Understanding Fashion Trends from Street Photos via Neighbor-Constrained Embedding Learning

Xiaoling Gu
College of Computer Science
Zhejiang University
belizabeth@zju.edu.cn

Yongkang Wong
Interactive & Digital Media Institute
National University of Singapore
yongkang.wong@nus.edu.sg

Pai Peng
YoutuLab
Tencent Technology (Shanghai) Co., Ltd
popeyepeng@tencent.com

Lidan Shou
College of Computer Science
Zhejiang University
should@zju.edu.cn

Gang Chen
College of Computer Science
Zhejiang University
cg@zju.edu.cn

Mohan S. Kankanhalli
School of Computing
National University of Singapore
mohan@comp.nus.edu.sg

ABSTRACT

Driven by the increasing popular image-dominated social networks, such as Instagram, Pinterest and Chictopica, sharing of daily-life street photos now plays an influential role in fashion adoption between fashion trend-setters and followers. In this work, we propose a deep learning based fine-grained embedding learning approach for street fashion analysis by leveraging user-generated street fashion data. Specifically, we present QuadNet, an effective CNN based image embedding network driven by both multi-task classification loss and neighbor-constrained similarity loss. The latter loss function is computed with a novel quadruplet loss function, which considers both hard and soft positive neighbors as well as a negative neighbor for each anchor image. The embedded feature learned from co-optimization is effective for both fine-grained classification task and image retrieval task. Quantitative evaluation on a newly collected large-scale multi-task street photo dataset shows that our QuadNet outperforms the state-of-the-art triplet network by a significant margin. In order to further evaluate the effectiveness of the learned embedding, we analyze and trace the fashion trends of New York City from 2011 to 2016. In our analysis, we are able to identify some short-term and long-term fashion styles.

CCS CONCEPTS

• **Information systems** → **Social tagging**; *Evaluation of retrieval results*; • **Applied computing** → *Online shopping*;

KEYWORDS

Quadruplet Loss; Street Fashion Analysis; Fashion Trends Analysis

1 INTRODUCTION

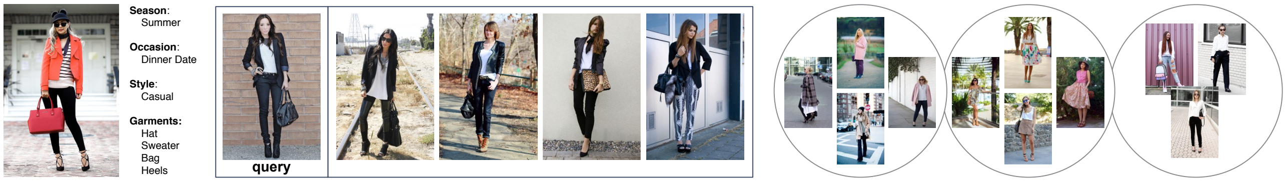
Mainstream fashion often appropriates street fashion trends as influences. As an increasing number of users are obsessed with sharing their fashionable street photos on social media platforms, street style has become a driving force of fashion trends. By distributing dressing ideas through the shared street photos, trend setters (e.g. celebrities and fashion bloggers) are believed to play an important role in influencing the personal fashion styles of their followers. Popular fashion brands even have designed their products based on influential bloggers[1]. Meanwhile, the availability of large-scale street data motivates researchers to analyze street fashion due to its societal and economic impact. The analytics and applications cover a wide range of topics, such as fashion photo segmentation [19, 35–37], fashion style classification [13, 17, 25], street-to-shop clothing retrieval [14, 16, 20], and data analysis [2, 24, 29, 34].

Street photos collected from image-dominated social network often consists of weakly-labeled metadata describing fine-grained image attributes. While most of the previous work focuses on a particular recognition task, such as clothing parsing [36] and fashion style classification [13], in fact, street photos contain multiple high-level attributes that are crucial for fashion analysis. For example, the particular combination of garments in the given image is strongly associated with different styles and a particular season. That is to say, fashion image recognition is genuinely a multi-task classification problem where a given image has distinct fine-grained attributes, which can be categorized into single label attributes (i.e. season and style) or multi-label attributes (i.e. garment categories).

On the other hand, street fashion analysis is highly related to image recognition and image retrieval tasks. More specifically, in order to provide insights on the noisy street photos, there exist three fundamental and specific tasks, namely *automated street photo annotation*, *street photo retrieval*, and *fashion trends analysis*. An illustration of these tasks is shown in Figure 1. Although it is possible to design a specialized model for each task, a more efficient approach is to learn a general and robust feature embedding model from large-scale user-generated street photos. Once the embedding has been produced, various analytics and applications can be conducted by performing clustering or other algorithms in the embedding space. A straightforward approach to learning such embedding is to use a

[1] http://fashionista.com/2016/03/style-bloggers-2016

MM '17, October 23–27, 2017, Mountain View, CA, USA
© 2017 Association for Computing Machinery.
ACM ISBN 978-1-4503-4906-2/17/10…$15.00
https://doi.org/10.1145/3123266.3123441



(a) Street photo auto-annotation (b) Street photo retrieval (c) Fashion trends analysis: mining popular dressing patterns

Figure 1: Three fundamental tasks for street fashion analysis.

fine-tuned CNN that minimizes the classification loss [12], i.e. extracting features from the last fully connected layer as embedded feature. However, this simple solution suffers from two main drawbacks. First, the learned embedding is usually high-dimensional. For example, VGG-16 network [26] has a dimensionality of 4096. Thus, it incurs high overheads for retrieval task. Second, although the learned embedding has sufficient discriminative capacity that can be used for the classification task, it is not good enough for similarity metric. This is because the objective of a classification task is to learn a clear decision boundary between different classes. Therefore, the distance between two embeddings may not precisely reflect the semantic similarity of the corresponding photos.

To address the above issues, we propose a novel embedding learning network called QuadNet for street fashion analysis, which produces lower dimensional embeddings by incorporating a new neighbor-based similarity constraint. Specifically, QuadNet consists of four identical CNNs to perform embedding learning, where the shared CNN is jointly optimized with a *multi-task classification loss* and a neighbor-constrained *quadruplet loss*. The multi-task classification loss is designed for learning the discriminative feature representation that can perform label prediction for multiple tasks. The proposed quadruplet loss is designed for similarity metric learning, where the distance constraints between the embeddings of an anchor image and its different types of neighbors are encoded. Quantitative evaluation on the multi-task street photo dataset shows that our QuadNet outperforms the state-of-the-art triplet network on both classification and retrieval tasks. In addition, we demonstrate the efficacy of the learned embedding by analyzing the fashion trends of New York City from 2011 to 2016.

