DeepProduct: Mobile Product Search with Portable Deep Features

YU-GANG JIANG, Fudan University
MINJUN LI, Fudan University
XI WANG, Fudan University
WEI LIU, Columbia University
XIAN-SHENG HUA, Alibaba Group

Features extracted by deep networks have been popular in many visual search tasks. This paper studies deep network structures and training schemes for mobile visual search. The goal is to learn an effective yet portable feature representation that is suitable for bridging the domain gap between mobile user photos and (mostly) professionally taken product images, while keeping the computational cost acceptable for mobile based applications. The technical contributions are two-fold. First, we propose an alternative of the contrastive loss popularly used for training deep Siamese networks, namely *robust contrastive loss*, where we relax the penalty on some positive and negative pairs to alleviate overfitting. Second, a simple multi-task fine-tuning scheme is leveraged to train the network, which not only utilizes knowledge from the provided training photo pairs, but also harnesses additional information from the large ImageNet dataset to regularize the fine-tuning process. Extensive experiments on challenging real-world datasets demonstrate that both the robust contrastive loss and the multi-task fine-tuning scheme are effective, leading to very promising results with a time cost suitable for mobile product search scenarios.

Categories and Subject Descriptors: I.4.10 [**Computing Methodologies**]: Image Processing and Computer Vision—*Image Representations*; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search Process*

General Terms: Algorithms, Design, Experimentation

Additional Key Words and Phrases: Mobile Product Search, Deep Learning, Efficiency, Contrastive Loss.

ACM Reference Format:
Jiang, Y.-G., Li, M., Wang, X., Liu, W., Hua, X.-S., DeepProduct: Mobile Product Search with Efficient Deep Features *ACM Trans. Multimedia Comput. Commun. Appl.* 9, 4, Article 39 (March 2010), 18 pages.
DOI: 0000001.0000001

1. INTRODUCTION

Shopping online on mobile devices has become increasingly popular in our daily life, where consumers typically use keywords to search products on online shopping applications. Only using textual words, however, is not always enough for all different types of goods in an ever expanding product database. For example, one may be interested in a dress with the brand unknown. In such a case, it would be very helpful if a shopping application supports visual search, so that the users can easily take a photo with their phones and search visually the same products online.

Many companies, like Amazon, Google and Alibaba, have implemented similar functions in their services. Nevertheless, automatically matching user photos to online product images remains a challenging problem due to the following reasons. Firstly, consumers usually use photos captured with their mobile phones as search queries,

Author's addresses: Y.-G. Jiang, M. Li and X. Wang, Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, 825 Zhangheng Road, Shanghai, China. E-mail: {ygj, minjunli13, xwang10}@fudan.edu.cn. W. Liu, Columbia University, New York, USA. E-mail: wliu@ee.columbia.edu. X.-S. Hua, Alibaba Group, Hangzhou, China. E-mail: huaxiansheng@gmail.com.

DOI: 0000001.0000001

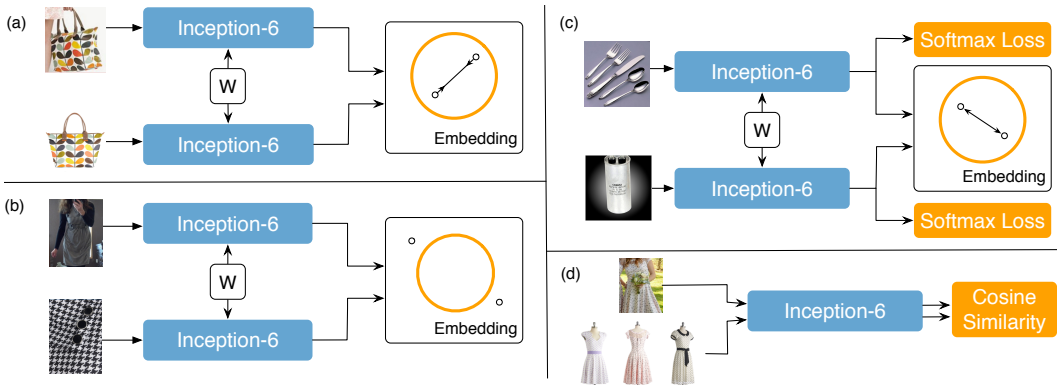

Fig. 1. An illustration of our approach. (a) Given positive pairs with a distance smaller than the pre-defined margin (orange circle), the network is optimized based on the contrastive loss. (b) Given positive pairs with a distance larger than the pre-defined margin, the network ignores them to avoid overfitting. (c) Given image pairs from ImageNet, the network is optimized based on both contrastive loss and softmax loss. (d) In the search process, we compute the cosine similarity of the extracted features to obtain the nearest samples to the query as the search results.

which are usually taken under unconstrained environments, while the product images in online shopping databases are often shot in professional studios. Secondly, products in user photos often have cluttered backgrounds or even partial occlusions. Lastly, some online product images only contain part of the goods to show details to consumers. This vast domain gap between consumer photos and in-shop images makes the consumer-to-shop image search task highly challenging.

The choice of feature representation is critical for developing an effective product image search system. In contrast to hand-crafted local image descriptors such as SIFT, using deep neural networks to learn feature representations has recently been popularly adopted in many areas. In particular, for tasks like object detection [He et al. 2016; Ren et al. 2015], image segmentation [Farabet et al. 2013] and video classification [Wu et al. 2015], the convolutional neural networks (CNNs) have produced solid performance. Unlike visual recognition problems, where using a typical CNN structure is normally sufficient, the problem to be tackled in this paper is cross domain image matching. In such a setting, training images are often provided in pairs (same/different product image pairs), while a typical CNN network is trained over the training images provided in classes. To solve such an issue, we employ the Siamese network architecture [Hadsell et al. 2006; Bell and Bala 2015; Jiang and Wang 2016] previously used in the face image matching problem [Chopra et al. 2005], which can directly rank the similarities between input image pairs by a contrastive loss function.

In this paper, we introduce a simple and effective alternative loss called *robust contrastive loss* on the adopted Siamese network, to bridge the vast domain gap between consumer photos and online shopping images. The fundamental difference is that we ignore the positive/negative image pairs that are visually too different in the training process, as such pairs may incur overfitting and a poor generalization ability of the learned model in our problem setting. Moreover, we propose a multi-task fine-tuning method to tune the parameters of the Siamese network, which combines the training of product images with general images from the ImageNet corpus. We show that optimizing the network both to match image pairs and to recognize images can improve the performance of product search. This work is extended from a conference paper [Wang et al. 2016] with a modified loss along with new evaluations and expanded discussions on the efficiency of our proposed method.

