Street-to-Shop Shoe Retrieval

Huijing Zhan[1]
zh0069ng@e.ntu.edu.sg
Boxin Shi[2]
boxin.shi@aist.go.jp
Alex C. Kot[1]
eackot@ntu.edu.sg

[1] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore

[2] Artificial Intelligence Research Center National Institute of AIST, Japan

Abstract

In this paper we aim to find exactly the same shoes from the online shop (shop scenario), given a daily shoe photo (street scenario). There are large visual differences between the street and shop scenario shoe images. To handle the discrepancy of different scenarios, we learn a feature embedding for shoes via a multi-task view-invariant convolutional neural network (MTV-CNN), the feature activations of which reflect the inherent similarity between any two shoe images. Specifically, we propose a new loss function that minimizes the distances between images of the same shoes captured from different viewpoints. To train the proposed MTV-CNN in a more efficient way, an attribute-based hard example weighting and mining strategy is developed to smartly select the hard negative examples. To evaluate the performance of the proposed MTV-CNN, we collect a large-scale multi-view shoe dataset with semantic attributes (MVShoe) from the daily life and online shopping websites. Experiments on the dataset demonstrate the effectiveness of our proposed method.

1 Introduction

In recent years, due to the huge profits from the online fashion market, there is increasing attention in the visual fashion analysis, including clothing retrieval [9], parsing [10, 11], handbag recognition [20], *etc*. However, visual research on shoes is still not well-explored while of equal importance in terms of necessity in real life and great potentials the online shoe market has brought. The following occasion often occurs in our daily life: when we are walking on the street and see a pair of beautiful shoes displayed on the shop window or worn on others' feet, we would like to find the exact same shoes from the online store. As several words are far from enough to depict the overall appearances of their specified shoe item, it is essential to develop a visual shoe retrieval system. In this paper, we are developing such a system, namely the street-to-shop shoe retrieval, where the goal is to return exactly the same shoe item according to a daily shoe photo.

Designing such a system is highly challenging in three aspects: I) cross-domain differences. It can be seen from the left of Fig. 1 that even the exactly matched shoe items (image pairs in the dashed grid) have large visual discrepancies in the background, viewpoint, scale, illumination, *etc*; II) subtle differences in appearances. The shoes belonging to the same style might differ from each other in fine-grained details as shown in the top row on the right

Fig. 1: Illustrations of main challenges in the street-to-shop shoe retrieval. Challenge I) cross-domain differences; Challenge II) subtle differences in appearances; Challenge III) viewpoint variation for the same shoe item.

of Fig. 1; III) viewpoint variation for the same shoe item. Shoes are usually displayed in a variety of views as demonstrated in the last row on the right of Fig. 1. Although the system is capable of searching for the exactly matched online store images with similar viewpoint as that of the query, it might still fail to find the same item with a less similar view. To deal with the aforementioned challenges, we need to learn an efficient feature embedding to 1) reduce the feature distance between the same shoe in different scenarios, 2) distinguish fine-grained differences, and 3) reduce ambiguous feature representation from different views.

In this paper, we propose a novel multi-task view-invariant CNN (MTV-CNN) to learn such feature embedding for shoes and our overall framework is illustrated in Fig. 2. The triplet examples are fed forward into the proposed MTV-CNN to minimize the distances of features of the same shoes and enlarge that of different shoes. To capture the local fine-grained details of the shoes and pull closer the feature embeddings of different view images in the positive bag, we incorporate another two auxiliary tasks, attribute prediction (L_a) and view invariance (L_v). In order to learn a discriminative feature representation that differentiates shoes with similar low-level features but with different high-level semantics, we develop an attribute-based hard example weighting and sampling strategy to smartly select the hard negative images in the triplet examples. More specifically, the triplets are weighted by the hierarchical properties of the multi-class attribute and higher weights are assigned to harder examples. Finally, a semantic-aware view-invariant shoe representation is obtained from the proposed MTV-CNN.