Our contributions are summarized as follows:

- We propose a novel neighbor-constrained embedding learning network called QuadNet for street fashion analysis. A new quadruplet loss is designed for similarity metric learning by considering both hard and soft positive neighbors as well as a negative neighbor for each anchor image.
- We collected a new multi-task street photos dataset, namely *Street Fashion Style* (SFS) dataset, with a total of 293,105 posts from Chictopia for fashion analysis. Each image consists of weakly-labeled multi-task ground-truth, including season, occasion, fashion style, garment categories, geographical and year information.
- Quantitative evaluation on SFS dataset shows the proposed QuadNet outperforms the conventional triplet network by a significant margin. A fashion trends analysis of New York city further demonstrates the efficacy of the learned embedding.

2 RELATED WORK

2.1 Street Data Analysis

In recent years, influenced by large-scale street data and the demand in the fashion market, researchers have made active research progress in street data analysis. The early work focuses on clothing parsing which simultaneously performs garment item segmentation and labeling on fashion photos [19, 35–37]. This task is extremely challenging due to the wide variety of garment items, possible variations in combination, layering, and occlusion. Another stream of research focuses on the cross-domain clothing retrieval problem, which aims to find the exact or similar clothing from online stores from a given a daily street photo [16, 20]. Very recently, Jiang *et al.* [14] proposed a deep cross-triplet embedding algorithm to jointly solve the bi-directional shop-to-street and street-to-shop clothing retrieval problems. One unique classification problem on fashion analysis is the fashion style classification, where fashion photos are classified into five categories [13, 17, 25]. Fashion analysis with street data has received increasing attentions in recent year [2, 24, 29, 34]. Vittayakorn *et al.* [29] analyzed how fashion trends transfer from runway collections to the dressing patterns in real life. Yamaguchi *et al.* [34] presented a vision-based approach to quantitatively evaluate the influence of social factors and content factors on popularity in a large real-world fashion social network. Chen *et al.* [2] discovered fashion trends in New York City by utilizing semantic clothing attributes (e.g. color, cut, head accessories, *etc.*). Simo-Serra *et al.* [24] utilized a conditional random field model to predict how fashionable a person looks on a photograph and suggest improvements to improve her/his appeal.

Different from most recent work, we propose to learn a neighbor-constrained embedding for street fashion analysis, which enables many practical tasks such as street photo annotation, street photo retrieval and fashion trends analysis. Our work is similar to [25]. The main difference is that we propose a unified framework for both classification and feature embedding learning by jointly optimizing the classification loss and a new neighbor-constrained quadruplet loss, whereas [25] used the classification loss to aid the learning of the embedded feature and an independent classifier is trained with the learned embedding.

2.2 Multi-label Annotation and Classification

Multi-label image annotation is a fundamental and challenging research problem. Given an unseen image, the goal is to predict multiple semantic labels. Makadia *et al.* [21] proposed a simple nearest neighbor-based tag transfer approach. Guillaumin *et al.* [8] proposed a discriminative framework that combines a weighted nearest-neighbor model with metric learning capabilities for this task. Weston *et al.* [32] introduced a scalable model by learning a joint representation of images and annotations that optimize precision at the top of the ranked list of annotations for a given image. Chen *et al.* [3] presented an image tagging method with two simple linear mappings that are co-regularized in a joint convex loss function. These above works focus on designing hand-crafted visual features to improve the accuracy of multi-label annotation. In contrast to aforementioned approaches, CNNs have been studied to solve multi-label annotation problem [7, 15, 33]. Johnson *et al.* [15] proposed a parametric visual model based on CNNs, where image metadata is exploited to generate neighborhoods of images. More recently, image annotation has been formulated as a classification problem [7, 33]. Gong *et al.* [7] trained CNNs with the weighted approximated-ranking loss. Wu *et al.* [33] addressed the annotation problem with a semi-supervised approach by jointly optimizing a weakly weighted pairwise ranking loss and a triplet similarity loss. Similar with [7], we formulate street photo annotation as a multi-task classification problem.

2.3 Deep Metric Learning

Deep metric learning has achieved promising results in various tasks using CNNs based on contrastive loss [1, 9] or triplet loss [4, 10, 23, 30, 38]. The objective is to learn a lower dimensional feature embedding that captures the semantic similarity among images with the distances evaluated in the embedding space. Cui *et al.* [4] proposed a framework for fine-grained visual categorization and dataset bootstrapping using deep metric learning with humans in the loop. Hoffer *et al.* [10] described a triplet network architecture for deep metric learning. Wang *et al.* [30] proposed a deep ranking model that employs a triplet-based ranking loss and an efficient online triplet sampling method to learn fine-grained image similarity. Zhang *et al.* [38] jointly optimized both classification loss and triplet loss to learn fine-grained feature representation. Bell and Bala [1] presented a method to learn the visual similarity metric for interior design with Siamese networks. Deep neural networks with Siamese or triplet architecture have also been applied to the problem of face verification, alignment, and recognition [11, 23]. Extending from pairwise or triplet-wise approaches, Law [18] introduced an image similarity learning framework with the quadruplet-wise constraints, while Ustinova [27] presented a Histogram loss for learning deep embeddings. Likewise, our proposed quadruple also extends from triplet loss. It should be noted that our quadruple loss has different definition compared with the quadruplet-wise constraints proposed in [18]. Different from the triplet loss designed for minimizing the distance between an anchor image and a positive neighbor, and maximizing the distance between the anchor image and a negative neighbor, the proposed quadruple loss enhances the similarity constraint by differentiating hard positive neighbor and soft positive neighbor. The experimental results quantitatively show that our quadruple loss outperforms the conventional triplet loss.

3 PROPOSED ARCHITECTURE

We propose a unified CNN architecture called QuadNet for robust neighbor-constrained embedding learning by exploiting the weakly labeled street photos. For each street photo, its visual content, tags and neighbors are jointly considered. The goal is to use the learned embedding for both classification and similarity estimation. The architecture of QuadNet is illustrated in Figure 2. It computes the embedding of an image x, $f(x) \in \mathbb{R}^d$, where d is the dimensionality of the learned embedding. In the training phase, QuadNet requires four input images, including an anchor image $\boldsymbol{x}_i^a$, a hard positive neighbor $\boldsymbol{x}_i^{p^+}$, a soft positive neighbor $\boldsymbol{x}_i^{p^-}$, and a negative neighbor $\boldsymbol{x}_i^n$. These four input images are independently fed into four CNNs (with shared parameters). Two following loss layers are designed for the neighbor-constrained embedding learning. First, a multi-task classification loss layer is applied on the negative neighbor image network to predict the class labels, where three fully connected layers and corresponding softmax layers are added. Second, a novel quadruplet loss layer is applied to L_2-normalization embedding of all four networks. The quadruplet loss layer is designed for similarity metric learning, where the distance constraints between the embeddings of an anchor image and its different types of neighbors are encoded.