Figure 1 illustrates the proposed approach. The core component of the framework is a deep neural network called *Inception-6*, which is based on the Inception-BN network [Ioffe and Szegedy 2015] but has a more compact size designed for scenarios with limited computing resource. In the training process, given a pair of images as input, each goes through the Inception-6 network, optimization is performed based on the robust contrastive loss. Positive pairs with a distance larger than a pre-defined margin (e.g., situation (b) in Figure 1) are ignored in the training process to avoid overfitting. Moreover, for the image pairs from ImageNet (situation (c) in Figure 1), we apply the proposed multi-task fine-tuning approach, which jointly optimizes the network parameters relying on the contrastive loss of image pairs and the softmax loss of each image. After training, in the online search process (situation (d) in Figure 1), a feature representation can be quickly computed for a query image using the learned Inception-6 network, and the search results can be obtained by simply computing its similarities to the online product images. More details will be explained in Section 3.

With the robust contrastive loss and the multi-task fine-tuning scheme, the proposed Inception-6 network can achieve outstanding accuracy on the challenging Exact Street2Shop Dataset [Hadi Kiapour et al. 2015], Deepfashion Dataset [Liu et al. 2016], and the Alibaba Large-scale Product Image Dataset [1]. In the rest of the paper, we first review related works, then elaborate the proposed approach, and finally discuss experimental results on the three datasets.

2. RELATED WORKS

Product Image Search: Due to a growing number of online shopping applications, there has been increasing interest in developing effective product image search systems in both industrial and academic communities. For instance, a mobile visual search system using various local features and indexing methods has been proposed in [He et al. 2012]. Street-to-shop photo search for similar clothing items has been studied in [Liu et al. 2012] using a part alignment approach. Using clothing recognition and segmentation techniques, an approach which suggests multiple relevant clothing products based on some given images has been proposed in [Kalantidis et al. 2013]. Based on contrastive loss, a Siamese network structure has been deployed in [Bell and Bala 2015], which is similar to our approach but only used the regular contrastive loss.

Different from searching for similar products, [Hadi Kiapour et al. 2015; Liu et al. 2016] focused on finding exactly the same item in the street-to-shop scenario, which is a more challenging problem that is also closer to a practical applications. Such a task has also been studied in [Huang et al. 2015] using triplet loss [Schroff et al. 2015] and visual attributes. In this paper, we tackle the same problem of exact product search with a novel method tailored for undertaking this particular issue.

Feature Representations: The crucial part of product image search is extracting discriminative feature representations. For instance, [Kuo et al. 2012] proposed a semantic feature discovery approach through visual and textual clusters to derive semantically related feature representations. Compared with the hand-crafted features, learned feature representations using CNNs have demonstrated very impressive performance on many problems [Sharif Razavian et al. 2014]. The deep feature embedding trained to perform a specific task like object classification [Russakovsky et al. 2015] could generate competitive performance on a broad range of related tasks [Donahue et al. 2013] like fine-grained visual recognition, attribute detection, scene recognition, and general image retrieval. Motivated by this fact, we build our Inception-6 network based on the Inception-BN network [Ioffe and Szegedy 2015] and pre-train the network with the ImageNet dataset [Deng et al. 2009] in our proposed approach.

[1] https://tianchi.aliyun.com/competition/introduction.htm?raceId=231510

Similarity Learning: Using the Siamese network [Hadsell et al. 2006] to learn feature representation is related to similarity or metric learning. In this category, the Online Algorithm for Scalable Image Similarity (OASIS) [Chechik et al. 2010] is one of the most successful approaches, which learns a bilinear similarity measure over hand-crafted features. Recently, a two-layer neural network on the top of the pre-trained CNN network is employed to predict whether two features represent the same item [Hadi Kiapour et al. 2015]. An end-to-end neural network using the regular contrastive loss function for metric learning is applied in [Bell and Bala 2015]. Also based on the Siamese network, [Huang et al. 2015] proposed a Dual Attribute-aware Ranking Network (DARN) for feature learning.

Triplet loss [Schroff et al. 2015] has also been successfully applied to similarity learning. [Liu et al. 2016] introduced a VGG based deep network based on triplet loss, cross-entropy loss and l2 regression loss, which jointly predict clothing attributes and landmarks. [Simo-Serra and Ishikawa 2016] adopted a similar multi-task approach based on triplet loss, which designs a compact network for fast feature extraction. [Wu et al. 2013] introduced a deep similarity learning approach for image search called Online Multimodal Deep Similarity Learning (OMDSL) algorithm.

Due to the often existed label errors in the training data, a distance metric learning approach which is more robust to the training data noise has also been studied in [Lim et al. 2013]. Different from the previous approaches, our approach proposed in this paper is based on the Siamese network, using a novel loss function and a multi-task network fine-tuning scheme.

3. THE PROPOSED APPROACH

We first describe the Siamese structure based on Inception-6 and then introduce our proposed robust contrastive loss function, followed by the multi-task fine-tuning method with implementation details.

3.1. Siamese Network

Given an image, we use the feature extraction network to obtain the feature vector $\vec{X} = f(I, W)$, where function f denotes the feature extraction network structure, which computes the feature vector $\vec{X}$ based on network parameters W. We can have a relatively reliable W by adopting parameters from the network pre-trained on a general image dataset such as the ImageNet dataset, which has been examined in a recent work [Hadi Kiapour et al. 2015]. However, as the parameters are trained for general image classification, the feature distance between two images of different products with high visual similarity can be closer than two photos of the same product, which is undesirable for handling exact product search problem.

We design a Siamese network that contains two copies of the Inception-6 network with shared weights W, as shown in Figure 2, to learn a better feature representation that correctly maps the product image proximity to plausible feature distance. A contrastive loss function [Hadsell et al. 2006] is applied on the feature extraction layer of the Inception-6 network, which can be written as:

$$L_s(\vec{X}_p, \vec{X}_q) = Y||\vec{X}_p - \vec{X}_q||_2^2 + (1 - Y)\max(0, m^2 - ||\vec{X}_p - \vec{X}_q||_2^2),$$

where $(\vec{X}_p, \vec{X}_q)$ is a pair of input features, m is the pre-defined margin, and Y is a binary label indicating similarity of the given pair of input, which $Y = 1$ means that the input pair contains the same product (i.e., a positive pair), otherwise $Y = 0$. In the learning process, we apply the contrastive loss function to encourage the positive pair to have a smaller distance and the negative pairs to have a larger distance. For