Our contributions can be summarized as the following: 1) we develop a multi-task view-invariant triplet network that jointly minimizes the distance between the same shoe in different scenarios while maximizes that between different shoes, e.g., triplet network makes anchor and positive images of the same shoe from different scenarios have closer feature distance (conquering challenge I); 2) we incorporate the high-level semantic attributes and design an attribute-based hard example (e.g., image pairs with similar feature but large attribute distance) weighting and mining strategy to differentiate the subtle differences in appearances (conquering challenge II); 3) we propose a view-invariant loss that aims to minimize the feature distances of different view images for the same shoes and achieve an view-invariant representation (conquering challenge III).

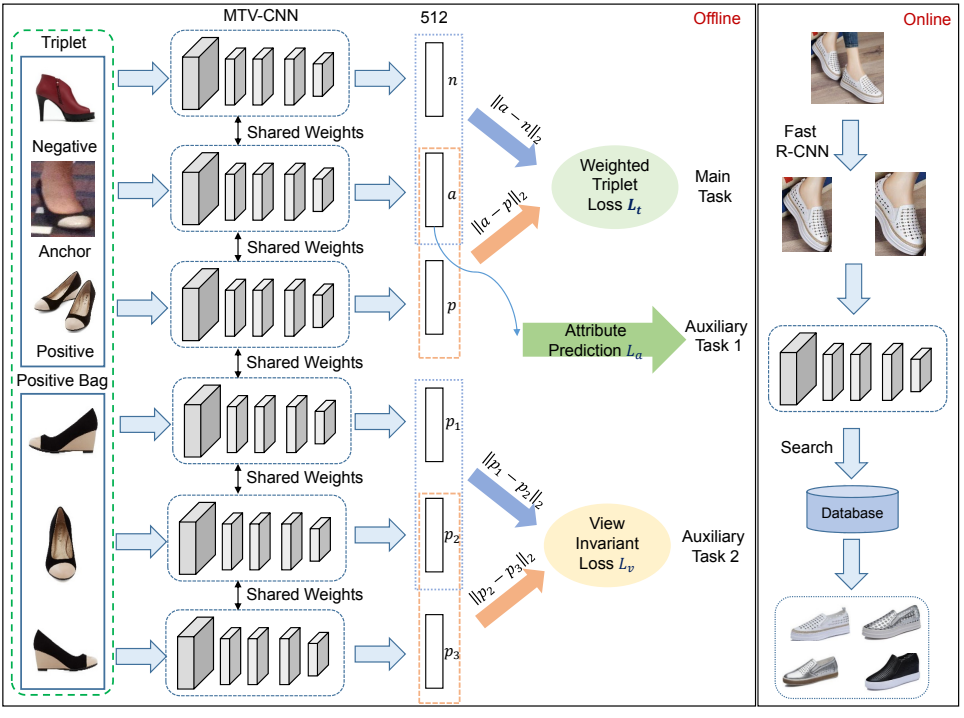

Fig. 2: The online and off-line modules for the schematic view of our street-to-shop shoe retrieval method. The proposed MTV-CNN consists of 6 parallel networks with shared parameters, the input of which contains the triplet examples (negative, anchor, and positive images) and different view images of the same shoe item (positive bag).

2 Related Work

Deep Convolutional Neural Network for Metric Learning: Recently, researchers have begun to integrate the metric learning into the framework of CNN, namely Siamese network and the triplet network. The former consists of two symmetric branches that take the image pairs as the inputs. It has been successfully applied to the problem of product design [2], sketch-based image retrieval [17], image-patch matching [21], *etc*. The triplet network is shown to be superior to the Siamese network [7] with three images as the input. Wang *et al* [18] built a multi-scale triplet network for learning the image similarity. The recent work by Zhang *et al.* [23] investigated the structure of the class label for fine-grained feature representation and Liu *et al.* [8] proposed a Coupled Cluster Loss (CCL) to pull the positive vehicle images closer. However, there still remains one common problem of these approaches: how to efficiently select the hard examples in the triplets. Existing hard example mining strategy involved in [14] only computes the feature-level distance to differentiate the hard negatives and doesn't fully exploit the semantic attribute information. In this work, the hard examples are evaluated in both the feature-level and attribute-level. Furthermore, we develop an attribute-based hard negative weighting and sampling strategy.