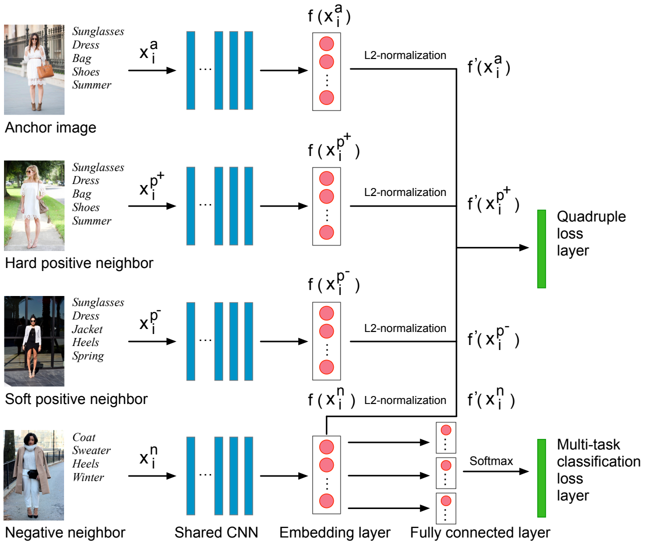


Figure 2: QuadNet: our proposed network architecture for neighbor-constrained embedding learning.

We jointly optimize the classification loss and the quadruplet loss in the shared CNN. The overall loss function can be expressed as:

$$L_{\texttt{Overall}} = L_{\texttt{Class}} + \lambda_{\texttt{Quad}} L_{\texttt{Quad}} \tag{1}$$

where $L_{\texttt{Class}}$ is the multi-task classification loss, $L_{\texttt{Quad}}$ is the proposed quadruplet loss, and $\lambda_{\texttt{Quad}}$ is the weight parameter to control the trade-off between the two losses terms. During the training phase, the loss is backpropagated to each layer of the CNN and their corresponding parameters are updated through Stochastic Gradient Descent (SGD). Given the learned model and an input image, the corresponding embedding can be extracted from the network for street photo classification, retrieval, or trend analysis.

3.1 Multi-task Classification Loss

Annotation of street photos is naturally a multi-task labeling problem, where some tasks belong to the binary classification problem (i.e. one label per image) and the others belong to the multi-label

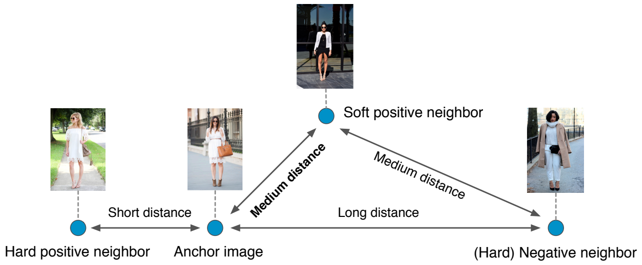


Figure 3: Illustration of distance constraints in quadruplet loss.

classification problem (i.e. one or more labels per image). For instance, the season annotation task can either be *spring*, *summer*, *fall* or *winter*, which is mutually exclusive. On the other hand, the garment annotation task is very common to have multiple labels, such as {*dress*, *hat*, *heels*, *bag*}.

Considering that these classification tasks may share common features, we utilize the shared CNN to jointly learn a unified discriminative feature embedding that can perform label prediction for multiple tasks. In this work, we define three annotation tasks, namely *season* annotation, *style* annotation, and *garment* annotation. Due to the difference in the classification problem, we adopt two loss functions. Specifically, the cross entropy loss [22] is defined for season and style classification, while a similar multi-label softmax loss [7] is defined for garment classification. Given N training images x_i and the corresponding season label Y^s, style label Y^t, and garment label Y^g, the loss functions of each task are defined as follows:

$$L_s = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_s} -Y^s_{i,j}\log\left(f^s_j(x_i)\right) \tag{2}$$

$$L_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_t} -Y^t_{i,j}\log\left(f^t_j(x_i)\right) \tag{3}$$

$$L_g = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K_g} -\bar{Y}^g_{i,j}\log\left(f^g_j(x_i)\right) \tag{4}$$

where K denotes the number of classes, i indexes images and j indexes classes. $Y^s_{i,j} \in \{0,1\}$, $Y^t_{i,j} \in \{0,1\}$ and $Y^g_{i,j} \in \{0,1\}$ denote the ground-truth season label, style label and garment label of image x_i and class j, respectively. Note that $\bar{Y}^g_{i,j} = Y^g_{i,j}/\sum_{j=1}^{K_g} Y^g_{i,j}$. $f_j(x_i)$ represents the softmax layer output. Finally, the overall classification loss function $L_{\mathtt{Class}}$ is defined as:

$$L_{\mathtt{Class}} = \lambda_1 L_s + \lambda_2 L_t + \lambda_3 L_g \tag{5}$$

where λ_1, λ_2 and λ_3 are three weight parameters to control the trade-off between the three losses terms. We constrain these parameters to be positive and sum-to-one.

3.2 Proposed quadruplet Loss

The second loss function is the neighbor-constrained similarity loss. The current state-of-the-art approach for similarity metric learning of street photos is to employ triplet loss in metric learning [23]. Motivated in the context of nearest-neighbor classification [31], triplet loss minimizes the distance between an anchor image and a positive neighbor, and maximizes the distance between the anchor image and a negative neighbor. Triplet loss is effective when the data on hand can be easily distinguished as (hard) positive or (hard) negative, such as face recognition [23] and object classification [4, 38].

However, the triplet loss is not the optimal solution for street photos as many samples consist of multiple class labels. Given a specific anchor image, except its hard positive and negative neighbors, there exists many images which are partially similar yet not identical to the anchor image, we refer them as *soft positive neighbor*. As illustrated in Figure 2, the input anchor image and its hard positive neighbor share more common labels and high visual similarity, whereas its soft positive neighbor share less common labels and visual similarity with anchor image. To this end, we propose a novel quadruplet loss, where the soft positive neighbor is differentiated from the hard positive neighbor. In this work, we define the similarity between two street photos by using the corresponding garment and season labels. The season label can be regarded as a soft constraint to remove the visual ambiguity of same category of garments. For example, a winter dress and a summer dress are both belong to the dress label yet its visual appearance is highly different due to the season. Formally, given an image pair x_1 and x_2, the corresponding similarity is defined using the Jaccard similarity function as:

$$S(\boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{T_{\boldsymbol{x}_1} \cap T_{\boldsymbol{x}_2}}{T_{\boldsymbol{x}_1} \cup T_{\boldsymbol{x}_2}} \tag{6}$$

where $T_{\boldsymbol{x}_i}$ are the ground-truth garment and season labels of image $\boldsymbol{x}_i$. This similarity function is used to sample different types of neighbors for an anchor image.

