ARTICLE IN PRESS

Computer Vision and Image Understanding xxx (xxxx) xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval

Huijing Zhan [a,*], Boxin Shi [b,*], Ling-Yu Duan [b], Alex C. Kot [a]

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[b] National Engineering Laboratory for Video Technology, School of EECS, Peking University, China

ARTICLE INFO

Communicated by M. Cord

MSC:
41A05
41A10
65D05
65D17

Keywords:
Street-to-shop
Shoe retrieval
Feature embedding
Multi-task
Weighted triplet loss

ABSTRACT

The difficulty of describing a shoe item seeing on street with text for online shopping demands an image-based retrieval solution. We call this problem street-to-shop shoe retrieval, whose goal is to find exactly the same shoe in the online shop image (shop scenario), given a daily shoe image (street scenario) as the query. We propose an improved Multi-Task View-invariant Convolutional Neural Network (MTV-CNN+) to handle the large visual discrepancy for the same shoe in different scenarios. A novel definition of shoe style is defined according to the combinations of part-aware semantic shoe attributes and the corresponding style identification loss is developed. Furthermore, a new loss function is proposed to minimize the distances between images of the same shoe captured from different viewpoints. In order to efficiently train MTV-CNN+, we develop an attribute-based weighting scheme on the conventional triplet loss function to put more emphasis on the hard triplets; a three-stage process is incorporated to progressively select the hard negative examples and anchor images. To validate the proposed method, we build a multi-view shoe dataset with semantic attributes (MVShoe) from the daily life and online shopping websites, and investigate how different triplet loss functions affect the performance. Experimental results show the advantage of MTV-CNN+ over existing approaches.

1. Introduction

The visual object retrieval (or search) becomes a more and more popular function in smart mobile devices with cameras. To improve such techniques, there are many research works studying the visual object search problem; representative works include landmark search (Radenović et al., 2016; Babenko and Lempitsky, 2015), logo search (Jiang et al., 2015; Das Bhattacharjee et al., 2015), *etc*. In recent years, great profits generated by online shopping attract increasing attention to the fashion domain. Existing works mainly fall into the clothing parsing (Liu et al., 2014), attribute prediction (Yamaguchi et al., 2015; Liu et al., 2016a), style recognition (Yamaguchi et al., 2015a; Kiapour et al., 2014) and retrieval (Liu et al., 2012; Fu et al., 2012), handbag recognition (Wang et al., 2015b), *etc*. However, visual search on shoes is not well-studied (Zhan et al., 2017b), which also has strong demand in daily life applications and great potentials to further enhance the online shoe shopping experience.

A pair of popular shoe shown in images from fashion magazines or being worn on others' feet may impress someone looking for exactly the same pair of shoes from online shops. As several words are far from enough to depict the accurate appearances of their desired shoe items, the text-based search engine usually fails in returning satisfactory retrieval results. Thus, it is of great interest to develop a visual shoe retrieval system, since a query image represents more than thousands of words. More practically, we hope the system could return exactly the same shoe item from online shop given a daily shoe image with cluttered background taken from street, which we call street-to-shop shoe retrieval problem. Designing such a system in a cross-scenario setup is non-trivial due to the following three challenges:

(I) Cross-domain differences: Even the exactly matched shoe image pairs show large visual discrepancies due to the different background, illumination, degree of occlusion, scale, viewpoint, *etc*, as in Fig. 1(a);

(II) Nuances in appearances: The shoes with similar styles may only show subtle or fine-grained differences. As the example shown in Fig. 1(b), the appearances of these two pairs of shoes just have slight differences from the small decorations on the shoe body.

(III) Viewpoint variation: Compared to other fashion products (*e.g.*, clothing, handbag) which are usually seen from the frontal view, shoes are usually captured from more arbitrary viewpoints in both the street and online scenarios. The system sometimes is capable of finding the exactly matched online shop images with similar viewpoint as that of the query, but it often fails to search for the same item from a less similar view, as demonstrated in Fig. 1(c).

To conquer the above mentioned three challenges, we propose an improved Multi-Task View-invariant CNN (MTV-CNN+) for street-to-shop shoe retrieval.

* Corresponding authors.
E-mail addresses: zh0069ng@e.ntu.edu.sg (H. Zhan), shiboxin@pku.edu.cn (B. Shi), lingyu@pku.edu.cn (L.-Y. Duan), eackot@ntu.edu.sg (A.C. Kot).

https://doi.org/10.1016/j.cviu.2019.01.001
Received 18 June 2018; Received in revised form 25 October 2018; Accepted 6 January 2019
Available online xxxx
1077-3142/

Fig. 1. Three challenges for the street-to-shop shoe retrieval problem. (a) The exactly matched shoe images in the street and online shop scenarios show scale, viewpoint, illumination, and occlusion changes. (b) The different shoes may only have fine-grained differences. (c) Given the query daily shoe image, it is easy to retrieve the shoes in the green dashed box with similar viewpoints, but difficult to find the shoe images in the red dashed box with less similar viewpoints. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

- The distance of features for the same shoe from different scenarios is minimized and that for different shoes is increased via the triplet network;
- We incorporate two auxiliary tasks, attribute prediction and style identification, to capture the fine-grained details of shoes. Moreover, according to the hierarchical properties of the multi-class attribute, we develop an attribute-based hard example weighting and sampling strategy to select the hard negatives as well as anchor images.
- We integrate another auxiliary task of viewpoint invariance to simultaneously pull the feature embeddings of different viewpoint images in the positive bag closer.

Our main contributions are summarized as the following:

1. An improved multi-task view-invariant CNN, optimized with one main task and three auxiliary tasks, is proposed to learn the effective shoe feature representation. It minimizes the distances between images of exactly the same shoe item from different scenarios.
2. We incorporate the high-level semantic attributes, and further design an attribute-based hard example weighting as well as a stage-wise hard example mining strategy to differentiate the subtle differences in appearances.
3. The multiple view images of the same shoes are treated as a positive bag and their mutual distances are pulled closer for the effective retrieval of the same shoe from arbitrary viewpoints.

Preliminary versions of this work were presented in Zhan et al. (2017d,c). Compared to its earlier versions, the current work has the following major improvements: We design and incorporate more auxiliary tasks — style identification in addition to attribute prediction in Zhan et al. (2017c) to ensure that the learnt feature representation is not only view-invariant but also semantic-aware. We also develop a three-stage hard negative and anchor/positive image mining process to allow for more efficient parameter updating for the proposed MTV-CNN+ over its simplified two-stage version of MTV-CNN in Zhan et al. (2017c). Moreover, besides the ratio loss, we investigate the usage of the margin loss to further improve the system's performance. More experimental results and discussions are presented to show the improvements in performance.

2. Related work

2.1. Deep feature representation for retrieval

Deep learning aims to capture the hierarchical representations of data with a cascade of layers. Leveraging the deep feature representation activated from CNN has shown its effectiveness in image retrieval (Gu et al., 2018; Babenko et al., 2014; Yue-Hei Ng et al., 2015), compared with the performance using hand-crafted global features (*e.g.*, GIST (Oliva and Torralba, 2001)) and local features with encoding methods (*e.g.*, DenseSIFT with Fisher encoding (Jegou et al., 2012) or SURF with VLAD (Arandjelovic and Zisserman, 2013)).

Recently, it becomes increasingly popular to integrate semantics into the feature learning process. The cross-modal retrieval works proposed by Salvador et al. (2017) and Carvalho et al. (2018) integrated semantic regularizer and the double-triplet scheme based on semantic relevance to address the problem of food image-to-recipe. McLaughlin et al. (2017) jointly optimized the task of the human attribute recognition and person identification to learn the deep semantic feature for the problem of person re-identification. Similar ideas have also been proposed in learning the attribute-based face representation (Liu et al., 2015). In the fashion area, Liu et al. (2016a) proposed a FashionNet which learnt the clothing feature by simultaneously regressing the clothing landmark location and predicting the clothing attributes.

Other categories of fashion objects such as clothes and handbags are near-planar object with representative patterns (*e.g.*, checkerboard of classical LV bag and printed alphabets of Supreme T-shirt). In contrast, shoes show larger appearance variation with the change in viewpoint and have less distinctive patterns on the surface, which make the feature matching lose its effectiveness for the shoe retrieval. Existing works (Yu et al., 2016; Song et al., 2016) on shoes focus on the sketch-based shoe retrieval: given a query sketch of shoes, the color image of that particular type of shoes can be retrieved. Yu et al. (2016) adapted the sketch-a-net (Yu et al., 2015) into a deep triplet ranking network to learn the domain-invariant representation of shoes. Song et al. (2016) proposed a deep triplet network with two additional attribute-based tasks and the hard example sampling strategy.

Different from these works that consider the attribute classification independently in the CNN learning, we design a novel definition of "shoe style", which is a combination of different part-aware semantic shoe attributes, and propose the style identification loss correspondingly.

2.2. Deep convolutional neural network for metric learning

Over these years, researchers have integrated the metric learning into the framework of CNN, namely Siamese network (Bell and Bala, 2015) and the triplet network (Simo-Serra and Ishikawa, 2016). The former consists of two symmetric branches that take the matched pairs (same class) or non-matched pairs (different classes) as the input. It has been successfully applied to the task of product design (Bell and Bala, 2015), sketch-based image retrieval (Wang et al., 2015a), image-patch matching (Zagoruyko and Komodakis, 2017), *etc*.