Multi-Task Learning: Multi-Task Learning (MTL) is about jointly optimizing several tasks, which has achieved better performances and improved learning efficiency, compared

to solving multiple tasks independently. Its effectiveness has been proven in a variety of computer vision areas such as face landmark localization [24], face detection [22], *etc*. Nevertheless, it is still new in the fashion area, especially on shoes, possibly due to the absence of accurate attribute labels. The most relevant work is [12], which proposed to learn the clothes feature by jointly optimizing the clothing landmark detection and attribute prediction. Our work differs from [12] in the sense that the retrieval of shoes has its unique challenge due to the various viewpoints of shoes in the fashion image. To address this problem, we design a view-invariant loss over the multi-task framework and obtain a view-invariant semantic shoe feature representation.

3 Street-to-Shop Shoe Retrieval

To learn a semantic-aware and view-invariant shoe feature representation, we fully make use of the attribute-level annotations of shoes and leverage the multi-task learning to train the triplet network with two auxiliary tasks: 1) attribute prediction (L_a); 2) preserving the viewpoint invariance within same shoes (view-invariant loss L_v). The main task is to the pull images of the same shoe item closer while push images of different shoes far from each other (weighted triplet loss L_t).

3.1 Triplet Network

The triplet network aims to learn an effective embedding and a similarity metric so that images of the same shoe item are clustered and images of different shoes are pushed away. Let us denote a training set as $X = \{x_i^k | i = 1,2,...,N\}$, where N is the total number of shoe items and x_i^k indicates the k-th image of the i-th 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_j^n (negative) belongs to another item. Here the anchor is randomly selected from the training shoe item set X, the positive image shares the same shoe item ID as the anchor while the negative image belongs to a different shoe item. We wish to learn a deep embedding network $f_{\mathbf{w}}(\cdot)$, which minimizes Euclidean distances $d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_i^p))$ of the matched pairs (x_i^a, x_i^p) while maximizes the distances $d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_j^n))$ between the non-matched pairs (x_i^a, x_j^n). Inspired by [15], we normalize the feature distances to get 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 $[0,1]$.

$$l_+ = \frac{\exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_i^p))\right)}{\exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_i^p))\right) + \exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_j^n))\right)}, \tag{1}$$

and

$$l_- = \frac{\exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_j^n))\right)}{\exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_i^p))\right) + \exp\left(d(f_{\mathbf{w}}(x_i^a), f_{\mathbf{w}}(x_j^n))\right)}. \tag{2}$$

Thus the triplet loss L to be minimized is defined as:

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

Compared to the triplet network with the contrastive loss [14], the ratio triplet loss in Eq. (3) is less sensitive to the margin parameter, which has a large impact on the convergence

speed and optimization of the network. Here different triplet examples are assigned with equal weights and the proposed attribute-based hard example weighting strategy is introduced in Section 3.4.

3.2 Attribute Prediction

Because 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 each shoe item. Through this work, we deal with the multi-class attributes and M is the number of the multi-class attributes. Let $a_i = \{a_i^1, a_i^2, ...a_i^m, ..., a_i^M\}$ represent the attribute labels for the i-th shoe item and $a_i^m = \{a_i^{m,1}, a_i^{m,2}, ..., a_i^{m,G_m}\}$ denotes a vector indicating the attribute membership for the m-th multi-class attribute. Here G_m is the number of attributes in the m-th multi-class attribute. As the number of training data with specific attribute labels varies, the weighted softmax loss for the task of m-th attribute prediction L_m is formulated as follows:

$$L_m = -\sum_{g=1}^{G_m} \bar{w}_{m,g} \mathbf{1}(a_i^{m,g}) \log p_i^{m,g}, \quad m = 1, 2, ..., M, \tag{4}$$