Given a quadruplet set, where each training sample consists of an anchor image $\boldsymbol{x}^{\mathsf{a}}_i$, a hard positive neighbor $\boldsymbol{x}^{\mathsf{p}^+}_i$, a soft positive neighbor $\boldsymbol{x}^{\mathsf{p}^-}_i$, and a negative neighbor $\boldsymbol{x}^{\mathsf{n}}_i$, our goal is to learn an embedding that assigns a short distance for $< \boldsymbol{x}^{\mathsf{a}}_i, \boldsymbol{x}^{\mathsf{p}^+}_i >$ pairs, a medium distance for both $< \boldsymbol{x}^{\mathsf{a}}_i, \boldsymbol{x}^{\mathsf{p}^-}_i >$ pairs and $< \boldsymbol{x}^{\mathsf{p}^-}_i, \boldsymbol{x}^{\mathsf{n}}_i >$ pairs, and a long distance for $< \boldsymbol{x}^{\mathsf{a}}_i, \boldsymbol{x}^{\mathsf{n}}_i >$ pairs (see Figure 3). Since we require to compute distances in the learned feature space for retrieval task, all features are normalized by L_2-normalization to eliminate the scale differences. The feature normalization can be achieved via $f'(\boldsymbol{x}) = f(\boldsymbol{x})/\sqrt{f(\boldsymbol{x})^T f(\boldsymbol{x})}$.

Formally, the intuition illustrated in Figure 3 can be expressed:

$$\begin{aligned}
\|f'(\boldsymbol{x}^{\mathsf{a}}_i) - f'(\boldsymbol{x}^{\mathsf{p}^+}_i)\|_2^2 + m_1 &< \|f'(\boldsymbol{x}^{\mathsf{a}}_i) - f'(\boldsymbol{x}^{\mathsf{n}}_i)\|_2^2 \\
\|f'(\boldsymbol{x}^{\mathsf{a}}_i) - f'(\boldsymbol{x}^{\mathsf{p}^+}_i)\|_2^2 + m_2 &< \|f'(\boldsymbol{x}^{\mathsf{a}}_i) - f'(\boldsymbol{x}^{\mathsf{p}^-}_i)\|_2^2 \\
\|f'(\boldsymbol{x}^{\mathsf{a}}_i) - f'(\boldsymbol{x}^{\mathsf{p}^-}_i)\|_2^2 + m_3 &< \|f'(\boldsymbol{x}^{\mathsf{a}}_i) - f'(\boldsymbol{x}^{\mathsf{n}}_i)\|_2^2 \\
&s.t. : m_1 > m_2 \\
&s.t. : m_1 > m_3 \\
&\forall\ (\boldsymbol{x}^{a}_i, \boldsymbol{x}^{\mathsf{p}^+}_i, \boldsymbol{x}^{\mathsf{p}^-}_i, \boldsymbol{x}^{\mathsf{n}}_i) \in \Omega
\end{aligned} \tag{7}$$

where m_1, m_2, m_3 are the margin parameters that regularize the gap between the squared Euclidean distance of the two corresponding image pairs. Ω is a specific quadruplet set and has cardinality N.

Similar with the triplet loss, the quadruplet loss over N samples can be defined as follows:

$$\begin{aligned} L_{\mathsf{Quad}} = \frac{1}{N}\sum_{i=1}^{N}\Big[& \max\Big(0, m_1 + \|f'(\boldsymbol{x}_i^a) - f'(\boldsymbol{x}_i^{\mathsf{p}^+})\|_2^2 - \|f'(\boldsymbol{x}_i^a) - f'(\boldsymbol{x}_i^{\mathsf{n}})\|_2^2\Big) \\ & + \max\Big(0, m_2 + \|f'(\boldsymbol{x}_i^a) - f'(\boldsymbol{x}_i^{\mathsf{p}^+})\|_2^2 - \|f'(\boldsymbol{x}_i^a) - f'(\boldsymbol{x}_i^{\mathsf{p}^-})\|_2^2\Big) \\ & + \max\Big(0, m_3 + \|f'(\boldsymbol{x}_i^a) - f'(\boldsymbol{x}_i^{\mathsf{p}^-})\|_2^2 - \|f'(\boldsymbol{x}_i^a) - f'(\boldsymbol{x}_i^{\mathsf{n}})\|_2^2\Big)\Big] \\ & \qquad s.t.: m_1 > m_2 \\ & \qquad s.t.: m_1 > m_3 \\ & \qquad \forall(\boldsymbol{x}_i^a, \boldsymbol{x}_i^{\mathsf{p}^+}, \boldsymbol{x}_i^{\mathsf{p}^-}, \boldsymbol{x}_i^{\mathsf{n}}) \in \Omega \end{aligned} \tag{8}$$

4 EXPERIMENT

4.1 Dataset

Currently, there exists a few public street fashion datasets, such as Fashionista [36] and Fashion 144k [24]. However, these datasets do not provide the required labels for our proposed street fashion analysis. In order to facilitate this study, we collected a new street photos dataset from Chictopia[2], namely *Street Fashion Style* (SFS) dataset[3], where a total of 293,105 user posts are crawled. In each post, a user usually publishes the photograph of her/his worn outfit along with associated tags. Generally, these tags include current season, the suitable occasion, fashion style, the detailed garment information (e.g. category, color and brand), the geographical and year information. The geographical and year information is used in the fashion trends analysis (Section 4.3).

In this work, we select the labels of season, style and category of garments as the ground-truth of the classification task. The color labels are not considered due to the data sparseness. The season classification task consists of four classes, namely *spring*, *summer*, *fall*, and *winter*. For style classification, classes with small amount of ground-truth labels are removed and 15 dominant classes are retrained[4]. For garment classification, we remove small accessories classes (e.g. bracelet, necklace and ring) as it covers little visual region for image recognition, and retain a total of 24 dominant classes[5]. To alleviate the missing label problem, we apply a simple yet effective filtering mechanism called *minimal dressing pattern*, which indicates the minimum amount of labels required to form a dressing style. We define two basic dressing patterns: (1) dress with shoes; and (2) upper garment with bottom garment and shoes. Based on these constraints, we assembled a total of 85,720 images and associated labels for our classification and retrieval analysis.

To generate the quadruplet set, we consider each of the assembled images as an anchor image, and the corresponding hard positive neighbor, soft positive neighbor, and negative neighbor are uniformly sampled with Eq. 6 using the similarity threshold 0.9, 0.3, and 0.001, respectively. The similarity threshold for sampling the soft positive neighbors is denoted as σ_s, where the influence of its selection will be discussed in Section 4.2.4. Examples of the sampled quadruplet set are shown in Figure 4. In our experiments, we split the data into the same training set, validation set, and test set with a ration of 7:1:2 for both classification and retrieval experiments.