The triplet network is shown to be superior to the Siamese network with three images as the input. Wang (etal Wang et al., 2014) built a multi-scale triplet network for learning the image similarity. The success of FaceNet (Schroff et al., 2015) on face verification and recognition becomes the driven force that motivates the following work on tasks of person re-identification (Liu et al., 2016c), vehicle retrieval (Liu et al., 2016b), and unsupervised video representation learning (Wang and Gupta, 2015). The recent work by Zhang et al. (2016) investigated the structure of the class label for fine-grained feature representation and Bui et al. (2017) designed a compact descriptor via a novel triplet ranking convolutional neural network to deal with the problem of sketch-based image retrieval. Jiang et al. (2016) proposed a scenario-based triplet weighting approach to address the clothing retrieval problem.

Please cite this article as: H. Zhan, B. Shi, L.-Y. Duan et al., DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval. Computer Vision and Image Understanding (2019), https://doi.org/10.1016/j.cviu.2019.01.001.

As most of the triplets may not contribute to the network parameter update (Schroff et al., 2015), there still remains one problem of these approaches: how to efficiently differentiate and select the hard negative examples in the triplets for efficient back propagation of the gradients? Existing hard negative example mining strategy (Schroff et al., 2015) only computes the feature-level distance to differentiate the hard negative and does not fully exploit the semantic information. In our proposal, the hard negative examples are evaluated in both the feature-level and attribute-level. Furthermore, we develop an attribute-based hard negative weighting and sampling strategy for network training.

2.3. Multi-Task Learning

Multi-Task Learning (MTL) jointly optimizes several tasks to achieve better performances and improves learning efficiency, compared with solving multiple tasks independently. Under the big umbrella of the deep learning model, the features from different tasks might complement with each other. Its effectiveness has been proven in a variety of computer vision areas such as face landmark localization (Zhang et al., 2014), face detection (Zhang and Zhang, 2014), pedestrian pose estimation (Gkioxari et al., 2014), *etc*. The feature learning process also benefits from exploiting the intrinsic relationships (differences and commonalities) between different tasks (Ding et al., 2015; Li et al., 2015). Nevertheless, it is still new in the fashion area, especially on shoes, partly due to the absence of accurate attribute labels. Because the descriptions of fashion images crawled from the online website are not reliable. The most relevant work is Liu et al. (2016a), which learns the clothes feature by jointly optimizing the clothing landmark detection and attribute prediction. Our work combines the multi-task learning with the triplet loss network for learning a more semantic feature representation for shoes. And the attribute annotations are provided in our newly-built dataset. Our work differs from Liu et al. (2016a) in the sense that the retrieval of shoes has its unique viewpoint variation challenge. To address this problem, we design a viewpoint invariant loss over the multi-task framework and obtain a view-invariant shoe feature representation.

3. Street-to-shop shoe retrieval

3.1. The overall framework

The overall framework of our street-to-shop shoe retrieval approach is demonstrated in Fig. 2. Our system takes the tuplet, including the triplet examples (anchor, positive and negative cropped[1] images) and the positive bag (images of different viewpoints for the same shoe item as the anchor) to train the MTV-CNN+. The semantic attribute label of the anchor image is also used as the input to compute both the attribute prediction L_a (see Section 3.2.2) and style identification loss L_s (see Section 3.2.3). Therefore, the learnt feature embedding is supposed to carry semantic information. Moreover, a view-invariant feature representation is achieved by optimizing the viewpoint invariant loss L_v (see Section 3.2.4) with images in the positive bag. Thus, multiple tasks are optimized jointly to obtain a semantic and view-invariant deep shoe feature representation. During the training procedure, hard examples are weighted based on the hierarchical grouping of the multi-class attributes, the loss of which is denoted as L_t (see Section 3.3). The corresponding hard example sampling strategy (see the right part of Fig. 2) is further developed for the fine-tuning of MTV-CNN+. It is a stage-wise procedure which progressively selects the hard negatives and anchor images with increasing difficulties.

In the online retrieval stage, given a real-world query shoe image, firstly we utilize Fast R-CNN (Girshick, 2015) to predict the shoe bounding boxes and the top-2 scored candidate boxes are returned as the region of interests (ROIs). Subsequently, they are fed forward into the MTV-CNN+ for query feature extraction. For images in the reference gallery, the same procedures are performed. Finally, the Euclidean distance computed between the query and reference gallery images, is used to return the rank list.

[1] The cropped images throughout the paper are the sub-images detected by Fast R-CNN model (Girshick, 2015).

3.2. Representation network learning via multi-task view-invariant CNN

Street-to-shop shoe retrieval matches exactly the same shoe item in two scenarios. To do this, we need to learn an effective embedding and a similarity metric so that images of the same shoe item are pulled closer, while images of different shoes are pushed away. Apart from the main task of triplet ranking, we also make fully use of the attribute-level annotations and the multi-task learning to train the triplet network with three auxiliary tasks: (1) attribute prediction; (2) style identification; (3) viewpoint invariance. The goal of training a multi-task triplet network is to learn a more semantic-aware and view-invariant embedding that captures both the low-level and high-level similarity.

3.2.1. Triplet ranking task

Let us denote a training set as $X = \{x_i^k | i = 1, 2, \dots, N\}$, where N is the total number of shoe items and each item has several images in different viewpoints. And x_i^k indicates the kth image of the ith shoe item. Given a triplet of shoe images $\{x_i^a, x_i^p, x_j^n\}$, x_i^a (anchor) and x_i^p (positive) have the same shoe item label, while x_k^n (negative) belongs to another item. Here the anchor is randomly selected from X, the positive image shares the same shoe item ID as the anchor while the negative image belongs to a different shoe. Fig. 3 illustrates the basic architecture of the triplet network. We wish to learn a deep embedding network $f_{\mathbf{w}}(\cdot) \in \mathbb{R}^D$, which maps an input image x to a point in the D-dimensional Euclidean space. Here $\mathbf{w}$ indicates the weights and biases of the L-layer deep embedding network. The learnt $f_{\mathbf{w}}(\cdot)$ should minimize the Euclidean distances $d_{ap} = \left\| f_{\mathbf{w}}(x_i^a) - f_{\mathbf{w}}(x_i^p) \right\|_2$ of the matched pairs (x_i^a, x_i^p) while maximize the distances $d_{an} = \left\| f_{\mathbf{w}}(x_i^a) - f_{\mathbf{w}}(x_j^n) \right\|_2$ between the non-matched pairs (x_i^a, x_j^n). The loss proposed in the triplet embedding network can be categorized into two groups, the ratio loss (Simo-Serra and Ishikawa, 2016) and the margin loss (Schroff et al., 2015).

Ratio Loss: The ratio loss is defined by normalizing the feature distances to have the unit norm rather than the features. A softmax layer is built on top of the feature distances so that they are within the range of $[0, 1]$.

$$l_+ = \frac{e^{d_{ap}}}{e^{d_{ap}} + e^{d_{an}}}, \tag{1}$$

and

$$l_- = \frac{e^{d_{an}}}{e^{d_{ap}} + e^{d_{an}}} \quad . \tag{2}$$

Thus the triplet-based ratio loss L_R to be minimized is defined as:

$$L_R(x_i^p, x_i^a, x_j^n) = 0.5 \times [(1 - l_-)^2 + l_+^2] = l_+^2. \tag{3}$$

Margin Loss: The margin loss is used as a convex alternative to the 0–1 ranking loss which is infeasible to compute the gradient descent. It is also used in the siamese network (Bell and Bala, 2015) which computes the loss based on the pair of examples that violates the pre-defined margin.

$$L_M(x_i^p, x_i^a, x_j^n) = \max(0, d_{ap}^2 - d_{an}^2 + \sigma), \tag{4}$$

where σ is the margin threshold that controls the distance $d_{ap}^2 - d_{an}^2$ between the matched pairs (anchor vs. positive) and the non-matched pairs (anchor vs. negative). If their relative distance is smaller than $-\sigma$, then the loss L_M is 0; otherwise, the loss is calculated and the parameters of the network are updated based on the gradient decent.

Please cite this article as: H. Zhan, B. Shi, L.-Y. Duan et al., DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval. Computer Vision and Image Understanding (2019), https://doi.org/10.1016/j.cviu.2019.01.001.