where $p_i^{m,g} = \frac{\exp(z_i^{m,g})}{\sum_{g=1}^{G_m} \exp(z_i^{m,g})}$ indicates the probability of assigning x_i to the g-th class in the m-th multi-class attribute and $z_i^{m,g}$ is the output from the softmax layer of g-th node. Here $\mathbf{1}(\cdot)$ is the indicator function. Take the "Toe Shape" (m-th multi-class attribute) as an example, $p_i^{m,g}$ is to compute the probability of the given shoe item x_i belonging to the "Pointy Toe" class (g-th class). The weights are inversely proportional to the number of training samples defined as $\bar{w}_{m,g} = \frac{\frac{1}{N_{m,g}}}{\sum_{g=1}^{G_m} \frac{1}{N_{m,g}}}$, where $N_{m,g}$ is the number of training samples with the attribute label g in the m-th multi-class attribute. Thus, the attribute prediction loss over all the M multi-class attributes is computed as below:

$$L_a = \frac{1}{M} \sum_{m=1}^{M} L_m. \tag{5}$$

3.3 View Invariance

To deal with the challenge III in Section 1, we design a novel view-invariant loss L_v and our goal is to minimize L_v so that the shop-scenario images in varied viewpoints are drawn closer. Assume that the i-th shoe is depicted by $X_{i,s} = \{x_{i,s}^1, ..., x_{i,s}^{N_s}\}$, which contains N_s street shoe images, and $X_{i,o} = \{x_{i,o}^1, ..., x_{i,o}^{N_o}\}$ which contains N_o online shoe photos. Here we treat the online shoe photos for the same shoes as a positive bag and 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 $(x_{i,o}^j, x_{i,o}^k)$ in the online shoe photo set $X_{i,o}$, which can be written as:

$$L_v(X_{i,o}) = \frac{1}{2 \times n_d} \sum_{j,k}^{n_d} d^2(f(x_{i,o}^j), f(x_{i,o}^k)), \tag{6}$$

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 store shoe images inside the positive

bag into the batch for computation. However, due to the memory bottle neck 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. Therefore, the multi-task loss function J over a mini-batch to be minimized is a weighted combined loss of the above mentioned main task and three auxiliary tasks:

$$J_0 = \frac{1}{T}\sum_{i,j\in X}^{T}[L(x_i^p, x_i^a, x_j^n) + \alpha_v L_v(X^o) + \beta_a L_a], \tag{7}$$

where α_v, β_a, are the weights to indicate the importance of each auxiliary tasks and T is the number of triplet examples in a mini-batch. What is worth to mention is that the conventional triplet loss in Eq. (3) assigns the equal importance to each triplet example, and the following Section 3.4 will introduce how to impose the weights on the triplet examples to compute the weighted triplet loss L_t (as shown in Fig. 2).

3.4 Attribute-based Hard Triplet Weighting Strategy

To further learn a semantic and discriminative shoe feature representation, we design a novel attribute-based hard triplet weighting and mining scheme. 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. (7) and the triplet examples are assigned with equal weight. The anchor as well as the negative examples in the triplets are randomly chosen from the training set X.

Stage 2: Hard Negative Weighting and Sampling. After the step 1 of the random selection, the training shoe items are forward through the learnt embedding network $f_{\mathbf{w}}(\cdot)$ and the feature representation for the training shoes is denoted as $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_N\} \in \mathbf{R}^{N\times 512}$. Then the pairwise Euclidean distance matrix $\mathbf{D}_0$ and the neighboring affinity map $\tilde{\mathbf{L}}_0$ can be computed. Each entry $\tilde{\mathbf{L}}_0(i,j)$ represents the j-th closest neighboring item to the particular i-th item in the embedding space. The top 40% closest neighboring items (for each i-th item) of the affinity map $\tilde{\mathbf{L}}_0$ are chosen as the *Rate-1 hard negative pool* $\mathbf{N}_1$, where the negatives are sampled from.