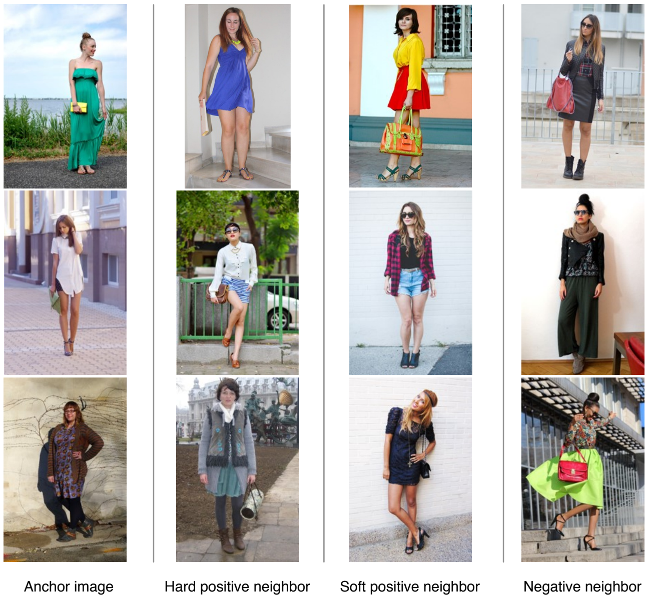


Figure 4: Examples of the quadruplet set.

4.2 Evaluation

QuadNet extends VGG-16 network [26] by adding an additional fully-connected layer (128D) on its 6th fully-connected layer. The CNN model is fine-tuned with the initial weights that pre-trained on the ImageNet dataset. We apply global normalization with zero mean and unit variance in the preprocessing phase. No data augmentation or whitening is applied in our implementation. The optimal parameters are selected with the following steps: First, we choose the optimal values for weight parameters λ_1, λ_2 and λ_3. During the fine-tuning stage, we set $\lambda_{\mathsf{Quad}} = 0$ to ignore the influence of quadruplet loss. Based on the intensive preliminary experiment, the parameters are fixed as $\lambda_1 = 0.3$, $\lambda_2 = 0.3$ and $\lambda_3 = 0.4$. Second, we evaluate the weight parameter λ_{Quad} from [0.1, 1.0] with step size of 0.1, where $\lambda_{\mathsf{Quad}} = 0.2$ achieves the best performance. We observed that if $\lambda_{\mathtt{Qual}}$ is too large, the classification accuracy is greatly reduced, whereas a small value leads to the slow convergence of the similarity loss. On third step, we search the optimal values for m_1, m_2 and m_3. We test m_1 from 0.1 to 1, while m_2, m_3 are set around $\lfloor m_1/2 \rfloor$. Through our experiments, we empirically set $m_1 = 0.5$, $m_2 = 0.2$, $m_2 = 0.3$ for best performance. And as 128 is the common dimension for feature embedding learning, here, we empirically set the parameters feature dimension $d = 128$.

In this work, we mainly compare QuadNet with two baseline methods. The first baseline is the deep model sharing the same extended CNN architecture but using conventional triplet loss. The second baseline is the proposed method in [25]. In the following section, we evaluate our QuadNet in three tasks, namely classification with deep models, classification with SVM models and image retrieval.

4.2.1 Fashion Photo Classification with Deep Models.

Due to different types of classification, different evaluation metrics should be employed. For season and style classification, we adopt the commonly used accuracy metric in multi-class classification. Following

[2] http://www.chictopia.com/

[3] available via https://doi.org/10.5281/zenodo.833051

[4] *rocker, casual, comfortable, basic, eclectic, trendy, classic, chic, urban, romantic, elegant, bohemian, sexy, preppy, denim*

[5] *bag, blouse, blazer, boots, cardigan, coat, dress, hat, heels, jacket, jeans, leggings, pants, sandals, shirt, shoes, shorts, skirt, sunglasses, sweater, tights, top, t-shirt, vest*

Table 1: The classification results of trained deep models.

Method	Dim	Season	Style	Garment			
		accuracy	accuracy	per-class precision	per-class recall	overall precision	overall recall
VGG_STL	4096	0.4113	0.2015	0.4277	0.4693	0.4966	**0.5970**
VGG_MTL	4096	0.4546	0.2602	0.4286	0.4459	0.4853	0.5719
VGG_128_MTL	128	0.4524	0.2468	0.3962	0.3434	0.4672	0.5240
Simo-Serra *et al.* [25]	128	0.4355	0.2119	0.2858	0.2706	0.3827	0.4317
TripletNet	128	0.5400	0.3518	0.4796	0.4347	0.5406	0.5213
QuadNet	128	**0.6450**	**0.4338**	**0.6279**	**0.5276**	**0.6414**	0.5954

the previous work [7], the performance of garment classification is quantified with the following metrics:

$$\begin{aligned} P_c &= \frac{1}{c}\sum_{i=1}^{c}\frac{N_i^c}{N_i^p}, & R_c &= \frac{1}{c}\sum_{i=1}^{c}\frac{N_i^c}{N_i^g} \\ P_o &= \frac{\sum_{i=1}^{c} N_i^c}{\sum_{i=1}^{c} N_i^p}, & R_o &= \frac{\sum_{i=1}^{c} N_i^c}{\sum_{i=1}^{c} N_i^g} \end{aligned} \tag{9}$$

where P_c is per-class precision, R_c is per-class recall, P_o is overall precision, and R_o is overall recall.

Below are the details of the comparison methods:

(1) **VGG_STL**: Here, we adopt the first six layers of VGG-16 models [26] as the first baseline. The VGG-16 model is fine-tuned with single-task classification loss.
(2) **VGG_MTL**: Sharing the same structure with VGG_STL but fine-tuned with multi-task classification loss
(3) **VGG_128_MTL**: To validate the benefits of the co-optimization of classification loss and similarity loss, we train the same extended CNN model using only the multi-task classification loss.
(4) **Simo-Serra** *et al.* **[25]**: We use the network and the ranking loss on triplets proposed in [25] to learn the embedding for comparison.
(5) **TripletNet**: We jointly optimize conventional triplet loss and multi-task classification loss with triplet network, sharing the same CNN architecture with QuadNet.