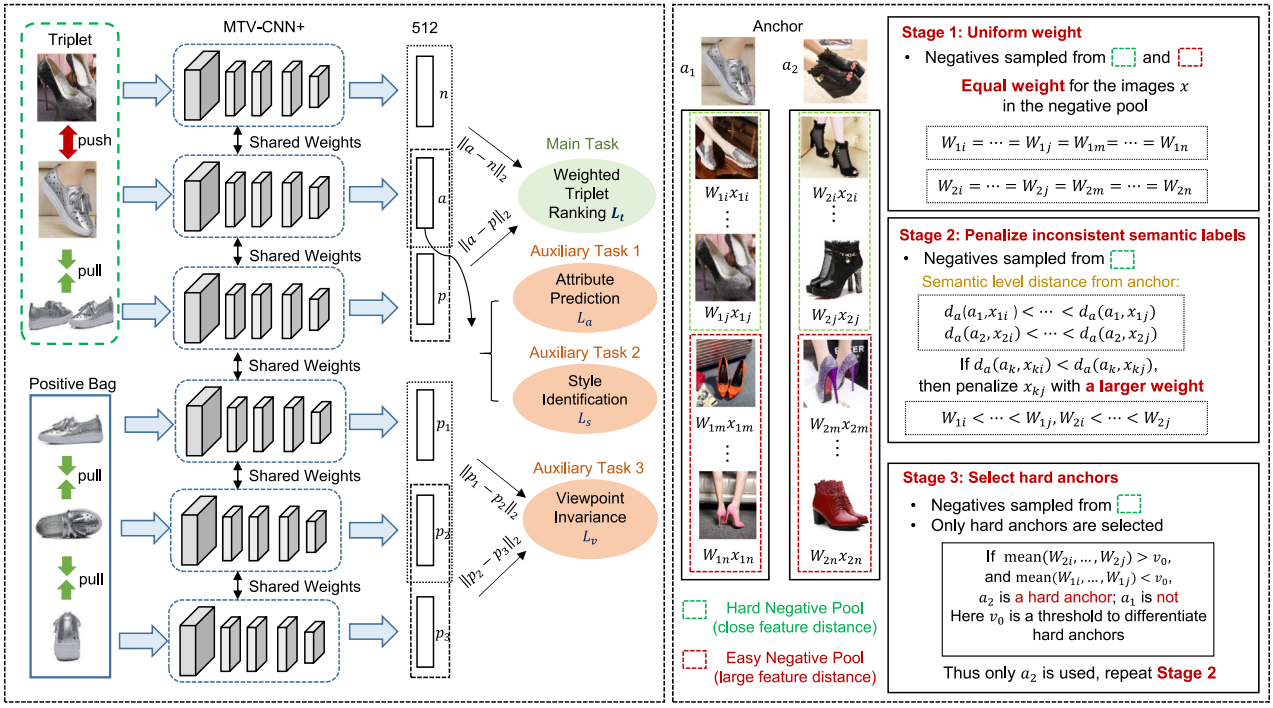

Fig. 2. The framework of our street-to-shop shoe retrieval system. On the left part, we show the overall network architecture for the training procedure with the tuplet, including the triplet examples and the corresponding positive bags. Here a, p, n indicate the 512-d feature vectors for the anchor, positive, and negative images, respectively. The feature representations for different viewpoint images of the anchor image, named as positive bag, are denoted as p_1, p_2, p_3. On the right part, we show the three-stage hard example selection process for effective network parameter learning. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

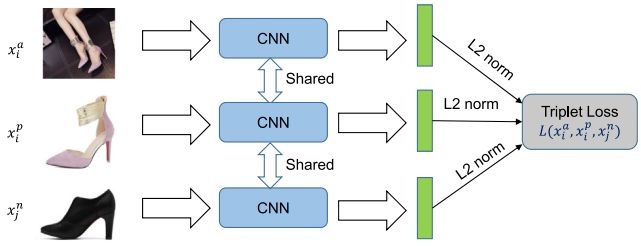

Fig. 3. The basic architecture of the conventional triplet network with three images (x_i^a, x_i^p, x_j^n) as the input.

3.2.2. Attribute prediction

The fully-connected layer involves a large number of parameters. For ease of network optimization, only the anchor image is fed into the fully-connected layer to compute the attribute prediction loss. The attributes are annotated based on per shoe item. For example, given N_i images (including daily and online images) of a particular shoe item x_i, they share the attribute labels $\mathbf{a}_i = \{\mathbf{a}_i^1, \mathbf{a}_i^2, \ldots, \mathbf{a}_i^m, \ldots, \mathbf{a}_i^M\}$. Through this work, we deal with the multi-class attributes and M is the number of the multi-class attributes. Let $\mathbf{a}_i^m = \{a_i^{m,1}, a_i^{m,2}, \ldots, a_i^{m,G_m}\}$ denote a vector indicating the attribute membership, where G_m is the number of attributes in the mth $(1 \leqslant m \leqslant M)$ multi-class attribute. Note that for each multi-class attribute, only one attribute label is nonzero, which means $\left\|\mathbf{a}_i^m\right\|_0 = 1$. As the number of training data with specific attribute labels varies, the weighted softmax loss for the task of mth attribute prediction L_m is formulated as follows:

$$L_m = -\sum_{g=1}^{G_m} \bar{w}_{m,g} \mathbf{1}(a_i^{m,g}) \log p_i^{m,g}, \quad m = 1, 2, \ldots, M, \tag{5}$$

where

$$p_i^{m,g} = \frac{\exp(z_i^{m,g})}{\sum_{g=1}^{G_m} \exp(z_i^{m,g})}, \tag{6}$$

$\mathbf{1}(\cdot)$ is the indicator function and $p_i^{m,g}$ indicates the probability of assigning x_i to the gth class in the mth multi-class attribute and $z_i^{m,g}$ is the output from the softmax layer of gth node. Take the "Toe Shape" (mth multi-class attribute) as an example, Eq. (6) is to compute the probability of the given shoe item x_i belonging to the "Pointy Toe" class (gth class). The weights are inversely proportional to the number of training samples defined as below:

$$w_{m,g} = \frac{1}{N_{m,g}}, \quad \bar{w}_{m,g} = \frac{w_{m,g}}{\sum_{g=1}^{G_m} w_{m,g}} \tag{7}$$

where $N_{m,g}$ is the number of training samples with the attribute label g in the mth multi-class attribute. Note that we only consider $N_{m,g} \geqslant 50$ to compute the respective weights $w_{m,g}$ and $\bar{w}_{m,g}$. For the minority attributes with $N_{m,g} < 50$, its $\bar{w}_{m,g}$ is directly set to 1 for simplicity. Thus, the attribute prediction loss over all the M multi-class attributes is computed as below:

$$L_a = \frac{1}{M} \sum_{m=1}^{M} L_m. \tag{8}$$

3.2.3. Style identification

For each shoe image, the crawled textual descriptions from the web also include the style of the shoes. Nevertheless, due to subjectivity of human's fashion perception, these descriptions are quite noisy and cannot be directly utilized to train the network. Furthermore, it is quite challenging to clean the noisy descriptions by manual annotations, which involves extensive human labor to obtain un-biased human judgments of style (Kiapour et al., 2014). Due to these limitations, we develop a novel style definition for shoes, which can guide the network to learn style-sensitive features.

Our shoe style definition is based on the combination of part-aware semantic shoe attributes (as illustrated in Fig. 4). Shoe images with similar part structure and appearances are grouped closer to form a style-specific cluster (style class). By considering both the distribution of attribute labels and their appearances within the multi-class attribute, the "Toe Shape" and "Heel Shape" are chosen to represent the shoe

Please cite this article as: H. Zhan, B. Shi, L.-Y. Duan et al., DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval. Computer Vision and Image Understanding (2019), https://doi.org/10.1016/j.cviu.2019.01.001.

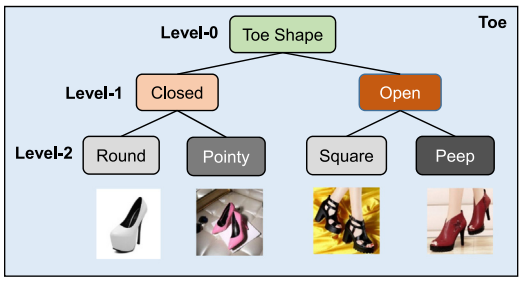

Fig. 4. Hierarchical grouping of the tree-structured semantic shoe attributes for "Toe Shape".

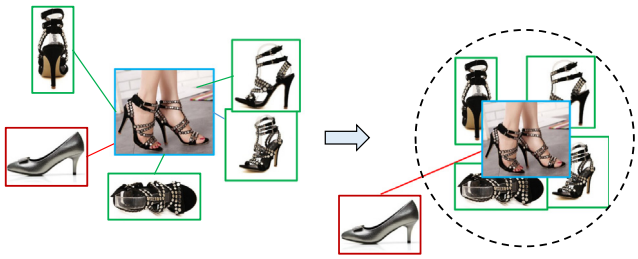

Fig. 5. The illustration for the viewpoint invariant loss. In the right figure, the positive examples of the same shoes (green box) are pulled closer. By minimizing the viewpoint invariant loss as an auxiliary task, the anchor (blue box) and positive bag (green box) are clustered while the negative (red box) are pushed away. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

styles. For example, the combination of "flat heel" and "peep toe" is considered as a casual style. Meanwhile, we may consider the combination of "flat heel" and "pointy toe" as one style which is both professional and casual. The shoes with "high heel" and "pointy toe" are considered as professional and sexy.

Let a^H and a^T denote the Heel Shape attribute set and Toe Shape attribute set, respectively. In this scenario, there are $|a^H|\times|a^T|$ different attribute combinations, where $|\cdot|$ is the cardinality of the attribute sets. Each combination is treated as a shoe style class k_s, where $1 \le k_s \le |a^H|\times|a^T|$. Let $s_i = \{s_i^1, s_i^2, \ldots, s_i^{k_s}, \ldots, s_i^{|a^H|\times|a^T|}\}$ denote the style label vector for each ith shoe sample and each entry of the s_i can be denoted as:

$$s_i^{k_s} = \mathbf{1}(a_i^{H,k_H}) \times \mathbf{1}(a_i^{T,k_T}), \tag{9}$$

where $1 \le k_H \le |a^H|$, $1 \le k_T \le |a^T|$. Each shoe item belongs to a style with one non-zero value activated in s_i. After filtering out the styles with less than 10 shoe items in the training data, the clean version of the shoe style list has 22 different style classes. The style classification is similar to the attribute prediction, and it branches out 22 nodes after the pool5 layer. Due to the imbalanced distribution of style class labels, the weighted softmax loss is also adopted for style classification task. Similar as Eq. (5), the style classification loss L_s is to minimize the softmax loss between the style label s_i and the predictions.

3.2.4. Viewpoint invariance

There is still a problem of the traditional triplet loss: it does not constrain the distances between the online shop images of the same shoe item, which usually orient to multiple viewpoints. Thus, the features for the online shop shoe images of different viewpoints may be scattered in the embedding space. Even we find the counterpart image as the query with similar viewpoint, it may fail to retrieve other images of the same item but in less similar viewpoints. To deal with this problem, we design a novel viewpoint invariant loss L_v so that the online shop scenario images from different viewpoints are drawn closer.