The rational behind our attribute-based weighting scheme is that if the anchor and negative in the triplet have less resemblance in semantic attributes but are close in the feature space, then we rate this triplet with a higher level of difficulty. In other words, the particular triplet is imposed with higher weight. Fig. 3 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*. For the comparison pair (i,j), the attribute-based weighting matrix W_a in the semantic space is computed as follows:

$$\mathbf{W}_a(i,j) = \sum_{m=1}^{M_p} d_a^m(a_i^m, a_e^m), \tag{8}$$

where the lower index $e = \tilde{\mathbf{L}}_0(i,j)$ indicates the j-th closest neighboring item, M_p is the number of part-aware semantic attributes and d_a^m is the pairwise distance of the m-th attribute in the semantic space, defined as follows:

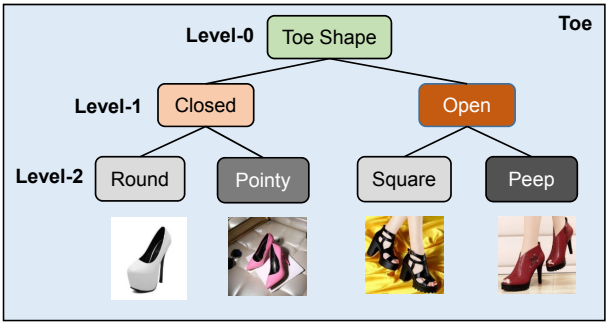

Fig. 3: Hierarchical grouping of the tree-structured semantic shoe attributes for "Toe Shape".

$$d_a^m(a_i^m, a_e^m) = \begin{cases} 0 & \text{if } [a_i^m]_1 = [a_e^m]_1, [a_i^m]_2 = [a_e^m]_2, \\ w_1 & \text{if } [a_i^m]_1 = [a_e^m]_1, [a_i^m]_2 \neq [a_e^m]_2, \\ w_2 & \text{if } [a_i^m]_1 \neq [a_e^m]_1, [a_i^m]_2 \neq [a_e^m]_2. \end{cases} \tag{9}$$

If the i-th and e-th 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. Based on the weighting matrix W_a, the attribute-based weighted triplet loss L_t over the *Rate-1 hard negative pool* $\mathbf{N}_1$ is expressed as:

$$L_t(x_i^p, x_i^a, x_j^n) = \mathbf{W}_a(i,j) L(x_i^p, x_i^a, x_j^n),\ i \in X, j \in \mathbf{N}_1 \tag{10}$$

4 Dataset Construction

We collect a novel large-scale multi-view shoe dataset (MVShoe) with online-offline image pairs for training and evaluation of our proposed approach. The images are crawled from several shopping websites (*e.g.*, *Jingdong*[1], *Amazon*[2], and *6pm*[3]). As the images from *Amazon* or *6pm* are online clean photos with completely white background, we mainly use them as the reference gallery photos. Apart from the images, we also crawl the associated metadata descriptions of the shoe items and these tags can be summarized into 31 types of semantic shoe attributes (*e.g.*, color, toe shape, heel shape, *etc*). As the proposed MTV-CNN largely depends on the attribute labels, we re-annotate these crucial attributes for cleaner annotation. Eventually, we collect about 9800 and 31050 images from the street and online shop scenarios in multiple viewpoints with annotated semantic attributes [4]. Some example images in MVShoe are demonstrated in Fig. 4.

5 Implementation Details

We adopt the high performance VGG 16 Layers model [16] pre-trained on the ImageNet [13] for initialization because the filters with small receptive field in VGG 16 Layers model

[1] https://www.jd.com/
[2] https://www.amazon.com/
[3] https://www.6pm.com/
[4] The MVShoe dataset is available for download from https://sites.google.com/view/deepshoe

Fig. 4: Examples of cropped street and shop scenarios images in our MVShoe dataset.