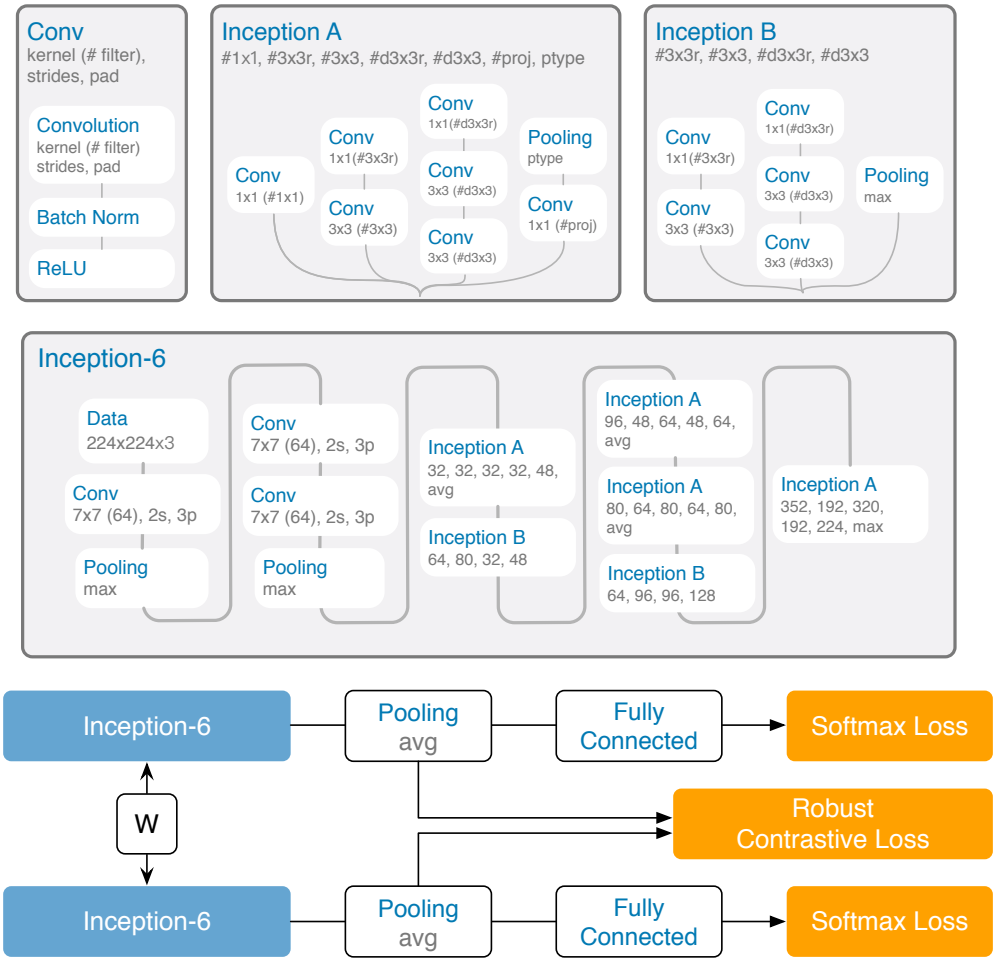

Fig. 2. The Inception-6 network and Siamese network architectures. In our experiments, the output of the average pooling layer has 1,024 dimensions, and the output of the fully connected layer has 21,841 dimensions.

a certain negative pair already having a distance greater than the margin, the loss function should not impose any further penalty.

3.2. Robust Contrastive Loss

The robust contrastive loss also runs over pairs of inputs, which is similar to the regular contrastive loss. The loss is written as:

$$L(\vec{X}_p, \vec{X}_q) = Y \min(m^2, ||\vec{X}_p - \vec{X}_q||_2^2) + \lambda(1-Y)\max(0, m^2 - ||\vec{X}_p - \vec{X}_q||_2^2),$$

where $(\vec{X}_p, \vec{X}_q)$ is a pair of input features, m is the pre-defined margin, Y is a binary label indicating similarity, and λ is a parameter balancing the trade-off of positive and negative pairs.

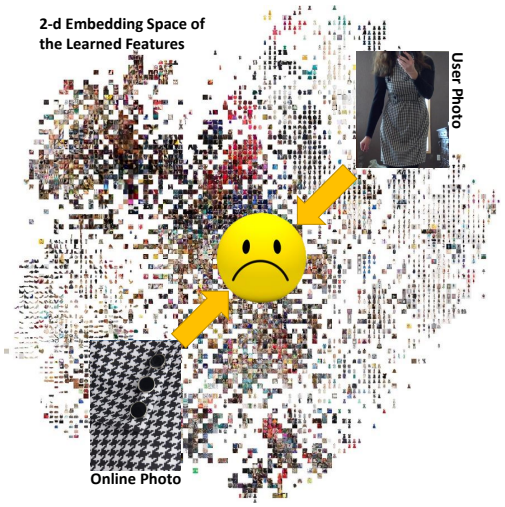

Fig. 3. Positive image pairs with small visual similarity values incur "undesirable" large gradient via the original contrastive loss function, which may result in overfitting.

The main novelties of this loss function lie in the following two points. First, not only the sufficiently separated negative pairs are less penalized, but also the penalty of positive pairs which have large feature distances is *constrained* in the robust contrastive loss function. Second, we introduce a parameter λ in the loss function to balance the penalties of positive pairs and negative pairs, which further enhances the flexibility of this robust contrastive loss function.

The major reason behind the first extension is that positive pairs with large feature distances often exist in real-world datasets, for example, the same object in significantly different scales, as shown in Figure 3. Optimizing the contrastive loss with such pairs requires a huge modification to the network parameters, which is likely to produce an overfitted model. In other words, forcing a model to fit some rare difficult cases will result in poorer generalization capability and lower overall prediction accuracy. By excluding these training pairs with the pre-defined margin, this problem can be largely alleviated. In addition, the reason of adding the λ parameter is to alleviate the effect of imbalanced positive and negative training pairs, which always exist in real-world applications. As will be shown in the experiments, balancing the contributions of positive and negative training pairs appear to be very important.

As shown in Figure 1 (c), in our proposed Siamese structure, the softmax loss is also incorporated to learn the object classification results of the input images. Using an additional loss can also help prevent overfitting caused by using the regular contrastive loss, which has been verified in [Bell and Bala 2015]. We show that, by combining our robust contrastive loss with the softmax loss, the generalization ability of the learned network can be tremendously improved.

3.3. Multi-Task Fine-Tuning

The authors of [Bell and Bala 2015] used the category information of product images in the softmax loss to regularize the fine-tuning procedure. Here we introduce a different multi-task fine-tuning scheme to incorporate additional training samples.