All models are individually tuned to achieve the best performance for fair comparisons. The classification results are reported in Table 1. All VGG-16 based models (i.e. VGG_STL, VGG_MTL, VGG_128_MTL, TripletNet, QuadNet) outperform Simo-Serra *et al.*, which demonstrates that the network used in [25] is inferior to VGG-16 network in terms of classification. By comparing the results of VGG_STL and VGG_MTL, we observe that multi-task learning improves the season and style classification. However, the garment classification performance is reduced. We also observe that VGG_MTL outperforms VGG_128_MTL across all tasks. This may be caused by the fact that the lower dimensional feature embedding (in our case is 128) reduces the discriminative capacity. QuadNet and TripletNet surpass VGG_128_MTL, which proves that joint learning of classification loss and similarity loss does improve the classification results substantially while the lower feature embedding does not sacrifice the classification results. Further, we found that, for season and style classification tasks, and the pre-class precision and overall precision of garment, QuadNet and TripletNet outperform VGG_STL and VGG_MTL, which is a strong evidence of the advantages of joint learning. Finally, by comparing the classification results of QuadNet and TripletNet, we found that our QuadNet outperforms the state-of-the-art triplet network by a significant margin.

4.2.2 Fashion Photo Classification with SVM Models. To evaluate the quality of embedded features extracted from each model, we use the embeddings from the above deep models to train a set of linear SVMs with L2 regularization and L2 loss [6]. It is worth noting that the training set, test set and validation set are same as that used in previous classification tasks. The training data is balanced when training the garment classifier. The results are reported in Table 2. To test the impact of L_2-normalization towards classification, we introduce TripletNet_l2 and QuadNet_l2, which add a L_2-normalization layer on TripletNet and QuadNet, respectively. We first compare the classification results between Table 1 and Table 2. For each embedding network, the season accuracy and style accuracy in Table 2 are both very close to its corresponding terms in Table 1. It demonstrates the effectiveness of embedded features trained with deep models. For the garment classification, the *per-class precision* and *overall precision* drops from Table 1 to Table 2, while *per-class recall* and *overall recall* increases for each embedding model. This may be caused by the different training processes. For these deep models, we train a single softmax classifier, where the SVM models consist of 24 independent binary classifiers for each garment label. The results indicate that deep models are better at dealing with multi-label classification compared with SVMs. Then, by comparing the classification results in Table 2, our QuadNet and QuadNet_l2 achieve the best performance in each evaluation metric. By comparing the classification results between two pairs, (TripletNet, TripletNet_l2) and (QuadNet, QuadNet_l2), we found that L_2-normalization has little impact on classification.

4.2.3 Fashion Photo Retrieval. To evaluate how well the learned embedding can be used as a distance metric, we use an *image-to-image* retrieval evaluation, where the Euclidean distance between two embeddings is used to calculate the semantic similarity of the corresponding photos. Specifically, 5,000 images are randomly selected from the test set as query set. For each query image, its K-nearest neighbors are retrieved from the remaining test set. To decide whether two images are relevant or not, one way is to use crowdsourcing strategy. To avoid labor-intensive manual labeling, we use the existing image labels. Based on the previous similarity function Eq. 6 that samples these neighbors, we set the threshold to

Table 2: The classification results of trained SVM models.

Method	Dim	Season accuracy	Style accuracy	Garment per-class precision	per-class recall	overall precision	overal recall
VGG_STL	4096	0.4169	0.2121	0.3478	0.6075	0.3797	0.6061
VGG_MTL	4096	0.4434	0.2315	0.3230	0.5969	0.3449	0.5921
Simo-Serra *et al.* [25]	128	0.4594	0.2048	0.2772	0.3524	0.3050	0.3395
TripletNet	128	0.5978	0.3411	0.3730	0.5981	0.4138	0.5858
TripletNet_l2	128	0.5951	0.3348	0.3756	0.5964	0.4176	0.5838
QuadNet	128	**0.6537**	**0.4280**	**0.4401**	0.6667	**0.4992**	0.6650
QuadNet_l2	128	0.6522	0.4161	0.4365	**0.6685**	0.4940	**0.6664**

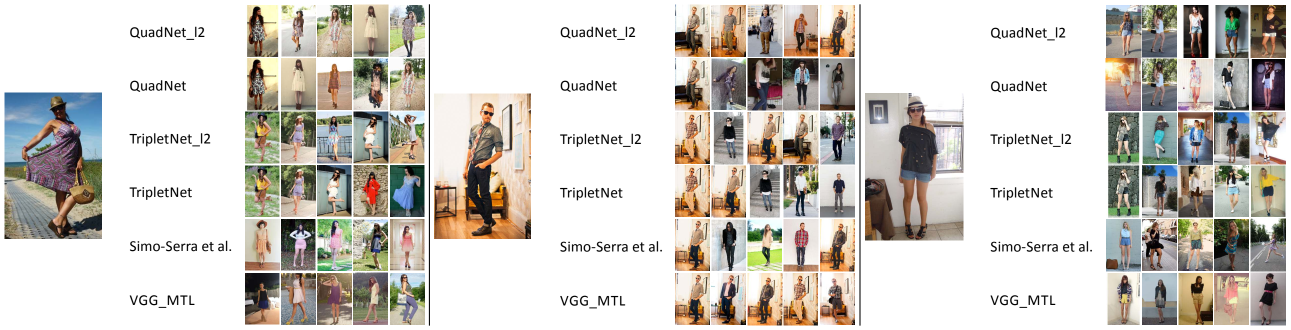


Figure 5: Some examples of top-5 retrieval results.

judge the relevance of two images as 0.3, which is the same value for sampling soft positive neighbors. The intuition is that the relevant results are at least the positive neighbors of a query image. The performance is shown in Figure 6, where Precision@K is utilized as the evaluation metric and six embeddings are compared. As shown, VGG_MTL achieves the worst performance, which indicates that fine-tuned CNN embeddings are not suitable for measuring the semantic similarity between street photos. Overall, QuadNet_l2 achieves the best performance, followed by QuadNet, TripletNet_l2, TripletNet, and Simo-Serra *et al.* [25], which demonstrates the effectiveness of our proposed quadruple loss. By comparing the feature embeddings with or without L_2-normalization, we found that L_2-normalization is essential for measuring image similarity. Examples of top-5 retrieval results are shown in Figure 5.

4.2.4 Discussions. In our QuadNet, the classification loss can be applied on one of the four embedded layers. Here, we evaluate the impact of the sensitivity of this selection. From the experimental results, we found that QuadNet gives the best performance when the classification is applied to the negative neighbor, followed by the hard positive neighbor and the soft positive neighbor. Optimizing the classification loss with the anchor image gives the lowest performance. Similar results are observed with TripletNet. Then, we test the influence of σ_s. With the optimal parameter setting, we evaluate σ_s in the range of $[0.2, 0.8]$ with step size of 0.1. From the classification results of these deep models, we found that the variations of each evaluation metric (i.e. accuracy of season, accuracy of style, precision and recall of garment) are within $4\% - 5\%$ difference.