capture fine-grained appearances of shoes, and its output of last convolutional layer makes semantically-close parts have close distance, both of which benefit the shoe retrieval process. And the experiments are implemented based on Torch [4]. For optimization, Stochastic Gradient Descent (SGD) [3] algorithm is used for back propagation and the mini-batch size is 8 triplets and the corresponding 8 positive bags ($8 \times 6 = 48$ images in total). The learning rate is initialized at 10^{-3}, momentum of 0.9 and weight decay of 10^{-4}. We set $\alpha_v = 0.05$, $\beta_a = 0.05$ in our experiments. For the training of Fast R-CNN, we follow the experimental settings in [5] and the positive examples are those cropped images which satisfy IoU[5] > 0.7; otherwise, they are used as negative examples. The attribute-based weighting parameters in Eq. (9) are experimentally set as $w_1 = 0.7, w_2 = 1.5, w_3 = 0.4$.

6 Experiment Settings

6.1 Training and Testing Data

Our proposed framework consists of two core components: shoe detection and cross-scenario shoe retrieval. 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 photos. 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 [5] and the input region proposals are generated by Edgebox [25]. For the shoe retrieval procedure, about 3130 shoe items with several daily shoe photos and online store images of different views 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 image pairs for the network training. About 25000 online store shoe images are used as reference gallery and 4400 daily shoe photos are

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

Table 1: Top-20 retrieval accuracy on the MVShoe dataset.

Method	Top@1(%)	Top@10(%)	Top@20(%)
DSIFT + Fisher Vector [6]	6.16	13.93	16.32
HessianSIFT + VLAD [1]	7.77	16.52	18.77
Pre-trained VGG 16 Layers CNN (conv5) [16]	14.97	28.83	32.27
Siamese Network [2]	23.47	50.60	58.53
Triplet Embedding Network [14]	33.65	59.24	65.69
Proposed MTV-CNN	**37.42**	**66.87**	**73.26**

used as the queries where each of them has a counterpart in the reference set.

6.2 Comparison Methods and Evaluation Protocol

To analyze the performance of our proposed MTV-CNN, we compare our method with a variety of approaches. These approaches can be divided into two groups. One group extracts the hand-crafted feature for retrieval, including (1) Dense SIFT feature followed by fisher vector encoding (DSIFT + Fisher Vector) [6] with the codebook size set as 64; 2) Hessian feature with VLAD encoding [1]. The other group utilizes the deep learning methods for retrieval feature learning. This group includes the following competitors: (1) Pre-trained VGG 16 Layers CNN: Deep pool5 feature activated from the VGG 16 layers model [16]; (2) Siamese Netowrk [2]: Two branches of input with the contrastive loss; (3) Triplet Embedding Network [14]: Three branches of input with the hinge loss. The performance is evaluated by the Top-K accuracy, which means that if the exactly matched shoes is found in the top-K list, then it is a successful result.

6.3 Results

6.3.1 Comparison with state-of-the-art methods

Table. 1 demonstrates the Top-K accuracy of our method and the compared methods when $K = 1$, 10 and 20. From the experimental results, we find that the deep learning based feature representation has a significant improvement over conventional state-of-the-art retrieval pipeline. We also compare the MTV-CNN with several other works that utilize the deep metric learning, Siamese network [19] and triplet network [14]. The CNN model with the triplet loss is better than that with the siamese network using the contrastive loss by about 9% in Top-20 accuracy, which indicates the triplet network is more suitable for our problem. Our proposed MTV-CNN achieves the best performances among all the competitors, outperforming the triplet embedding methods by about 6.5% in Top-20 accuracy, which verifies the effectiveness of our framework learning the semantic view-invariant deep shoe feature representation in a multi-task manner.

6.3.2 Evaluation of Components

To evaluate the specific contribution of each auxiliary task, the individual tasks are stacked to the framework one by one and the corresponding retrieval performance is reported as shown in Table 2. The simplest version is using the triplet network (TN) directly. The other two

Table 2: Contributions of the auxiliary tasks to the proposed MTV-CNN.

Auxiliary Tasks	Top-1(%)	Top-10(%)	Top-20(%)
TN	33.65	59.24	65.69
TN + AP	33.45	61.90	67.96
TN + VI	35.13	63.60	69.71
TN + AP + VI	**35.33**	**64.60**	**71.05**

Table 3: Contributions of each stage to the proposed MTV-CNN.