As previously mentioned in Section 1, we apply the Inception-6 network presented in Figure 2 in the Siamese structure. Compared with the traditional CNN networks like AlexNet [Krizhevsky et al. 2012] and GoogLeNet [Szegedy et al. 2015], the batch

normalization layers [Ioffe and Szegedy 2015] are added to the Inception-6 network, which can help the network achieve comparable or even better accuracy with much fewer training iterations. These extra layers eliminate the necessity of dropout layers, which can also be viewed as a regularizer. Compared with the original Inception-BN, the major modification of our Inception-6 is reducing the number of Inception layers and also eliminating the number of filters in most of the layers, which lead to a compact network architecture that is very suitable for resource-constrained mobile-based applications. For pre-training of the Inception-6 network, we employ the ImageNet dataset[2] [Deng et al. 2009] which has 21,841 object categories. Our pre-trained network achieves 0.315 top-1 accuracy over a randomly sampled validation dataset. After pre-training on ImageNet, we construct the Siamese network by directly creating a copy of the pre-trained Inception-6 and connecting both networks with the robust contrastive loss. The robust contrastive loss function is applied on the average pooling layers, as shown in Figure 2, while two classification softmax losses are simply adopted from the pre-trained Inception-6 network.

Our multi-task fine-tuning scheme uses the images from ImageNet and their category labels to "regularize" the fine-tuning procedure, which is different from the previous methods. The benefits of our proposed scheme are two-fold. First, our approach no longer requires annotated category labels for product images. Secondly, the softmax loss, by keeping the network to preserve the learned ImageNet embedding, can serve as an auxiliary regularizer to prevent overfitting in the training process.

In the experiments, we construct three kinds of input image pairs to form the training and validation sets for the proposed multi-task fine-tuning. The *positive pairs* are generated by using the ground-truth image item labels. The *hard negative pairs* are constructed based on the top-retrieved false positives results. And, for the *background negative pairs*, we randomly sample images from different ImageNet categories, ensuring that no pair contains images of the same object. Since only the background negative pairs have category labels, we set the gradient of the softmax loss to 0 for the other two types of training pairs. Note that most of the background negative pairs have feature distances greater than the pre-defined margin. In the training process, the remaining background negative pairs serve as regularizers to prevent overfitting. These three kinds of image pairs are generated with a ratio of approximately $1:1:4$ to construct the training and validation sets.

To fine-tune the Siamese network, we set the learning rate to 5×10^{-5}, the momentum to 0.9, and the weight decay to 10^{-4}, respectively. The margin m and balance factor λ in the robust contrastive loss function are set to 40.0 and 1.5 respectively in our following experiments if there is no additional description. The size of each mini-batch is 128, which means 128 pairs of product images are processed by the network in a single batch. The training procedure lasts 5 epochs maximally, and early stopping is used based on the results on the validation set. For the input consumer photos, after resizing smaller edge to 256, we use the 224×224 center-cropped regions as the inputs. For the online product images with bounding box labels, assuming that we have a $w \times h$ bounding box in a $W \times H$ image, the edge length of the cropped input is $min(max(w, h), min(W, H))$.

After fine-tuning, the outputs of the last average-pooling layer (1,024 dimensions) of the Inception-6 network are used as the feature representations of the input images. We adopt the Cosine similarity to measure the proximity of two images. Notice that other similarity measures or fast similarity search approaches (e.g., hashing [Lai et al.

[2]ImageNet Fall 2011 release:
http://www.image-net.org/archive/stanford/fall11_whole.tar

2015; Liu et al. 2017]) are all applicable on top of our features. However, investigating this is beyond the scope of our work.

The proposed approach is named as *R. Contrastive* in the following experiments. We also evaluate the performance of a network fine-tuned by the robust contrastive loss without the softmax loss and a network fine-tuned by the traditional contrastive loss with the softmax loss, which are named as *R. Contrastive w/o Softmax* and *Contrastive* respectively.

4. EXPERIMENTS

4.1. Experimental Setup

4.1.1. Dataset and Evaluation Measure. In most of our experiments, we use the recently released *Exact Street2Shop Dataset* [Hadi Kiapour et al. 2015] and *Deepfashion Consumer-to-Shop Dataset* [Liu et al. 2016], which focus on matching real-world street/consumer photos and online product images of clothing items. Both datasets contain two types of images: 1) *street/consumer photos*, which are taken by normal end-users under natural, uncontrolled settings, and 2) *shop photos*, which are clothing item pictures from online shopping sites, mostly taken under more professional conditions.

The Street2Shop dataset contains 20,357 street photos and 404,683 shop photos including 204,795 different clothing items from 11 broad categories. It also provides 39,479 pairs of exact matching items between the street and the shop photos. Using the settings from [Hadi Kiapour et al. 2015], we divide the street-to-shop pairs into training and testing sets with a ratio of approximately $4:1$ in each category. In the experiments, each search query includes a street photo with a bounding box indicating the location of the target item in the photo and its category label. According to [Hadi Kiapour et al. 2015], all search queries are executed within the corresponding item category.

The Deepfashion Consumer-to-Shop Dataset contains 194,165 consumer photos and 45,392 shop photos. In total there are 239,557 clothes images from 33,881 clothing items, provided with 195,540 annotated pairs between consumer and shop photos. Each image is labeled with an object bounding box, 303 different types of fashion attributes and 8 fashion landmarks. Using the setting provided by [Liu et al. 2016], we split the pairs into training, validation and testing sets with a ratio of approximately $2:1:1$. Different from the Street2Shop dataset where searches are performed within each category, search tests in Deepfashion Consumer-to-Shop Dataset are executed against all shop images from the validation and testing sets, according to [Liu et al. 2016].

In addition, to evaluate the generalization capability of our proposed method, we also adopt the *Alibaba Large-scale Product Image Dataset* (see footnote 1 for its URL). Since the testing labels of this dataset are not publicly available, we randomly split the training street-to-shop pairs (1,417 products and 92,572 manually annotated positive pairs) into smaller training and testing sets with a ratio of $2:1$ without product overlap and evaluate the search accuracy. Similar to the Deepfashion Consumer-to-Shop dataset, we do not impose any category constraint and all queries are compared against all the online product images.

Top-k accuracy is adopted to measure the search performance. Given a search query, the result is considered successful if at least one exact match item can be found from the top-k returned images.

4.1.2. Alternative Methods for Comparison. We compare with the following alternative methods to verify the effectiveness of our approach.