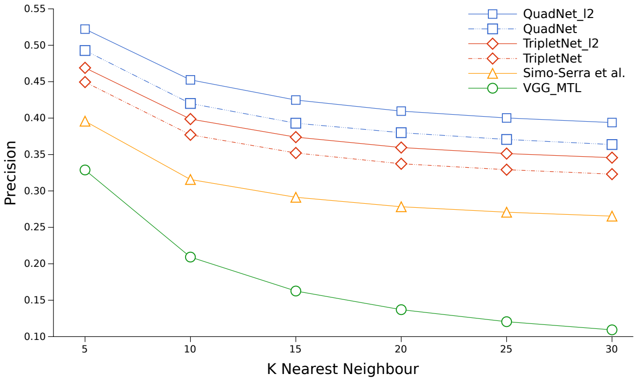


Figure 6: Evaluation results of image-to-image retrieval task.

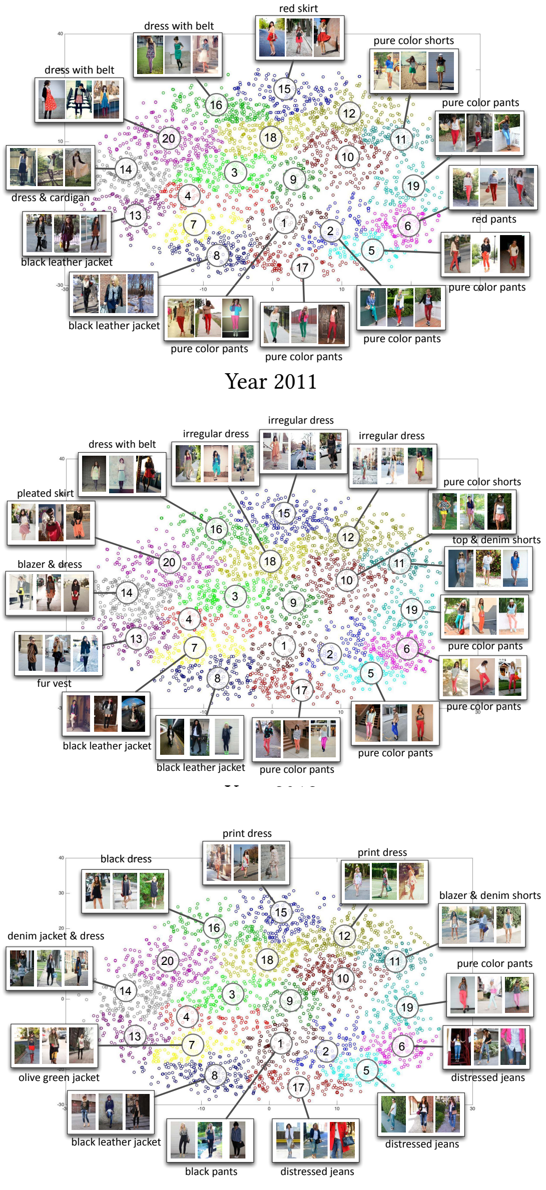
4.3 Street Fashion Trends Analysis

Mining the trends of popular fashion items and dressing patterns from social media data are crucial for fashion marketers and designers. Here, we attempt to trace the fashion trends of New York City from 2011 to 2016 with the learned embedding and additional information (i.e. geographical and year information). New York City is chosen for our analysis because of its reputation as the fashion capital of the world[6]. Note that we only consider the fashion trends for female users due to the scarcity of male users' posts. By utilizing the geographical and year information from SFS dataset, we obtain

[6] http://www.vogue.co.uk/article/new-york-is-crowned-fashion-global-capital

a total of 12,835 posts of New York City from 2011 to 2016 (*2011*: 2,595 posts; *2012*: 2,900 posts; *2013*: 2,812 posts; *2014*: 1,994 posts; *2015*: 1,033 posts; *2016*: 1,501 posts). These posts are all mapped into a unified embedding space using only their visual information with QuadNet_l2. To observe the common dressing patterns or fashion items from the generated embedding space for each year, we produce a visualization tool for fashion trends analysis called *fashion trend map*. First, we reduce the original 128 dimension embeddings to 2 dimensions with t-SNE [28]. Then, we employ agglomerative hierarchical clustering algorithm [5] to cluster these posts. Based on preliminary experiments, we set the maximum number of clusters to 20. Thus, similar dressing patterns and fashion items are mapped into the same cluster or adjacent clusters. If a specific dressing pattern or item is popular, it may occur in multiple clusters. Due to space constraints, we only display the fashion trend map from 2011-2013 (see Figure 7). More visualization results are provided in the supplementary material.

By leveraging the fashion trend map from 2011 to 2016, we made the following observations: **(1)** In these generated clusters, except some noisy clusters that have no common fashion item or dressing pattern, we can observe one or two dominant fashion items and dressing patterns from each cluster as shown with the representative photos. For example, cluster 5, 6, 17, 19 are the regions of pants and jeans, while the popular dressing patterns are jeans/pants with shirt/top. **(2)** Some dressing pattern or item will be repeated for a few years, such as pure color pants (2011-2013), dress with belt (*2011-2012*), pure color shorts (*2011-2012*), distressed jeans (*2013-2016*), print dress (*2013-2016*), print pants (*2015-2016*), and black leather jacket (*2011-2016*). It is an interesting finding that the fashionable women of New York City have a preference for the black leather jacket, matched with either dress or pants. **(3)** Some fashion styles just appear for a single year, such as red skirt (*2011*), dress & cardigan (*2011*), irregular dress (*2012*), olive green jacket (*2013*), plaid blazer or coat (*2015*), off-the-should top (*2016*), off-the-should dress (*2016*), and color fur coat (*2016*). **(4)** Some fashion styles will become popular again after a few years, such as pleated skirt (*2012*, *2016*). **(5)** Denim related fashion products, which include jeans, denim shorts and denim jacket, take an important role in street fashion, whereas the popular dressing patterns involve jeans with jacket/shirt/top, denim shorts with top or blazer, denim jacket with dress. From the evolution of above fashion trends, we can conclude that some fashion styles just appear for a short period of time while some will last for a few years.



Year 2011

Year 2013

Figure 7: The fashion trend map from 2011 to 2013.

5 CONCLUSION

In this work, we propose a novel neighbor-constrained fine-grained embedding learning approach for street photo representation and analysis. Specifically, we present an effective CNN-based network structure jointly optimized with a multi-task classification loss and a new quadruplet loss. The multi-task classification loss is designed to learn the discriminative feature representation while the quadruplet loss is designed for similarity metric learning. Thus, the learned embedding can be used for fine-grained categorization and similarity measures simultaneously. Quantitative evaluation shows the effectiveness of our proposed architecture. For future work, we consider adding other kinds of fashion data such as runway photos into the embedding learning framework, which enables a variety of interesting fashion analytics and applications.