Assume that the ith shoe is depicted by $X_{i,s} = \{x_{i,s}^1, \ldots, x_{i,s}^{N_s}\}$, which contains N_s street shoe images, and $X_{i,o} = \{x_{i,o}^1, \ldots, x_{i,o}^{N_o}\}$ which contains N_o online shop shoe images. Here we treat the online shop shoe images for the same shoes as a positive bag and aim to make the features of the images in the positive bag similar to each other. This can be formulated to minimize the mutual distance of any image pair $(x_{i,o}^j, x_{i,o}^k)$ denoted as $d_{jk} = \left\| f_{\mathbf{w}}(x_{i,o}^j) - f_{\mathbf{w}}(x_{i,o}^k) \right\|_2$ in the online shoe image set $X_{i,o}$. Thus L_v can be written as:

$$L_v(X_{i,o}) = \frac{1}{2\times n_d} \sum_{j,k}^{N_o} d_{jk}^2, \tag{10}$$

where n_d is the number of pairs sampled from the positive bag X^o and $d(\cdot)$ is the Euclidean distance. Indeed, it is desirable to feed all the online shop shoe images inside the positive bag into the batch for computation. However, due to the memory bottleneck of deep triplet network, three shop-scenario examples are randomly selected to form the mini-batch. Thus in this paper we set $n_d = 3$ in our experiments. The concept of our viewpoint invariant triplet loss is illustrated in Fig. 5.

As the input of the proposed MTV-CNN+ is forwarded with a $(3+n_d)$-tuplet, to generate more shop-scenario images of different viewpoints for achieving the view-invariant representation, we augment the data by in-plane rotation and flipping operation. Note that this augmentation technique is only applied to for those shoe items with $N_o < 4$. We assume the rotation between the street image and online shop shoe image is a relative shift, thus the rotation is merely performed on the online shop shoe images. In the training stage, each shoe item should have at least three online shop shoe images to guarantee that there are enough images to forward into the MTV-CNN+ for computing the viewpoint invariant loss L_v.

Therefore, the multi-task loss function J_0 is a weighted combined loss of the above mentioned main task and three auxiliary tasks:

$$J_0 = \frac{1}{T} \sum_{i,j\in X}^{T} [L(x_i^p, x_i^a, x_j^n) + \alpha L_a + \beta L_s + \gamma L_v(X^o)], \tag{11}$$

where α, β, γ are the weights to indicate the importance of each auxiliary tasks and T is the number of triplet examples in a mini-batch. The conventional triplet loss in Eqs. (3) and (4) is a special case of our multi-task view-invariant CNN where $\alpha = 0$, $\beta = 0$ and $\gamma = 0$. Each triplet example is assigned with equal importance. The following section will introduce how to impose weights on the triplet examples to compute the weighted triplet loss L_t (as shown in Fig. 2) and mine the hard examples in a stage-wise procedure.

3.3. Attribute-based hard triplet weighting strategy

Existing works like Wang and Gupta (2015); Schroff et al. (2015) mainly sample the hard examples based on the low-level feature distance, which do not consider the high-level knowledge, *e.g.*, semantic-level distance. Moreover, the hard negatives are treated equally without considering their learning difficulties within different triplets.

To further learn a semantic and discriminative shoe feature representation, we design a novel attribute-based hard triplet weighting and mining scheme. Moreover, we define the difficult levels of the triplet examples in a progressive way and select the hard negatives as well as hard anchors in a stage-wise manner. The whole procedure can be summarized into the following three stages:

Stage 1: Random Selection In the first 10 epoches, the MTV-CNN+ is trained with the multi-task loss J_0 in Eq. (11) and the triplet examples (images in both the green and red rectangles in Fig. 2) are assigned with equal weight. The anchors as well as the negative examples in the triplets are randomly chosen from the training set X. Here the number of stopping epoch for the Stage 1 model training is empirically determined based on the early-stopping performance on the validation set.

Stage 2: Hard Negative Weighting and Sampling The feature representations for the training shoe items are computed through the

Please cite this article as: H. Zhan, B. Shi, L.-Y. Duan et al., DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval. Computer Vision and Image Understanding (2019), https://doi.org/10.1016/j.cviu.2019.01.001.

learnt embedding network $f_{\mathbf{w}}(\cdot)$. For the ith item, its online shop image in the left profile view is viewed as the representation for that item denoted as $\mathbf{z}_i$ and $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_N\} \in \mathbb{R}^{N\times 512}$ indicates the embedding feature matrix for all the N shoe items in the training set. We use $\tilde{\mathbf{z}} = [\|\mathbf{z}_1\|_2^2, \|\mathbf{z}_2\|_2^2, \dots, \|\mathbf{z}_N\|_2^2]^\top$ to denote the column vector of squared norms of all vectors in $\mathbf{Z}$ and the pairwise Euclidean distance matrix $\mathbf{D}_0$ can be efficiently computed.

Based on $\mathbf{D}_0$, the neighboring affinity map $\tilde{\mathbf{L}}_0$ are computed with each entry $\tilde{\mathbf{L}}_0(i,j)$ representing the jth closest neighboring item to the particular ith item in the embedding space. For example, $\tilde{\mathbf{L}}_0(100,2)$ is the exact shoe item that is closest to the 100th shoe item in the Euclidean space apart from itself. To ensure the sample diversity without the degradation of the performance, empirically, the top 40% closest neighboring items (for each ith item) of the affinity map $\tilde{\mathbf{L}}_0$ are chosen as the *Rate-1 hard negative pool* $\mathbb{N}_1$ (images in the green rectangles on the right part of Fig. 2).

The rationale behind the proposed attribute-based weighting scheme is that, if the anchor and negative are close in the feature level but far away in the semantic level, then the particular triplet is penalized with higher weight. To this end, we incorporate the hierarchical property of semantic attribute labels for hard triplet weighting.

Fig. 4 demonstrates the hierarchical tree-structured grouping of the multi-class "Toe Shape" attribute. It is composed of three levels, *Level-0*, *Level-1* and *Level-2*, which indicate the similarity in a hierarchical manner. Here for the *Level-0* attribute type, the *Level-1* attributes are grouped semantically close attributes that have similar physical structure according to the *Level-2* attributes. For a comparison pair (i,j), the attribute-based weighting matrix $\mathbf{W}_a$ in the semantic space is computed as follows:

$$\mathbf{W}_a(i,j) = \sum_{m=1}^{M_p} d_a^m(a_i^m, a_e^m), \text{with } e = \tilde{\mathbf{L}}_0(i,j), \tag{12}$$

where the subscript e indicates the jth closest neighboring item, M_p is the number of part-aware semantic attributes and d_a^m is the pairwise distance of the mth attribute in the semantic space, defined as follows:

$$d_a^m(a_i^m, a_e^m) = \begin{cases} 0 & \text{if } [a_i^m]_1 = [a_e^m]_1, [a_i^m]_2 = [a_e^m]_2 \\ w_1 & \text{if } [a_i^m]_1 = [a_e^m]_1, [a_i^m]_2 \neq [a_e^m]_2 \\ w_2 & \text{if } [a_i^m]_1 \neq [a_e^m]_1, [a_i^m]_2 \neq [a_e^m]_2. \end{cases} \tag{13}$$

If the ith and eth item share the *Level-1* attribute, then $[a_i^m]_1 = [a_e^m]_1$; Otherwise, $[a_i^m]_1 \neq [a_e^m]_1$. For the *Level-2* attribute, it follows the same definition.

For the non part-aware attribute like color, we do not consider the *Level-2* attribute because empirically we find that the learnt feature embedding is good at capturing the color than the part shape. Thus we impose smaller penalty to compute the color attribute-based distance $d_a^{\tilde{m}}$, defined as below:

$$d_a^{\tilde{m}}(a_i^{\tilde{m}}, a_e^{\tilde{m}}) = \begin{cases} 0 & \text{if } [a_i^{\tilde{m}}]_2 = [a_e^{\tilde{m}}]_2 \\ w_3 & \text{if } [a_i^{\tilde{m}}]_1 \neq [a_e^{\tilde{m}}]_1. \end{cases} \tag{14}$$

Based on the weighting matrix $\mathbf{W}_a$, the attribute-based weighted triplet loss L_t over the *Rate-1 hard negative pool* $\mathbb{N}_1$ is expressed as:

$$L_t(x_i^p, x_i^a, x_j^n) = \mathbf{W}_a(i,j) L(x_i^p, x_i^a, x_j^n),\ i \in X, j \in \mathbb{N}_1, \tag{15}$$

and the multi-task loss J_1 is denoted as below and the negatives are sampled from $\mathbb{N}_1$:

$$J_1 = \frac{1}{T} \sum_{i\in X, j\in\mathbb{N}_1}^{T} [L_t(x_i^p, x_i^a, x_j^n) + \alpha L_a + \beta L_s + \gamma L_v(X^o)], \tag{16}$$

Stage 3: Hard Anchor and Negative Selection In the final stage, we not only select the hard negatives but also the hard anchors. Similar to the previous stage, the Euclidean distance matrix $\mathbf{D}_1$ and the neighboring affinity map $\tilde{\mathbf{L}}_1$, the top 40% of the closest shoe neighboring items are selected to form *Rate-2 hard negatives pool* $\mathbb{N}_2$. Following Eq. (12), the distance matrix $\mathbf{W}_a(i,k)$ is re-calculated and updated. The hard anchors are selected based on the attribute inconsistency $\bar{\mathbf{w}}_a(i)$ between each ith shoe item and its neighboring items, which is computed as:

$$\bar{\mathbf{w}}_a(i) = \frac{1}{T_n} \sum_{k=1}^{T_n} \mathbf{W}_a(i,j), \tag{17}$$

where $T_n = \lfloor 0.4N \rfloor$ is the number of hard negative examples per shoe item.