(1) *F. T. Similarity* [Hadi Kiapour et al. 2015], which utilizes category-specific two-layer neural networks to predict whether two features extracted by the AlexNet represent the same product item. Selective search algorithm [Van de Sande et al. 2011] is applied to extend the training and testing sets. Note that the result of this method on the Deepfashion dataset is evaluated on a network trained across all categories, which is adopted in [Liu et al. 2016].
(2) *AlexNet*, for which we directly report the results in [Hadi Kiapour et al. 2015]. The activations of the fully-connected layer FC6 (4,096-d) are used as the feature representation. The network is trained on the 1000-category subset of the ImageNet corpus [Russakovsky et al. 2015].
(3) *FashionNet* [Liu et al. 2016] consists of a VGG based network with multiple losses, which learn clothes categories, attributes, clothes similarity and fashion landmark simultaneously. Different from our approach, this method applied triplet loss. Networks are pre-trained on a subset of 300,000 images of the Deepfashion dataset and then fine-tuned on the same dataset.
(4) *Inception-6*: our pre-trained Inception-6 network, where the outputs of the last average-pooling layer (1,024-d) are used as the image representation.
(5) *R. Contrastive w/o Lambda* [Wang et al. 2016], which is the model proposed in the previous conference version. The robust contrastive loss function without the λ parameter and softmax loss function is applied to fine-tune the Inception-6 network, which is equivalent to R. Contrastive with $\lambda = 1$. We compare our R. Contrastive with this setting to demonstrate the effectiveness of the λ parameter.
(6) *Inception-6 with Attribute*, which is our pre-trained Inception-6 network fine-tuned on the Deepfashion dataset, where a cross-entropy loss is used to learn the 303 attribute labels provided in the Deepfashion dataset.
(7) *R. Contrastive with Attribute*, which uses the attribute information to further tune the R. Contrastive model. Note that the cross-entropy loss from Inception-6 with Attribute is also preserved in the multi-task fine-tuning process.

Cosine similarity is applied for all these methods except F. T. Similarity, which uses the learned two-layer neural network to predict the similarity of two input images. All the models are trained using clothes bounding boxes, except that the FashionNet method uses the fashion landmark locations in the training process.

4.2. Results on the Street2Shop Dataset

Table I. Top-20 search accuracy (%) of our proposed approach and the alternative methods on the Exact Street2Shop Dataset. Notice that the F. T. Similarity method uses category-specific models and selective object proposals, while others use unified category-independent models and simple center crops of the images.

Category	#Queries	#Testing Images	AlexNet	F. T. Similarity	Inception-6	Contrastive	R. Contrastive w/o Softmax	R. Contrastive w/o Lambda	R. Contrastive
Dresses	3,292	169,733	22.2	37.1	31.0	44.0	57.6	56.9	**59.2**
Footwear	2,178	75,836	5.9	9.6	10.9	11.2	12.7	13.1	**14.8**
Tops	763	68,418	14.4	38.1	30.7	41.5	45.1	**48.0**	47.1
Outerwear	666	34,695	9.3	21.0	16.4	**21.5**	21.3	20.3	20.7
Skirts	604	18,281	11.6	**54.6**	39.1	40.6	51.7	50.8	49.8
Leggings	517	8,219	14.5	**22.1**	17.0	15.3	16.6	15.9	20.1
Bags	174	16,308	23.6	37.4	30.5	**46.6**	42.5	46.6	46.0
Eyewear	138	1,595	10.1	35.5	34.8	**39.1**	23.1	13.8	26.1
Pants	130	7,640	14.6	**29.2**	22.3	17.7	19.1	22.3	21.3
Belts	89	1,252	6.7	13.5	24.7	19.1	**25.6**	20.2	19.8
Hats	86	2,551	11.6	**38.4**	30.2	23.3	19.6	24.4	29.0
Overall	8637	404,528	14.7	29.0	24.4	30.9	37.4	37.2	**38.9**

Results of our approach and the compared methods on the *Exact Street2Shop Dataset* are summarized in Table I. Overall, our proposed approach achieves the best average performance among all the methods under comparison, outperforming F. T.

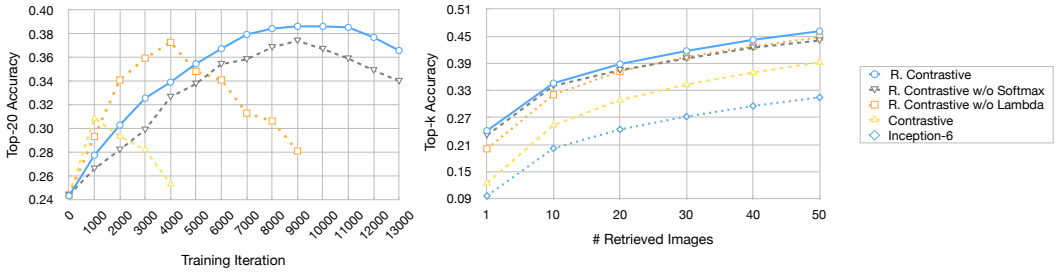

Fig. 4. Top-20 accuracy (left) and top-k accuracy (right) on the Street2Shop dataset under different experimental settings and training iterations.

Similarity by around 10 percents. In the category-specific performance, our proposed approach shows lower accuracy on product categories with fewer photos. This is because our approach uses a unified model, which has a bias to emphasize on the categories with more training data. Choosing a unified model instead of category-specific models is mainly due to speed consideration. Additionally, base on our observation that proposed Inception-6 network is clearly better than AlexNet, we can also use the F. T. Similarity in conjunction with Inception-6 for further improvements. However, training category-specific models requires more time and is difficult to be implemented in real-world applications as the category of the query is often unknown.

Comparing the last four columns of Table I, we can notice the significance contributions of our robust contrastive (8 absolute percents of gain) and multi-task training (around 2 absolute percents of gain) schemes. In the following, we further analyze various experimental results in detail.

The top-20 search accuracies of the last four approaches in Table I are visualized in Figure 4, under different numbers of training iterations. The R. Contrastive w/o Softmax approach consistently performs worse than the R. Contrastive approach, showing that it is necessary to apply multi-task approach by preserving softmax loss layer. Also, compared to the R. Contrastive w/o Lambda, the R. Contrastive approach achieves a better top-20 average accuracy, which verifies the effectiveness of using the λ parameter. Comparing to the Contrastive approach, all approaches applying robust loss have higher top-20 average accuracies. As aforementioned, positive pairs with large feature distances may incur overfitting of the Contrastive approach, which explains its poorer performance shown in the experiments. The performance of all approaches saturates at some time points and then decreases, as shown in the figure, which indicates that using early stopping strategy is also necessary to avoiding overfitting.

