CLOTHES RETRIEVAL (TOP-20). ALL THE METHODS IN THIS TABLE HAVE USED THE SAME TRAINING DATA AND THE SAME SIAMESE NETWORKS ARCHITECTURE.

Loss \ Dataset	Accuracy	
	DeepFashion	DARN
Softmax	0.32	0.46
SphereFace	0.55	0.59
ArcFace	0.57	0.61
CosFace	0.58	0.64
FAM	**0.62**	**0.73**

TABLE IV
THE RETRIEVAL PERFORMANCE OF DISCRIMINATIVE LOSS MARGIN WITH DIFFERENT DISCRIMINATION PARAMETERS λ_1 AND λ_2 IN CONSUMER-TO-SHOP CLOTHES RETRIEVAL (TOP-20).

Hyperparams (λ_1, λ_2) \ Dataset	Accuracy	
	DeepFashion	DARN
(0,5)	0.591	0.682
(10,15)	0.599	0.696
(20,25)	0.606	0.701
(30,35)	0.611	0.709
(40,45)	0.615	0.718
(50,55)	0.620	0.721
(60,65)	0.622	0.728
(70,75)	**0.624**	**0.733**
(80,85)	0.620	0.729
(90,95)	0.615	0.721
(100,105)	0.609	0.706

V. CONCLUSION

In this work, we proposed a cross-domain fashion adaptive margin loss (FAM) to train deep embedding features for consumer-to-shop fashion retrieval problem. FAM improves the performance of the CNNs in discriminative feature extraction. Unlike the large-margin softmax loss, FAM learns two different margins for negative and positive classes to boost the intra-class compactness and inter-class separability. The negative class margin is larger than the positive class margin, and, accordingly, FAM attempts to enhance the inter-class separation with particular focus on the negative intra-class compactness. For this reason, negative pairs with small visual differences are not considered as positive pairs, leading to improved retrieval performance. The extensive experimental

results on two public fashion datasets show clear advantages over the state-of-the-art methods and all the compared margin-based softmax functions.

REFERENCES

[1] Z. Kuang, Y. Gao, G. Li, P. Luo, Y. Chen, L. Lin, and W. Zhang, "Fashion retrieval via graph reasoning networks on a similarity pyramid," *arXiv preprint arXiv:1908.11754*, 2019.

[2] A. Chopra, A. Sinha, H. Gupta, M. Sarkar, K. Ayush, and B. Krishnamurthy, "Powering robust fashion retrieval with information rich feature embeddings," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[3] S. Park, M. Shin, S. Ham, S. Choe, and Y. Kang, "Study on fashion image retrieval methods for efficient fashion visual search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[4] J. Lasserre, C. Bracher, and R. Vollgraf, "Street2fashion2shop: Enabling visual search in fashion e-commerce using studio images," in *International Conference on Pattern Recognition Applications and Methods*. Springer, 2018, pp. 3–26.

[5] B. Gajic and R. Baldrich, "Cross-domain fashion image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1869–1871.

[6] M. Kucer and N. Murray, "A detect-then-retrieve model for multi-domain fashion item retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[7] Z. Wang, Y. Gu, Y. Zhang, J. Zhou, and X. Gu, "Clothing retrieval with visual attention model," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[8] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1096–1104.

[9] J. Huang, R. S. Feris, Q. Chen, and S. Yan, "Cross-domain image retrieval with a dual attribute-aware ranking network," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1062–1070.

[10] M. Hadi Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 3343–3351.

[11] Z. Li, Y. Li, Y. Gao, and Y. Liu, "Fast cross-scenario clothing retrieval based on indexing deep features," in *Pacific Rim Conference on Multimedia*. Springer, 2016, pp. 107–118.

[12] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3330–3337.

[13] X. Wang, Z. Sun, W. Zhang, Y. Zhou, and Y.-G. Jiang, "Matching user photos to online products with robust deep features," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 2016, pp. 7–14.

[14] Y. Kalantidis, L. Kennedy, and L.-J. Li, "Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM, 2013, pp. 105–112.

[15] S. Chopra, R. Hadsell, Y. LeCun *et al.*, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR (1)*, 2005, pp. 539–546.

[16] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[17] Y. Rao, J. Lu, and J. Zhou, "Learning discriminative aggregation network for video-based face recognition and person re-identification," *International Journal of Computer Vision*, vol. 127, no. 6-7, pp. 701–718, 2019.

[18] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1386–1393.

[19] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.

[20] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European conference on computer vision*. Springer, 2016, pp. 499–515.

[21] W. Liu, Y. Wen, Z. Yu, and M. Yang, "Large-margin softmax loss for convolutional neural networks." in *ICML*, vol. 2, no. 3, 2016, p. 7.

[22] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[23] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[24] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[25] S. Horiguchi, D. Ikami, and K. Aizawa, "Significance of softmax-based features in comparison to distance metric learning-based features," *arXiv preprint arXiv:1712.10151*, vol. 2, 2017.

[26] X. Ji, W. Wang, M. Zhang, and Y. Yang, "Cross-domain image retrieval with attention modeling," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1654–1662.

[27] H. Xuan, R. Souvenir, and R. Pless, "Deep randomized ensembles for metric learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 723–734.

[28] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6886–6895.