We define the set of hard anchors $\mathbb{P} = \{i | \bar{\mathbf{w}}_a(i) > v_0\}$ as the ensemble of shoe items with $\bar{\mathbf{w}}_a(i)$ larger than the median value over all the training shoe items, where v_0 computes the median value of $\bar{\mathbf{w}}_a$. Please refer to Fig. 2 that the anchor a_2 is a hard anchor to be used by Stage 3, while a_1 is not.

Then the weighted triplet loss over the set of $\mathbb{N}_2$ and $\mathbb{P}$ is:

$$L_t(x_i^p, x_i^a, x_j^n) = \mathbf{W}_a(i,j) L(x_i^p, x_i^a, x_j^n),\ i \in \mathbb{P}, j \in \mathbb{N}_2. \tag{18}$$

The multi-task loss J_2 is denoted as below and the negatives are sampled from $\mathbb{N}_2$ and anchors from $\mathbb{P}$:

$$J_2 = \frac{1}{T} \sum_{i\in\mathbb{P}, j\in\mathbb{N}_2}^{T} [L_t(x_i^p, x_i^a, x_j^n) + \alpha L_a + \beta L_s + \gamma L_v(X^o)], \tag{19}$$

For back propagation of the network parameters, the gradients of L_t with respect to the output of the softmax layer (l_+, l_-) is computed as:

$$\frac{\partial L_t}{\partial l_+} = \sqrt{\mathbf{W}_a} \frac{\partial L}{\partial l_+}, \qquad \frac{\partial L_t}{\partial l_-} = \sqrt{\mathbf{W}_a} \frac{\partial L}{\partial l_-}. \tag{20}$$

The stochastic gradient descent (SGD) algorithm is utilized to optimize the proposed loss over multiple tasks in a mini-batch. The details of the training strategy is illustrated in Algorithm 1. We stopped the training after about 40 epoches and the training curve is almost converged, from which we obtain the discriminative and semantic shoe feature representation.

Algorithm 1 Training of MTV-CNN+

Input: Training set $X = \{x_i\}_{i=1}^N$, the attribute labels $\{a_i\}_{i=1}^N$
Output: The parameters **w** of MTV-CNN+.
Initialization: The parameters **w** from the CNN model pre-trained on ImageNet;
for epoch = 1 to 10 **do**
  Forward randomly sampled triplets (x_i^a, x_i^p, x_j^n), positive bag $X_{i,o}$ and attribute labels a_i into the network according to **Stage 1** in Section 3.3;
  Compute the multi-task loss J_0 according to Eq. (11) and back-propagate the gradients to update **w**;
end for
Compute $\mathbb{N}_1$ and $\mathbf{W}_a$ according to **Stage 2**;
while not converged **do**
  Compute the total loss J_1 based on Eq. (16) with hard negatives sampled from $\mathbb{N}_1$ and back-propagate the gradients according to Eq. (20);
end while
Compute $\mathbb{N}_2$, $\mathbf{W}_a$ and $\mathbb{P}$ according to **Stage 3**;
while not converged **do**
  Compute the weighted and total loss J_2 based on Eq. (19) with hard negatives sampled from $\mathbb{N}_2$ and hard anchors from $\mathbb{P}$, then back-propagate the gradients;
end while
return The parameters **w**;

4. Dataset construction

We collect a novel multi-viewpoint street-to-shop scenario shoe dataset (MVShoe) for training and testing of the proposed approach. It is an extension of our preliminary version in Zhan et al. (2017a), which

Please cite this article as: H. Zhan, B. Shi, L.-Y. Duan et al., DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval. Computer Vision and Image Understanding (2019), https://doi.org/10.1016/j.cviu.2019.01.001.

Table 1
Semantic attribute groups in MVShoe dataset.

Group	Attributes
Color	Black, Blue, Coffee, Golden, Gray, Green Bronze, Brown, Multi-Color, Orange, Purple, Red Rice White, Silver, White, Yellow, Pink Multi-color, Khaki, Pearl Blue
Toe shape	Round Toe, Pointy Toe, Square Toe, Peep Toe
Heel shape	Wedge, Flat, Thick, Block, High-Thin, Wine-Shape Cone-Shape

provides 8021 daily shoe images from the street domain and 5821 shoe pictures from the online domain. Moreover, all the images are collected from Jingdong.com and we only choose the online shop images facing the left side 45 degree view.

The newly collected MVShoe dataset is a larger collection of tagged shoe images and the images are crawled from more shopping websites, such as Amazon.com or 6pm.com. As images from *Amazon* or *6pm* are online clean images with completely white background, we mainly use them as the reference gallery images. We also crawl the meta-data of the shoe item and these tags can be summarized into three types of multi-class semantic attributes, as shown in Table 1. As the proposed MTV-CNN+ largely depends on the attribute labels, we re-annotate these crucial attributes for cleaner annotation. Eventually, we have 14314 and 31048 images from the street and online shop scenario in multiple viewpoints with annotated semantic attributes. Some example images in MVShoe are demonstrated in Fig. 6.

Fig. 6. Examples of cropped street and shop scenarios images in our MVShoe dataset.

5. Experiment settings

5.1. Training and testing data

Our proposed framework consists of two core components: shoe detection and cross-scenario shoe retrieval based on the detected bounding boxes. For the training of shoe detection module, 2292 images are used to learn the Fast R-CNN model, including the real-world and online shop images. Each image (usually with a pair of shoes) is annotated with two ground truth bounding boxes indicating each single shoe of a pair. We follow the same parameter setting as in Zitnick and Dollár (2014) for the input region proposals generated by Edgebox.

For the shoe retrieval procedure, about 3130 shoe items with several daily shoe images and online shop images of different viewpoints (*i.e.*, 24000 images after augmentation) are fed into the proposed MTV-CNN+ for training. Each shoe item has an associated product item ID that can help us to find matched and non-matched pairs. About 30200 online shop shoe images with multiple viewpoints are used as reference gallery and 4401 daily shoe images are used as the queries where each of them has a counterpart in the reference set.

5.2. Implementation details

We adopt the high performance VGG 16 Layers model (Simonyan and Zisserman, 2014) pre-trained on the ImageNet (Russakovsky et al., 2015) for initialization and the experiments are implemented based on Torch (Collobert et al., 2011). For optimization, Stochastic Gradient Descent (SGD) (Bottou, 2012) algorithm is used. The mini-batch size is 8 triplets and the corresponding positive bags (altogether $8 \times 6 = 48$ images). The learning rate is initialized with 10^{-3}, momentum with 0.9 and weight decay of 10^{-4}. Usually, the image are resized to 256×256 and multiple crops with the size 224×224 are extracted for the network training. However, we find that the cropped shoe images lose discriminative details of the part appearances, which have a large impact on the overall appearances of shoes. To address this issue, we keep the aspect ratio of the original image (w, h) and rescale the image such that length of its longer side is 224. Then the scaled shorter side is padded by $224 \times [1 - \frac{\min(w,h)}{\max(w,h)}]$ white pixels. Task-importance parameters are set to $\alpha = 0.05$, $\beta = 0.05$, and $\gamma = 0.05$ to reach the same level of magnitude, which were empirically found to achieve best performance. The main triplet ranking task has the highest weight, and the rest auxiliary tasks all have the same lower weights. For the training of Fast R-CNN, we follow the experimental settings in Girshick (2015) and the positive examples are those cropped images which satisfy IoU[2] > 0.7; otherwise, they are used as negative examples. For the attribute-based weighting parameters in Section 3.3, firstly, we find the optimal values for w_1 and w_2 in the range of $[0, 2]$. Experimentally, we find that when $w_1 = 0.7$ and $w_2 = 1.5$, the retrieval performance achieved the best performance. To study the influence of w_3, with the value of w_1 and w_2 fixed, we search w_3 between 0 and 1, and find that $w_3 = 0.4$ worked best.

5.3. Comparison methods

To analyze the performance of our proposed MTV-CNN+, we compare with a variety of approaches. These approaches can be divided into four groups. The first group (red curve in Fig. 7) extracts the hand-crafted feature for retrieval, including (1) GIST feature (Oliva and Torralba, 2001) with 512-dimension, (2) dense SIFT feature followed by fisher vector encoding (Jegou et al., 2012) with the codebook size set as 64, denoted as DSIFT + Fisher Vector, and (3) Hessian affine detector with SIFT descriptor followed by the VLAD encoding (Arandjelovic and Zisserman, 2013) with the codebook size set as 128, denoted as HessianSIFT + VLAD. The second group (green curve in Fig. 7) exploits the commonly-used deep feature learning methods for the problem of instance retrieval and adapt them on shoes. This group includes the following methods: (1) Pre-trained fc6: deep feature activated from the first fully-connected layer of the VGG 16 layers model (Simonyan and Zisserman, 2014) with the cropped images as input; (2) Pre-trained conv5: deep feature activated from last convolutional layer the VGG 16 layers mode (Simonyan and Zisserman, 2014) followed by the max-pooling operation; (3) Siamese Network (Bell and Bala, 2015): the hinge loss with the siamese network; (4) Triplet Network (Simo-Serra and Ishikawa, 2016): the triplet network with the ratio loss in Eq. (3); The third group (magenta curve in Fig. 7) consists of our previously proposed methods on shoes: (1) Multi-scale Viewpoint Invariant Triplet Network (Zhan et al., 2017d) (MsVI-CNN): the triplet network considers the viewpoint invariance with the WI-CNN trained using the whole images and RP-CNN using the high-quality region proposals; (2) Multi-Task View-invariant CNN (MTV-CNN) (Zhan et al., 2017c): the triplet

[2] IoU is defined as the intersection of the candidate window with the ground truth box divided by the union of them.