We also plot the top-k accuracy of our approach in Figure 4, with k ranging from 1 to 50. As expected, by applying our proposed robust loss and multi-task training scheme, we obtain consistently increasing accuracies under all top-k settings. Notice that after applying R. Contrastive approach, nearly half of the queries succeed in the top-50 retrieved images.

Figure 5 further shows the top-20 accuracy under different λ parameters and training iterations, using the R. Contrastive approach. Results indicate that an appropriate λ parameter is critical to obtain a good performance in a reasonable training time. A small balance parameter cannot fully utilize the positive samples, while a large balance parameter may incur unnecessary longer training time. We leave more discussions on the selection of this key parameter after reporting the results on the Deepfashion dataset in the next subsection.

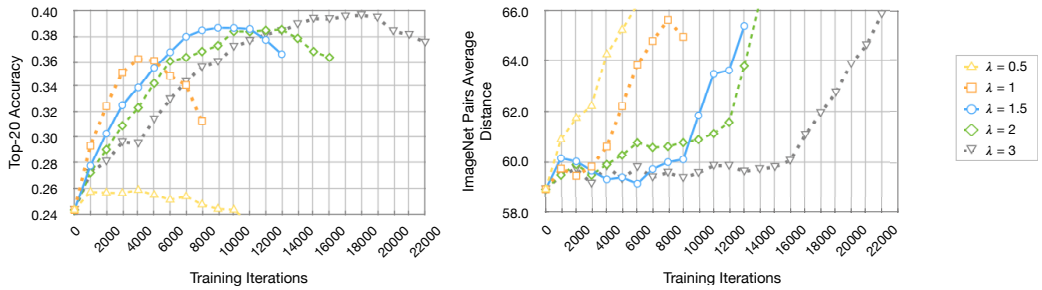

Fig. 5. Top-20 accuracy on the Street2Shop dataset (left) and average distance of the sampled ImageNet pairs (right) in the training process on proposed R. Contrastive models under different balance parameters and training iterations.

Fig. 6. Example search results of our proposed approach on the Street2Shop Dataset. Top three rows are successful queries (at least one correct match is found within top-20 returned images, marked by the green check mark icon) and search results, while the bottom three rows are failure queries with ground-truth matches and top-retrieved results.

We also monitor the average distance of the sampled ImageNet pairs during the fine-tuning process, as shown on the right of Figure 5. Since the ImageNet pairs act as the regularizer, their average distance should remain constant during an ideal fine-tuning process. Results show that by increasing the weight of the balance factor, we can better suppress the change of the ImageNet pair distance.

Some example search results on Street2Shop dataset are shown in Figure 6. Top three rows are the successful search examples, while the bottom three rows are failure

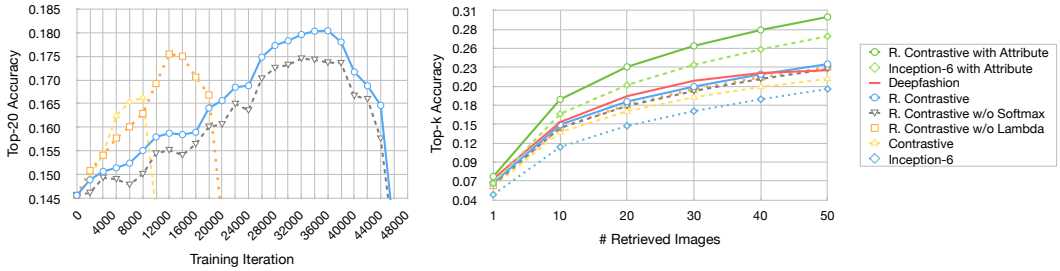

Fig. 7. Top-20 accuracy (left) and top-k accuracy (right) on the Deepfashion Consumer-to-Shop dataset under different experimental settings and training iterations.

cases. Reasons behind a failure case include poor lighting environment and defocused image, highly occluded target or simply the lack of visual characteristics of the specific product, shown in the 4^{th}, 5^{th} and 6^{th} row in Figure 6, respectively. From some successful queries, we observe that sophisticated patterns on the products are helpful as they are visually distinctive. (e.g., the query image of 3^{rd} row severely suffers from poor lighting, which is still correctly identified due to its unique pattern.)

4.3. Results on the Deepfashion Consumer-to-Shop Dataset

Table II. Top-20 search accuracy (%) of our proposed approach and the alternative methods on the Deepfashion Consumer-to-Shop Dataset. Notice that FashionNet method uses predicted fashion landmarks for object localization, while others simply use the center crops of the images.

Approach	Top-20 Accuracy
F. T. Similarity	6.3
Inception-6	14.6
Contrastive	16.7
R. Contrastive w/o Softmax	17.4
R. Contrastive w/o Lambda	17.5
R. Contrastive	18.0
FashionNet	18.8
Inception-6 with Attribute	20.4
R. Contrastive with Attribute	**23.0**

Table II summarizes the results of both our approach and the compared methods on the *Deepfashion Consumer-to-Shop Dataset*. Since the Deepfashion dataset comes with rich fashion labels, including bounding box, fashion attributes and fashion landmarks, experiments are conducted under two training label settings: without fashion attributes and with fashion attributes. Above the horizontal line are methods without using the attributes information, while below are methods fine-tuned using the 303 attributes in the dataset. Under both settings, our approach outperforms all the compared methods.

To provide a clear comparison of the overall performance of all the approaches, we also plot the top-k accuracy of our proposed approach and the compared methods in Figure 7. As shown in the figure, in the experiments without fashion attributes, similar to the results on the Street2shop dataset, the R. Contrastive approach achieves the best performance. Compared with the F. T. Similarity approach, the pre-trained Inception-6 improves the accuracy from 6.3% to 14.6%. Improvements of both robust contrastive and multi-task fine-tuning are around 1 to 2 absolute percents, which are

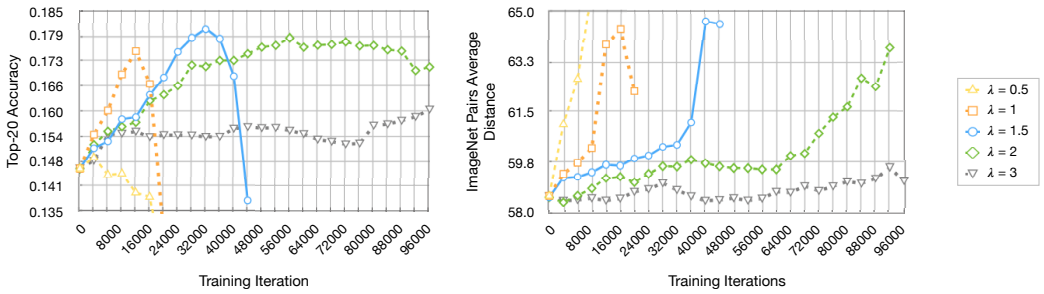

Fig. 8. Top-20 accuracy on the Deepfashion Consumer-to-Shop dataset (left) and average distance of the sampled ImageNet pairs (right) in the training process on proposed R. Contrastive models under different balance parameters and training iterations.

smaller but still fairly significant considering the average accuracy is lower than that on the Street2shop dataset. Despite the fact that we did not employ attribute information in the training process, the best performance of R. Contrastive (18.0%) is still close to the FashionNet approach (18.8%), which again highlights the effectiveness of our approach.