ACKNOWLEDGMENT

This research was carried out at the NUS-ZJU SeSaMe Centre. It is supported by the National Research Foundation, Prime Minister's Office, Singapore under its International Research Centre in Singapore Funding Initiative. This research is also supported by the National Basic Research Program of China (GrantNo. 2015CB352402) and National Science Foundation of China (GrantNo. 61672455, 61472348 and 61661146001).

REFERENCES

[1] Sean Bell and Kavita Bala. 2015. Learning visual similarity for product design with convolutional neural networks. *ACM Trans. Graph.* 34, 4 (2015), 98.

[2] Kuan-Ting Chen, Kezhen Chen, Peizhong Cong, Winston H. Hsu, and Jiebo Luo. 2015. Who are the Devils Wearing Prada in New York City?. In *ACM Multimedia*. 177–180.

[3] Minmin Chen, Alice X. Zheng, and Kilian Q. Weinberger. 2013. Fast Image Tagging. In *ICML*. 1274–1282.

[4] Yin Cui, Feng Zhou, Yuanqing Lin, and Serge J. Belongie. 2016. Fine-Grained Categorization and Dataset Bootstrapping Using Deep Metric Learning with Humans in the Loop. In *CVPR*. 1153–1162.

[5] Ian Davidson and SS Ravi. 2005. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In *European Conference on Principles of Data Mining and Knowledge Discovery*. 59–70.

[6] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* 9 (2008).

[7] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. 2013. Deep Convolutional Ranking for Multilabel Image Annotation. *CoRR* abs/1312.4894 (2013).

[8] Matthieu Guillaumin, Thomas Mensink, Jakob J. Verbeek, and Cordelia Schmid. 2009. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*. 309–316.

[9] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality Reduction by Learning an Invariant Mapping. In *CVPR*. 1735–1742.

[10] Elad Hoffer and Nir Ailon. 2015. Deep metric learning using Triplet network. *Lecture Notes in Computer Science* 9370 (2015), 84–92.

[11] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. 2014. Discriminative Deep Metric Learning for Face Verification in the Wild. In *CVPR*. 1875–1882.

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional Architecture for Fast Feature Embedding. In *ACM Multimedia*. 675–678.

[13] Shuhui Jiang, Ming Shao, Chengcheng Jia, and Yun Fu. 2016. Consensus Style Centralizing Auto-Encoder for Weak Style Classification. In *AAAI*. 1223–1229.

[14] Shuhui Jiang, Yue Wu, and Yun Fu. 2016. Deep Bi-directional Cross-triplet Embedding for Cross-Domain Clothing Retrieval. In *ACM Multimedia*. 52–56.

[15] Justin Johnson, Lamberto Ballan, and Fei-Fei Li. 2015. Love Thy Neighbors: Image Annotation by Exploiting Image Metadata. In *ICCV*. 4624–4632.

[16] M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, and Tamara L. Berg. 2015. Where to Buy It: Matching Street Clothing Photos in Online Shops. In *ICCV*. 3343–3351.

[17] M. Hadi Kiapour, Kota Yamaguchi, Alexander C. Berg, and Tamara L. Berg. 2014. Hipster Wars: Discovering Elements of Fashion Styles. In *ECCV*. 472–488.

[18] Marc Teva Law, Nicolas Thome, and Matthieu Cord. 2013. Quadruplet-Wise Image Similarity Learning. In *ICCV*. 249–256.

[19] Si Liu, Jiashi Feng, Csaba Domokos, Hui Xu, Junshi Huang, Zhenzhen Hu, and Shuicheng Yan. 2014. Fashion Parsing With Weak Color-Category Labels. *IEEE Trans. Multimedia* 16 (2014), 253–265.

[20] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*. 3330–3337.

[21] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. 2008. A New Baseline for Image Annotation. In *ECCV*. 316–329.

[22] K. Murphy. 2012. *Machine Learning: a Probabilistic Perspective.* MIT Press.

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*. 815–823.

[24] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. 2015. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *CVPR*. 869–877.

[25] Edgar Simo-Serra and Hiroshi Ishikawa. 2016. Fashion Style in 128 Floats: Joint Ranking and Classification Using Weak Data for Feature Extraction. In *CVPR*. 298–307.

[26] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[27] Evgeniya Ustinova and Victor S. Lempitsky. 2016. Learning Deep Embeddings with Histogram Loss. In *NIPS*. 4170–4178.

[28] Laurens van der Maaten and Geoffrey E. Hinton. 2014. User's Guide for t-SNE Software. (2014).

[29] Sirion Vittayakorn, Kota Yamaguchi, Alexander C. Berg, and Tamara L. Berg. 2015. Runway to Realway: Visual Analysis of Fashion. In *WACV*. 951–958.

[30] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning Fine-Grained Image Similarity with Deep Ranking. In *CVPR*. 1386–1393.

[31] Kilian Q. Weinberger, John Blitzer, and Lawrence K. Saul. 2005. Distance Metric Learning for Large Margin Nearest Neighbor Classification. In *NIPS*. 1473–1480.

[32] Jason Weston, Samy Bengio, and Nicolas Usunier. 2011. WSABIE: Scaling Up to Large Vocabulary Image Annotation. In *IJCAI*. 2764–2770.

[33] Fei Wu, Zhuhao Wang, Zhongfei Zhang, Yi Yang, Jiebo Luo, Wenwu Zhu, and Yueting Zhuang. 2015. Weakly Semi-Supervised Deep Learning for Multi-Label Image Annotation. *IEEE Trans. Big Data* 1, 3 (2015), 109–122.

[34] Kota Yamaguchi, Tamara L. Berg, and Luis E. Ortiz. 2014. Chic or Social: Visual Popularity Analysis in Online Fashion Networks. In *ACM Multimedia*. 773–776.

[35] Kota Yamaguchi, M. Hadi Kiapour, and Tamara L. Berg. 2013. Paper Doll Parsing: Retrieving Similar Styles to Parse Clothing Items. In *ICCV*. 3519–3526.

[36] Kota Yamaguchi, M. Hadi Kiapour, Luis E. Ortiz, and Tamara L. Berg. 2012. Parsing clothing in fashion photographs. In *CVPR*. 3570–3577.

[37] Wei Yang, Ping Luo, and Liang Lin. 2014. Clothing Co-parsing by Joint Image Segmentation and Labeling. In *CVPR*. 3182–3189.

[38] Xiaofan Zhang, Feng Zhou, Yuanqing Lin, and Shaoting Zhang. 2016. Embedding Label Structures for Fine-Grained Feature Representation. In *CVPR*. 1114–1123.