Please cite this article as: H. Zhan, B. Shi, L.-Y. Duan et al., DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval. Computer Vision and Image Understanding (2019), https://doi.org/10.1016/j.cviu.2019.01.001.

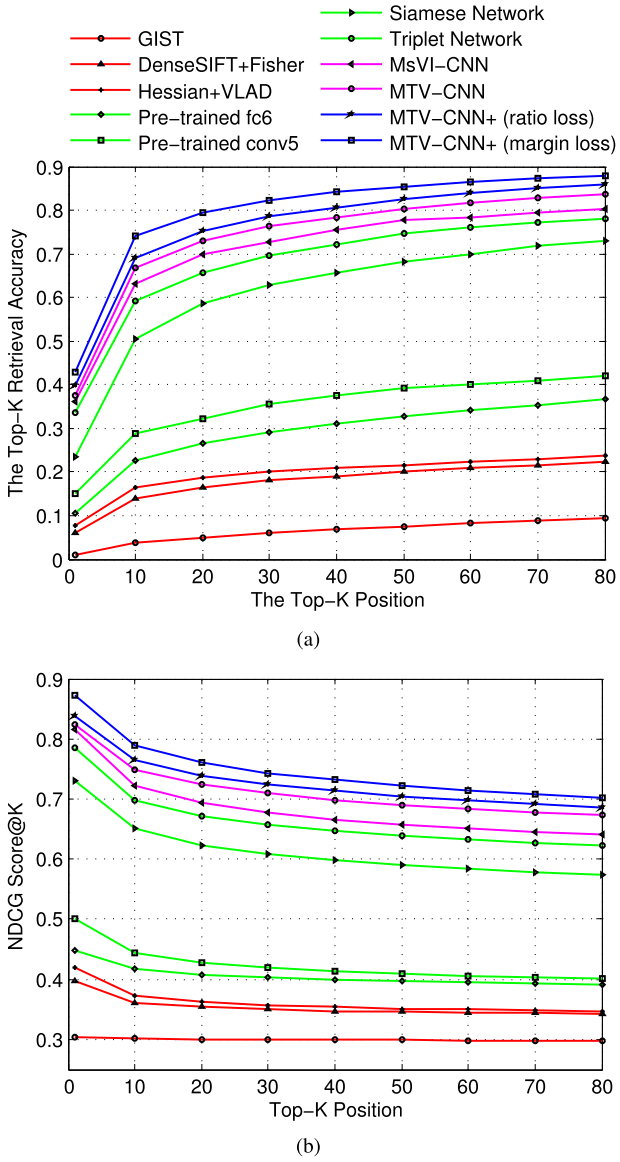

Fig. 7. (a) Top-K retrieval accuracy (b) NDCG@K score with K varying from 1 to 80. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

network with viewpoint invariant loss as well as the attribute prediction loss with two-stage hard negative mining. The last group (blue curve in Fig. 7) contains the proposed Multi-Task View-invariant Convolutional Neural Network (MTV-CNN+) with different types of losses.

5.4. Evaluation protocol

To evaluate the performance of our proposed street-to-shop retrieval system, we adopt two evaluation metrics.

5.4.1. Top-K retrieval accuracy

It computes the percentage of the query images whose top-K retrieval results include the same shoe item as the query images. If the top-K returned results contain the exact match as the query, then it is considered as a successful result; otherwise, it is a failure case.

5.4.2. Normalized Discount Cumulative Gain (NDCG)

NDCG score is used to evaluate the retrieval results in terms of attribute overlapping between the query image and reference gallery, defined as NDCG@K $= \frac{1}{Z}\sum_{k=1}^{K}\frac{2^{r(k)}-1}{\log(1+k)}$, where Z is a normalization term, and $r(k)$ assesses the relevance between the kth ranked reference gallery image and the query image by computing the number of their overlapping attributes in different levels. Note that because some multi-class attribute (*e.g.*, color) are too fine-grained to differentiate the specific classes so that we consider the attribute-level matching in the Level-1 sense. That is to say, if the query image and the retrieved image share the Level-1 similarity as shown in Fig. 4, then they are supposed to be matched for the particular attribute.

Table 2
Top-20 retrieval accuracy on the MVShoe dataset.

Method	Top-20(%)
GIST (Oliva and Torralba, 2001)	5.02
DSIFT + Fisher Vector (Jegou et al., 2012)	16.32
HessianSIFT + VLAD (Arandjelovic and Zisserman, 2013)	18.77
Pre-trained fc6 (Simonyan and Zisserman, 2014)	26.52
Pre-trained conv5 (Simonyan and Zisserman, 2014)	32.27
Siamese Network (Bell and Bala, 2015)	58.53
Triplet Network (Simo-Serra and Ishikawa, 2016)	65.69
Multi-scale Viewpoint Invariant Triplet Network (Zhan et al., 2017d)	69.82
MTV-CNN (Zhan et al., 2017c)	73.69
MTV-CNN+ (ratio loss)	75.25
MTV-CNN+ (margin loss)	**79.32**

5.5. Results

5.5.1. Comparisons with other baselines

Table 2 summarizes the top-20 ($K = 20$) retrieval accuracy of the proposed MTV-CNN+ (ratio loss and margin loss) and other compared methods. MTV-CNN+ achieves the best performances among all the baselines, which verifies the effectiveness of the learnt deep shoe feature representation in a multi-task manner. From the experimental results we find that the deep features (fc6 and conv5) with the pre-trained model outperform the state-of-the-art hand-crafted feature (HessianSIFT + VLAD) by about 7.8% and 13.5%, respectively. This verifies that the generic deep features pre-trained on the ImageNet is transferable to the target shoe dataset and they are also an effective representation for the cross-scenario shoe retrieval task. Furthermore, it can be seen that the retrieval results using the feature activated from the conv5 layer has an improvement of about 5.5% over that extracted from the fc6 layer, which inspires us to utilize the feature extracted from the conv5 layer as the building block for learning the feature embedding network.

We also compare the MTV-CNN+ with several state-of-the-art works that utilize the deep metric learning, Siamese network (Wang et al., 2016), the triplet network with ratio loss (Simo-Serra and Ishikawa, 2016). We can find the retrieval performance has been significantly boosted with the learnt feature representations fine-tuned on the target shoe dataset. Compared to the two-branch siamese network, the three-branch triplet network achieves better performance by about 6.9% (58.53% vs. 65.69%) improvements. One possible reason is that imposing the constraint between the similar and dissimilar image pairs within a triplet of three examples is more effective at capturing the semantic high-level features of shoes.

With the further extensions of the proposed MTV-CNN+ has about 7.6% (ratio loss) and 11.7% (margin loss) improvements over the preliminary version in Zhan et al. (2017d), which demonstrates the effectiveness of the added components in this work (*e.g.*, auxiliary task, weighted triplet loss, hard example mining). Moreover, MTV-CNN+ improves over Zhan et al. (2017c) by about 1.6% (ratio loss) and 5.7% (margin loss), which indicates the importance of style identification and Stage 3 hard anchor selection process. Compared to merely minimizing the color or shape attribute loss, the newly included style loss (the combination of multiple part-aware attributes) is a more strict constraint, which focuses on both of toe and heel shape of the shoes. Moreover, the anchor image selection described in Stage 3 focuses more

Please cite this article as: H. Zhan, B. Shi, L.-Y. Duan et al., DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval. Computer Vision and Image Understanding (2019), https://doi.org/10.1016/j.cviu.2019.01.001.

Table 3
NDCG score on the MVShoe dataset.

Method	NDCG@20
GIST (Oliva and Torralba, 2001)	0.301
DSIFT + Fisher Vector (Jegou et al., 2012)	0.354
DSURF + VLAD (Arandjelovic and Zisserman, 2013)	0.363
Pre-trained fc6 (Simonyan and Zisserman, 2014)	0.408
Pre-trained conv5 (Simonyan and Zisserman, 2014)	0.427
Siamese Network (Bell and Bala, 2015)	0.624
Triplet Network (Simo-Serra and Ishikawa, 2016)	0.672
Multi-scale Viewpoint Invariant Triplet Network (Zhan et al., 2017d)	0.694
MTV-CNN (Zhan et al., 2017c)	0.724
MTV-CNN+ (ratio loss)	0.740
MTV-CNN+ (margin loss)	**0.761**

on those samples with close feature-level distance but large semantic-level distance (subtle details in the localized shoe parts).

We also investigate the performance of our proposed MTV-CNN+ with both the ratio loss and margin loss. The experiments indicate that the proposed network with the margin loss improves over that with ratio loss by about 4.1%. It is possibly because that the hinge loss is capable of making use the similar/dissimilar examples within a triplet that violates the constraint and further updating the network parameters.

The results of NDCG scores are reported in Table 3 when $K = 20$ and we can see that our method is also the best among the compared methods in terms of attribute-level matching, which indicates our proposed network is able to learn the semantic-aware features. It can seen that the NDCG score of MTV-CNN+ with margin loss improves over that of MTV-CNN+ with ratio loss by 0.02 and that of MTV-CNN by 0.04. It demonstrates that the proposed MTV-CNN+ with the margin loss generally better at learning more semantic feature representations.