With fashion attributes, R. Contrastive and Inception-6 outperform the FashionNet approach by 1.6 and 4.2 absolute percentages, respectively. Compared with the R. Contrastive and Inception approaches, adding attribute information increases the performance by 5.0 and 5.8 absolute percents, respectively, which suggest the flexibility of our approach by using additional information in the multi-task fine-tuning stage. Notice that our approach only employs simple center cropping in testing, adding localization methods like selective object proposals or fashion landmark prediction will further improve the model performance.

Figure 8 shows the top-20 accuracy under different λ parameters and training iterations using the R. Contrastive approach. Based on the results in this figure and also in Figure 5, we observe that Lambda=1.5, which is actually proportional to the remaining positive:negative pair ratio[3], offers consistently good results. This result not only shows that the approach is not very sensitive to this parameter setting, but also provides a way to predict a good parameter before model training. The distance changes of the ImageNet pairs remain similar to the observations from the results on the Street2Shop dataset (Figure 5).

Comparing the trends of the search performance and the average distance in Figure 5 and Figure 8, we can see that the performance degradation is synchronized with the sudden growth of the average distance. This indicates that we can use the average distance of the testing set to monitor overfitting on-the-go while training, which is very helpful in real-world applications without sufficient annotated validation and testing data.

Figure 9 shows some example search results on the Deepfashion dataset. We can observe that, different from the Street2shop dataset, shop images in Deepfashion dataset have better visual similarity to the consumer photos, which explains the smaller improvements by the contrastive loss.

[3]By calculating the feature distance of all the training pairs and filtering the pairs that have larger distance than the pre-defined margin, the ratio of the remaining positive and negative pairs is 1.58:1 in the Street2Shop dataset and 1.34:1 in the Deepfashion dataset.

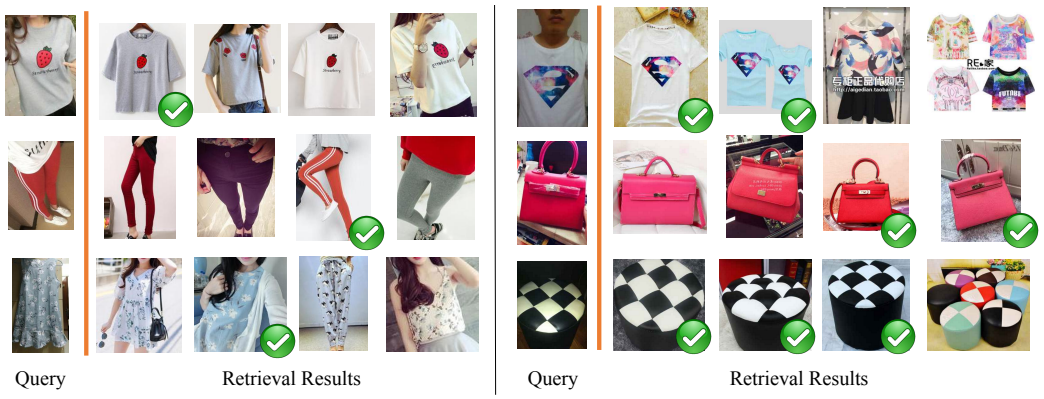

Fig. 9. Example search results of our proposed approach on the Deepfashion Dataset (left) and the Alibaba Large-scale Product Image Dataset (right).

Table III. Top-20 search accuracy (%) of our proposed method and the state-of-the-art results on the Alibaba Large-scale Product Image Dataset.

Approach	Top-20 Accuracy
AlexNet	62.0
Inception-6	65.7
Contrastive	69.5
R. Contrasive w/o Softmax	70.1
R. Contrasive w/o Lambda	71.9
R. Contrasive	**72.3**

4.4. Results on the Alibaba Dataset

To verify the generalization capability of the learned network, we also examine our proposed approach on the *Alibaba Large-scale Product Image Dataset*. As shown in Table III, we observe similar performance trend to that on the Street2Shop dataset and Deepfashion dataset. The improvement of both robust contrastive loss and softmax are around 2%, which again verifies the effectiveness of our approach. We also run the same R. Contrastive approach with only samples from this dataset to estimate the benefit from using the ImageNet samples, where the softmax loss is evaluated using the class labels (604 classes) of the Alibaba images. We reached 69.2% in top-20 accuracy, which is 3 absolute percents lower than R. Contrastive with ImageNet data (72.3%).

Similar to the Deepfashion dataset, we obtain relatively smaller improvement by using our R. Contrastive approach, which is mainly because the labels are cleaner with fewer errors in this dataset. In other words, our approach not only ignores the dissimilar positive pairs but also excludes wrongly labelled "false" positive pairs from training data to improve the results. Several example search results are shown in Figure 9.

4.5. Visualizing the Feature Embedding Space

To get a clear visual impression of our learned fashion embedding, we apply the t-SNE algorithm [Maaten and Hinton 2008] to visualize the embedding space. The features extracted by the Inception-6 network are projected to 50 dimensions using the PCA algorithm to reduce the calculation cost. Then we use the t-SNE algorithm to further compress the 50 dimensional features into a 2-d space. Lastly, we randomly select 5,000

Fig. 10. Visualization of image proximity using features from the ImageNet-trained Inception-6 network (left) and our proposed approach with the Siamese network (right). See texts for more discussions.

images from the Street2Shop dataset, including both street photos and shop photos, and place them into their corresponding 2-d locations.

Figure 10 shows the results of our visualization. The ImageNet pre-trained Inception-6 network tends to group the photos with high visual similarities together. In contrast, the fine-tuned Inception-6 based on our proposed approach tends to place the photos of the same products in close proximity, regardless of whether they are street photos or shop photos, which is important in product search applications.

4.6. Efficiency

Our experiments are conducted on a server with an Intel i7-6850K CPU and two Nvidia GTX Titan X Pascal GPUs. Fine-tuning a Siamese network with the proposed scheme normally requires 5 to 8 hours, depending on the size and complexity of the dataset. Compared with the model training and fine-tuning, which can be executed off-line, feature extraction time and the memory usage are way more important in real-time mobile visual search scenario.