Stage	Top-1(%)	Top-10(%)	Top-20(%)
Stage 1	30.74	58.60	65.78
Stage 1 + Stage 2	37.42	66.87	73.26

auxiliary task are simplified as AP for attribute prediction and VI for view invariance, respectively. We can find that the retrieval performance improves gradually as more auxiliary tasks are added to the framework, which verifies the importance of each auxiliary task. The model using TN + AP + VI obtains about 3% improvement in terms of Top-20 accuracy compared to that without VI module, which indicates the effectiveness of the view invariant loss. Table 3 demonstrates the performance improvements of the hard example weighting and sampling approach. It can be seen that after two stages of hard negative sampling and selection, the proposed MTV-CNN has about 2.6% improvements over TN+ AP + VI (71.05%) with randomly triplet selection.

7 Conclusion

In this paper, we address the problem of street-to-shop shoe retrieval via a multi-task view-invariant triplet network, which embeds the feature of images for the same shoes close to each other. We show that the proposed MTV-CNN is simultaneously solving the task of the attribute prediction, and preserving the view-invariance, and the main task of triplet loss is helpful for learning a semantically-aware deep shoe feature representation. Our training strategy is a stage-by-stage process, progressively selecting the hard examples and assigning the weights to the triplets based on the semantic attributes to penalize more the hard examples. Negative sample with increasingly challenging examples are mined. We also establish a novel multi-view cross-scenario large-scale shoe dataset, MVShoe. The experiments performed on our newly built dataset show the effectiveness of our proposed method.

Acknowledgement

This research was carried out at the Rapid-Rich Object Search(ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme. Boxin Shi is supported by a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Relja Arandjelovic and Andrew Zisserman. All about vlad. In *CVPR*, 2013.

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

[3] Léon Bottou. Stochastic gradient tricks. In *Neural Networks, Tricks of the Trade, Reloaded*, Lecture Notes in Computer Science (LNCS 7700), pages 430–445. 2012.

[4] Ronan Collobert, Koray Kavukcuoglu, and Clément Farabet. Torch7: A matlab-like environment for machine learning. In *NIPS Workshop*, 2011.

[5] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[6] Herve Jegou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Perez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9):1704–1716, 2012.

[7] BG Kumar, Gustavo Carneiro, Ian Reid, et al. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *CVPR*, 2016.

[8] Hongye Liu, Yonghong Tian, Yaowei Yang, Lu Pang, and Tiejun Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, 2016.

[9] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, 2012.

[10] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. Fashion parsing with weak color-category labels. *IEEE Transactions on Multimedia*, 16(1):253–265, 2014.

[11] Si Liu, Xiaodan Liang, Luoqi Liu, Ke Lu, Liang Lin, Xiaochun Cao, and Shuicheng Yan. Fashion parsing with video context. *IEEE Transactions on Multimedia*, 17(8): 1347–1358, 2015.

[12] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

[13] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

[14] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[15] E. Simo-Serra and H. Ishikawa. Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In *CVPR*, 2016.

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

[17] Fang Wang, Le Kang, and Yi Li. Sketch-based 3d shape retrieval using convolutional neural networks. In *CVPR*, 2015.

[18] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014.

[19] Xi Wang, Zhenfeng Sun, Wenqiang Zhang, Yu Zhou, and Yu-Gang Jiang. Matching user photos to online products with robust deep features. In *ACM ICMR*, 2016.

[20] Yan Wang, Sheng Li, and Alex C Kot. Deepbag: Recognizing handbag models. *IEEE Transactions on Multimedia*, 17(11):2072–2083, 2015.

[21] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *CVPR*, 2015.

[22] Cha Zhang and Zhengyou Zhang. Improving multiview face detection with multi-task deep convolutional neural networks. In *WACV*, 2014.

[23] Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang. Embedding label structures for fine-grained feature representation. In *CVPR*, 2016.

[24] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.

[25] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014.