5.5.2. Evaluation of different returned items

To give a thorough analysis of the advantage of our method over the compared methods, Figs. 7(a) and 7(b) illustrate the top-K retrieval results and NDCG@K score with K varying from $[1, 10, 20, \ldots, 80]$. We can observe that the proposed MTV-CNN+ achieves the best performance over other methods with different settings of K. Not only for its ability to find the exact match, it also returns semantically similar shoes in terms of overlapping attributes to the query. Moreover, the retrieval performance with the feature embedding network has a significant improvement compared to the hand-crafted features and generic deep features. When K reaches to 80, the system is capable of returning the exact match for almost 90% of the daily shoe images.

5.5.3. Qualitative evaluation of the system

To give a more intuitive illustration of the performance of our street-to-shop shoe retrieval system, Fig. 8 presents two daily life query examples and the corresponding top-5 retrieved shoe images. In each example, the results in the first row are generated utilizing the MTV-CNN (Zhan et al., 2017c). The second and third rows are achieved utilizing MTV-CNN+ with the ratio loss and margin loss, respectively. It can be seen that most of the returned results are quite similar as the exemplar query images in the global aspects (*e.g.*, style, color) and local aspects (*e.g.*, decorations, strap, buckle).

For the first example, the results based on the two types of losses find exactly the same shoes as the query in three different viewpoints, which validate the effectiveness of the viewpoint invariance. The returned results by MTV-CNN+ with margin loss (the third row) carry more semantic meaning because the top-5 retrieved images exhibit open structure without side cover, but with straps. What is more, our system is capable of finding the matched shoes with less similar viewpoints (the 5th returned shoe item in the third row) as the query.

For the second example, even though the query shoe image is captured in a top-to-down viewpoint, our system is still able to return visually similar shoes as the query and the ratio loss successfully returns

Fig. 8. Example retrieval results. The green boxes indicate the exactly matched items to the query. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
Comparisons of different CNN architectures on the retrieval performance in terms of Top-20 (T20) and NDCG@20 (N20) accuracy. Here ratio loss and margin loss are simplified as R and M, respectively.

CNN models	T20R/T20M	N20R/N20M
VGG CNN-S	49.35/**54.12**	0.592/0.615
VGG 16 Layers CNN	75.25/**79.32**	0.740/**0.761**

the exact match. Moreover, it can be seen that when compared to MTV-CNN (Zhan et al., 2017c) with only two-stage process, MTV-CNN+ using ratio and margin loss with three-stage procedure increasingly improves the capability of dealing with the fine-grained details such as the flat heel.

5.5.4. Evaluation of different deep architectures

Our proposed MTV-CNN+ is not limited to the VGG 16 Layers CNN. Other architectures of the convolutional neural network can also be fitted into our approach (*e.g.*, VGG CNN-S). Here we also conduct the experiments utilizing CNN-S as shown in Table 4 and analyze the choice of the CNN model to the retrieval performance. The VGG 16 Layers achieves much better results when compared to CNN-S model. This makes sense because usually deeper networks bring better performances. Again, the experimental results with different CNN structures indicate that the margin loss exhibits more capabilities of learning semantic attributes compared to the ratio loss.

5.5.5. Impacts of the auxiliary tasks

To evaluate the individual impact of each auxiliary task, the experiments are designed by removing one auxiliary task and its impact is evaluated on the overall retrieval accuracy as shown in Table 5.

We denote the triplet network alone as TN, the task of attribute prediction as AP, style identification as SS and the viewpoint invariance as VI. The individual tasks are stacked to the framework one by one and the corresponding retrieval performance is reported. We can find

Please cite this article as: H. Zhan, B. Shi, L.-Y. Duan et al., DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval. Computer Vision and Image Understanding (2019), https://doi.org/10.1016/j.cviu.2019.01.001.

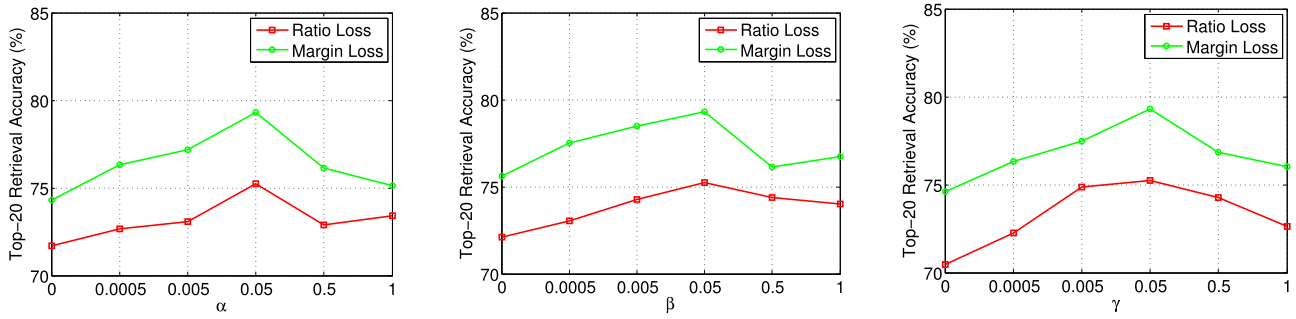

Fig. 9. Analysis of the joint loss weighting parameters: α, β, and γ.

Table 5
Impacts of the auxiliary tasks and each stage to the proposed MTV-CNN+ for different choices of the triplet loss function.

Auxiliary tasks	T20R/T20M	N20R/N20M
TN	65.69/70.46	0.672/0.650
TN + AP	67.96/73.48	0.682/0.717
TN + VI	69.71/71.67	0.678/0.703
TN + SS	68.69/72.64	0.692/0.728
TN + AP + SS	70.48/74.60	0.715/0.737
TN + AP + SS + VI	75.25/**79.32**	0.740/**0.761**
Stage 1	67.05/70.92	0.684/0.712
Stage 1 + Stage 2	73.69/76.10	0.724/0.745
Stage 1 + Stage 2 + Stage 3	75.25/**79.32**	0.740/**0.761**

that the retrieval performance improves the top-20 retrieval accuracy gradually as more auxiliary tasks are added to the framework. The results show that the combination of the auxiliary task modules (*e.g.*, TN + AP, TN + VI and TN + SS) improve over that achieved by merely the baseline module TN, which verifies that each of the task benefits the overall system. The model using TN + AP + SS + VI obtains 1.6% (72.01% vs. 70.48%, ratio loss) and 1.0% (75.64% vs. 74.60%, margin loss) improvement compared to that without VI module. It indicates the effectiveness of the viewpoint invariant loss.

From the NDCG@20 score metric, we can clearly find that the attribute prediction TN + AP and the style identification modules TN + SS are helpful for learning semantic feature representation compared to the viewpoint invariant module TN + VI. By stacking the AP and SS modules together, the retrieval performance has made further improvement.

5.5.6. Impacts of hard example weighting and sampling

The hard triplet weighting and sampling strategy are based on the attribute and performed in a progressive way. Here Stage 2 mainly searches for the hard negatives that the distance between the anchor image and negative example is small, while Stage 3 looks for hard anchor images which is equivalent to mining for hard positive images. Table 5 shows the retrieval performance of different stages utilized. We find that combining Stage 2 and Stage 3 together improves the retrieval performance for the top-20 accuracy by 6.6% and 1.6% (ratio loss) versus 5.2% and 3.2% (margin loss), respectively. It indicates the importance of both hard negative example (Stage 2) and hard anchor mining (Stage 3) in effectively filtering out triplets with limited information. Moreover, we find that both hard negative mining and hard anchor/positive selection have increased NDCG score, which means that the selection of the negative and anchor images is of vital importance to the ultimate goal of learning a semantic feature representation.

5.5.7. Parameter analysis

To further figure out the influence of parameter settings on the performance of our system, we evaluate the impact of the α, β, and γ on the retrieval performance in terms of Top-20 retrieval accuracy, as shown in Fig. 9. It can be seen that the retrieval accuracy curves grow monotonically in [0.0.05] and reach the best performance when $\alpha = 0.05$, $\beta = 0.05$, and $\gamma = 0.05$. When α, β, and γ are set to be 0, which means the model is trained with the attribute prediction loss, style identification loss, and viewpoint invariant loss, respectively, the performances become the worst. For the auxiliary task of attribute prediction with $\alpha = 0.05$, we obtain 4.13% (ratio loss) and 4.72% (margin loss) improvements. For the auxiliary task of style identification with $\beta = 0.05$, we obtain 3.13% (ratio loss) and 3.72% (margin loss) improvements. For the auxiliary task of the viewpoint invariance with $\gamma = 0.05$, we obtain 4.72% (ratio loss) and 4.77% (margin loss) improvements.

6. Conclusion

In this paper, we address the street-to-shop shoe retrieval via an improved multi-task view-invariant triplet network, which embeds the feature of images for the same shoes similar to each other. We show that the network learnt in a multi-task manner is helpful for learning a semantic-aware deep shoe feature representation. Our training strategy is a three-stage process, progressively selects the hard examples, assigns higher weights to penalize the difficult triplets based on the hierarchical properties of the multi-class semantic attributes. The experiments performed on our MVShoe dataset show that, in terms of top-20 accuracy, our proposed method has 13.7% improvement over the commonly used triplet network with the ratio loss. Moreover, not only on shoes, the proposed system could also be generalized to other object types, such as food, cars, *etc*.

There are still several directions for further improvements: 1) we aim to adapt the off-line stage-wise hard example selection to the online alternative, which involves less human interfere; 2) The scale of database is far less than that in the web-scale e-commerce situation. Thus we will increase the number of images to million orders of magnitude and develop the fastindexing methods.