Since the computational resource on mobile devices is limited, maintaining a good performance with minimal computational cost is crucial in mobile visual search. To evaluate the efficiency of our proposed Inception-6 network and robust contrastive loss, we compare with other network architectures by their feature extraction time on the GPU server and three mobile devices (iPhone 6 Plus, iPhone 7 and iPad 5th generation). The feature extraction code is based on the MXNet [Chen et al. 2015] 0.8 release.

As a baseline, we employ the AlexNet for its compact size. We also use the full Inception-BN network trained on the full ImageNet dataset released by DMLC (denoted as Inception-BN-21k) and fine-tuned with our R. Contrastive approach[4]. From Table IV, the Inception-BN-21k yields the best performance but is also the most time-consuming option, while AlexNet has the fastest speed but the lowest accuracy. Our proposed Inception-6 is well balanced between accuracy and time consumption, producing a competitive accuracy with acceptable time cost on the mobile devices.

Once the features are extracted, many approaches like hashing can be adopted to accelerate the search process, which is however beyond the scope of this work. Using

[4]We adopt the same training setting with our Inception-6 network, except that we scaled the margin to $m = 30$ based on the average distance of the randomly sampled ImageNet pairs

Table IV. Time consumption of our proposed approach and the alternative methods tested on the Exact Street2Shop Dataset.

Model	Time per Image (sec)				Top-20 Search Accuracy on the Street2Shop dataset
	Server	iPhone 6 Plus	iPhone 7	iPad 5th	
AlexNet	**0.0023**	**1.5**	**0.6**	**0.6**	14.7
Inception-6	0.0044	3.4	1.2	1.5	24.4
Inception-BN-21k	0.0117	7.3	3.2	3.7	23.5
Inception-6 (R. Contrasive)	0.0044	3.4	1.2	1.5	38.9
Inception-BN-21k (R. Contrasive)	0.0117	7.3	3.2	3.7	**40.4**

the brute-force exhaustive search, it requires 150 ms to finish the feature extraction and search of a query on the server (one single GPU for feature extraction and multi-thread CPU for search) for both Street2Shop and Deepfashion datasets.

5. CONCLUSION

In this paper, we have presented an approach based on neural networks to tackle the problem of matching a consumer-taken photo to the images of the same product in online shopping websites. To prevent the overfitting issue caused by some visually very different positive/negative pairs and meanwhile alleviate the effect of label noise, we proposed the robust contrastive loss to exclude such training samples in the network training process. We also proposed a multi-task fine-tuning scheme to provide additional data from the ImageNet dataset with a softmax loss for improved results. Our proposed approach clearly outperforms the compared methods on three real-world datasets. The time cost experiments further demonstrated that our approach is suitable for real-world applications with strict computational speed requirements.

While our results show that ignoring some positive pairs with large visual distances can help improve the overall results, the learned model will be limited when facing these ignored difficult cases during testing. We conjecture that using advanced techniques like better feature learning and fine-grained object segmentation may further improve the results, which is a promising direction of future research.

6. ACKNOWLEDGEMENTS

This work was supported by two grants from NSF China (#61622204 and #61572134).

REFERENCES

Sean Bell and Kavita Bala. 2015. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 98.

Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large scale online learning of image similarity through ranking. *Journal of Machine Learning Research* 11, Mar (2010), 1109–1135.

Tianqi Chen, Mu Li, Yutian Li, Min Lin, Naiyan Wang, Minjie Wang, Tianjun Xiao, Bing Xu, Chiyuan Zhang, and Zheng Zhang. 2015. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. *arXiv preprint arXiv:1512.01274* (2015).

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, Vol. 1. IEEE, 539–546.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 248–255.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531* (2013).

Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. 2013. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1915–1929.

M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE International Conference on Computer Vision*. 3343–3351.

Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, Vol. 2. IEEE, 1735–1742.

Junfeng He, Jinyuan Feng, Xianglong Liu, Tao Cheng, Tai-Hsu Lin, Hyunjin Chung, and Shih-Fu Chang. 2012. Mobile product search with bag of hash bits and boundary reranking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 3005–3012.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. (2016), 770–778.

Junshi Huang, Rogerio S Feris, Qiang Chen, and Shuicheng Yan. 2015. Cross-domain image retrieval with a dual attribute-aware ranking network. In *Proceedings of the IEEE International Conference on Computer Vision*. 1062–1070.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* (2015).

Yu-Gang Jiang and Jiajun Wang. 2016. Partial copy detection in videos: A benchmark and an evaluation of popular methods. *IEEE Transactions on Big Data* 2, 1 (2016), 32–42.

Yannis Kalantidis, Lyndon Kennedy, and Li-Jia Li. 2013. 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, 105–112.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

Yin-Hsi Kuo, Wen-Huang Cheng, Hsuan-Tien Lin, and Winston H Hsu. 2012. Unsupervised semantic feature discovery for image object retrieval and tag refinement. *IEEE Transactions on Multimedia* 14, 4 (2012), 1079–1090.

Hanjiang Lai, Yan Pan, Ye Liu, and Shuicheng Yan. 2015. Simultaneous feature learning and hash coding with deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3270–3278.

Daryl Lim, Brian McFee, and Gert R Lanckriet. 2013. Robust structural metric learning. In *Proceedings of the 30th International Conference on Machine Learning*. 615–623.

Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 3330–3337.

Wu Liu, Huadong Ma, Heng Qi, Dong Zhao, and Zhineng Chen. 2017. Deep learning hashing for mobile visual search. *EURASIP Journal on Image and Video Processing* 2017, 1 (2017), 17.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1096–1104.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, and others. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252. DOI:http://dx.doi.org/10.1007/s11263-015-0816-y

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 815–823.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 806–813.

Edgar Simo-Serra and Hiroshi Ishikawa. 2016. Fashion style in 128 floats: joint ranking and classification using weak data for feature extraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 298–307.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

Koen EA Van de Sande, Jasper RR Uijlings, Theo Gevers, and Arnold WM Smeulders. 2011. Segmentation as selective search for object recognition. In *International Conference on Computer Vision*. IEEE.

Xi Wang, Zhenfeng Sun, Wenqiang Zhang, Yu Zhou, and Yu-Gang Jiang. 2016. Matching User Photos to Online Products with Robust Deep Features. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 7–14.

Pengcheng Wu, Steven CH Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. 2013. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 153–162.

Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. 2015. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 461–470.

Received February 2007; revised March 2009; accepted June 2009