Acknowledgments

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore, supported by the National Research Foundation, Prime Ministers Office, Singapore, under NRF-NSFC grant NRF2016NRF-NSFC001-098. This project was also supported in part by the National Natural Science Foundation of China under Grant 61661146005 as well as National Science Foundation of China under Grant No. 61872012 and No. 61876007.

References

Arandjelovic, R., Zisserman, A., 2013. All about VLAD. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1578–1585.
Babenko, A., Lempitsky, V., 2015. Aggregating local deep features for image retrieval. In: Proc. of International Conference on Computer Vision, ICCV, pp. 1269–1277.

Please cite this article as: H. Zhan, B. Shi, L.-Y. Duan et al., DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval. Computer Vision and Image Understanding (2019), https://doi.org/10.1016/j.cviu.2019.01.001.

Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V., 2014. Neural codes for image retrieval. In: European Conference on Computer Vision, ECCV, pp. 584–599.

Bell, S., Bala, K., 2015. Learning visual similarity for product design with convolutional neural networks. ACM Trans. Graph. 34 (4), 98.

Bottou, L., 2012. Stochastic gradient tricks. In: Neural Networks, Tricks of the Trade, Reloaded. In: Lecture Notes in Computer Science, LNCS 7700, pp. 430–445.

Bui, T., Ribeiro, L., Ponti, M., Collomosse, J., 2017. Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network. Comput. Vis. Image Underst. 164, 27–37.

Carvalho, M., Cadène, R., Picard, D., Soulier, L., Thome, N., Cord, M., 2018. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In: SIGIR.

Collobert, R., Kavukcuoglu, K., Farabet, C., 2011. Torch7: A Matlab-like environment for machine learning. In: BigLearn, NIPS Workshop.

Das Bhattacharjee, S., Yuan, J., Tan, Y.P., Duan, L., 2015. Query-Adaptive logo search using shape-aware descriptors. In: Proceedings of the ACM International Conference on Multimedia, pp. 1155–1158.

Ding, C., Xu, C., Tao, D., 2015. Multi-task pose-invariant face recognition. IEEE Trans. Image Process. 24 (3), 980–993.

Fu, J., Wang, J., Li, Z., Xu, M., Lu, H., 2012. Efficient clothing retrieval with semantic-preserving visual phrases. In: Proc. of Asian Conference on Computer Vision, ACCV, pp. 420–431.

Girshick, R., 2015. Fast r-cnn. In: Proceedings of International Conference on Computer Vision, ICCV, pp. 1440–1448.

Gkioxari, G., Hariharan, B., Girshick, R., Malik, J., 2014. R-cnns for pose estimation and action detection, arXiv preprint arXiv:1406.5212.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., Chen, T., 2018. Recent advances in convolutional neural networks. Pattern Recognit. 77, 354–377.

Jegou, H., Perronnin, F., Douze, M., Sánchez, J., Perez, P., Schmid, C., 2012. Aggregating local image descriptors into compact codes. IEEE Trans. Pattern Anal. Mach. Intell. 34 (9), 1704–1716.

Jiang, Y., Meng, J., Yuan, J., Luo, J., 2015. Randomized spatial context for object search. IEEE Trans. Image Process. 24 (6), 1748–1762.

Jiang, S., Wu, Y., Fu, Y., 2016. Deep bi-directional cross-triplet embedding for cross-domain clothing retrieval. In: Proceedings of the 2016 ACM on Multimedia Conference, ACM MM. ACM, pp. 52–56.

Kiapour, M.H., Yamaguchi, K., Berg, A.C., Berg, T.L., 2014. Hipster wars: Discovering elements of fashion styles. In: p. 472–488. : Proc. of European Conference on Computer Vision, ECCV, pp. 472–488.

Li, Y., Tian, X., Liu, T., Tao, D., 2015. Multi-Task model and feature joint learning. In: IJCAI, pp. 3643–3649.

Liu, S., Feng, J., Domokos, C., Xu, H., Huang, J., Hu, Z., Yan, S., 2014. Fashion parsing with weak color-category labels. IEEE Trans. Multimed. 16 (1), 253–265.

Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X., 2016a. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1096–1104.

Liu, Z., Luo, P., Wang, X., Tang, X., 2015. Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision, ICCV, pp. 3730–3738.

Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., Yan, S., 2012. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 3330–3337.

Liu, H., Tian, Y., Yang, Y., Pang, L., Huang, T., 2016b. Deep relative distance learning: Tell the difference between similar vehicles. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 2167–2175.

Liu, J., Zha, Z.J., Tian, Q., Liu, D., Yao, T., Ling, Q., Mei, T., 2016c. Multi-Scale triplet CNN for person re-identification. In: Proc. of ACM on Multimedia Conference, ACM MM, pp. 192–196.

McLaughlin, N., del Rincon, J.M., Miller, P.C., 2017. Person reidentification using deep convnets with multitask learning. IEEE Trans. Circuits Syst. Video Technol. 27 (3), 525–539.

Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vis. 42 (3), 145–175.

Radenović, F., Tolias, G., Chum, O., 2016. CNN image retrieval learns from BoW: Unsupervised fine-tuning with hard examples. In: European Conference on Computer Vision. Springer, pp. 3–20.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Salvador, A., Hynes, N., Aytar, Y., Marin, J., Ofli, F., Weber, I., Torralba, A., 2017. Learning cross-modal embeddings for cooking recipes and food images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 815–823.

Simo-Serra, E., Ishikawa, H., 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, CVPR, pp. 298–307.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.

Song, J., Song, Y.Z., Xiang, T., Hospedales, T.M., Ruan, X., 2016. Deep multi-task attribute-driven ranking for fine-grained sketch-based image retrieval. In: Proc. of British Machine Vision Conference, BMVC. vol. 1, p. 3.

Wang, X., Gupta, A., 2015. Unsupervised learning of visual representations using videos. In: Proc. of International Conference on Computer Vision, ICCV, pp. 2794–2802.

Wang, F., Kang, L., Li, Y., 2015a. Sketch-based 3D shape retrieval using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1875–1883.

Wang, Y., Li, S., Kot, A.C., 2015b. DeepBag: Recognizing handbag models. IEEE Trans. Multimed. 17 (11), 2072–2083.

Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y., 2014. Learning fine-grained image similarity with deep ranking. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1386–1393.

Wang, X., Sun, Z., Zhang, W., Zhou, Y., Jiang, Y.G., 2016. Matching user photos to online products with robust deep features. In: ACM on International Conference on Multimedia Retrieval, ACM MM, pp. 7–14.

Yamaguchi, K., Kiapour, M.H., Ortiz, L.E., Berg, T.L., 2015a. Retrieving similar styles to parse clothing. IEEE Trans. Pattern Anal. Mach. Intell. 37 (5), 1028–1040.

Yamaguchi, K., Okatani, T., Sudo, K., Murasaki, K., Taniguchi, Y., 2015. Mix and match: Joint model for clothing and attribute recognition. In: Proc. of British Machine Vision Conference, BMVC, pp. 51–51,

Yu, Q., Liu, F., Song, Y.Z., Xiang, T., Hospedales, T.M., Loy, C.C., 2016. Sketch me that shoe. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 799–807.

Yu, Q., Yang, Y., Song, Y.Z., Xiang, T., Hospedales, T., 2015. Sketch-a-net that beats humans, arXiv preprint arXiv:1501.07873.

Yue-Hei Ng, J., Yang, F., Davis, L.S., 2015. Exploiting local features from deep networks for image retrieval. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops, ICCVW, pp. 53–61.

Zagoruyko, S., Komodakis, N., 2017. Deep compare: A study on using convolutional neural networks to compare image patches. Comput. Vis. Image Underst. 164, 38–55.

Zhan, H., Shi, B., Kot, A.C., 2017a. Cross-domain shoe retrieval using a three-level deep feature representation. In: International Symposium on Circuits and Systems, ISCAS.

Zhan, H., Shi, B., Kot, A.C., 2017b. Cross-domain shoe retrieval with a semantic hierarchy of attribute classification network. IEEE Trans. Image Process. 26 (12), 5867–5881.

Zhan, H., Shi, B., Kot, A.C., 2017c. Street-to-shop shoe retrieval. In: Proc. of British Machine Vision Conference, BMVC.

Zhan, H., Shi, B., Kot, A.C., 2017d. Street-to-shop shoe retrieval with multi-scale viewpoint invariant triplet network. In: Proc. of Internation Conference on Image Processing, ICIP.

Zhang, Z., Luo, P., Loy, C.C., Tang, X., 2014. Facial landmark detection by deep multi-task learning. In: Proc. of European Conference on Computer Vision, ECCV, pp. 94–108.

Zhang, C., Zhang, Z., 2014. Improving multiview face detection with multi-task deep convolutional neural networks. In: Proc. of IEEE Winter Conference on Applications of Computer Vision, WACV, pp. 1036–1041.

Zhang, X., Zhou, F., Lin, Y., Zhang, S., 2016. Embedding label structures for fine-grained feature representation. In: Proc. of IEEE Conference on Computer Vision and Pattern Recognition, CVPR, pp. 1114–1123.

Zitnick, C., Dollár, P., 2014. Edge boxes: Locating object proposals from edges. In: Proc. of European Conference on Computer Vision, ECCV, pp. 391–405.

Please cite this article as: H. Zhan, B. Shi, L.-Y. Duan et al., DeepShoe: An improved Multi-Task View-invariant CNN for street-to-shop shoe retrieval. Computer Vision and Image Understanding (2019), https://doi.org/10.1016/j.cviu.2019.01.001.