Pedestrian Color Naming via Convolutional Neural Network

Zhiyi Cheng Xiaoxiao Li Chen Change Loy

Department of Information Engineering, The Chinese University of Hong Kong

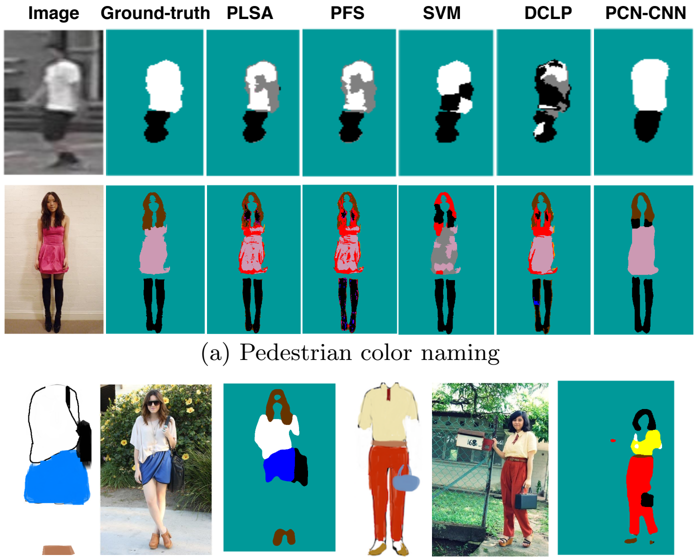
Abstract. Color serves as an important cue for many computer vision tasks. Nevertheless, obtaining accurate color description from images is non-trivial due to varying illumination conditions, view angles, and surface reflectance. This is especially true for the challenging problem of pedestrian description in public spaces. We made two contributions in this study: (1) We contribute a large-scale pedestrian color naming dataset with 14,213 hand-labeled images. (2) We address the problem of assigning consistent color name to regions of single object's surface. We propose an end-to-end, pixel-to-pixel convolutional neural network (CNN) for pedestrian color naming. We demonstrate that our Pedestrian Color Naming CNN (PCN-CNN) is superior over existing approaches in providing consistent color names on real-world pedestrian images. In addition, we show the effectiveness of color descriptor extracted from PCN-CNN in complementing existing descriptors for the task of person re-identification. Moreover, we discuss a novel application to retrieve outfit matching and fashion (which could be difficult to be described by keywords) with just a user-provided color sketch.

1 Introduction

Color naming aims at mapping image pixels' RGB values to a pre-defined set of basic color terms[1], *e.g.*, 11 basic color terms defined by Berlin and Kay [5] - black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow. Color names have been widely used as a type of color descriptor for a variety of applications such as image retrieval and image classification [38]. Recent studies [18, 25, 41] have applied color naming for the task of person re-identification [11, 13, 17, 21, 22, 43, 45] to achieve robust person matching under varying illuminations. Automatic color naming has also been exploited for cloth retrieval and fashion parsing [24].

In this study, we focus on the task of assigning consistent color names to pedestrian images captured from public spaces (see Figure 1). This task is non-trivial since the observed color of different parts of a pedestrian's body surface can look totally different under disparate illuminant conditions and view angles. In addition, strong highlights and shadows can make the RGB values of the same

[1] A basic color term is defined as being not subsumable to other basic color terms and extensively used in different languages.



(a) Pedestrian color naming

(b) Color sketch-based fashion retrieval

Fig. 1: (a) State-of-the-art non-learning based color naming method, DCLP [26] and learning-based approaches based on hand-crafted features, including SVM [24] and PLSA [38] fail to extract accurate color names for different regions. In contrast, the proposed Pedestrian Color Naming Convolutional Neural Network (PCN-CNN) generates color labels consistent with the ground truth (PCN-CNN generates color labels over its own predicted foreground region while other methods use ground-truth foreground mask). (b) A meaningful application of our method is the retrieval of outfit matching based on a simple user-provided color sketch (from left to right: sketch, retrieved image, and the corresponding estimated color names map). The application is demonstrated in Section 5.3.

surface span from light to dark. Creases and folds in clothing surface can also lead to drastically different predictions of color. Some examples are shown in Figure 1. Existing methods are not effective for this kind of challenging scenarios. Specifically, some of these approaches are non-learning-based methods [26], they thus cannot effectively capture the uncontrollable variations for specific scenarios. Some other methods rely on hand-crafted features and color histograms, *e.g.*, LAB color space [38], SIFT and HOG features [24], which may have limited expressive power to represent the image content (more details in Section 2).

We believe that the key to address the aforementioned problem is a model that is capable of extracting meaningful representation to achieve color constancy [2, 6, 10], *i.e.* the capability of inferring the true color distribution intrinsic to the surface. Such a representation needs to be learned from a large-scale training set to ensure robustness for real-world scenes. To this end, we make two main contributions:

– *A large-scale dataset* - Existing color naming datasets either lack of sufficient training samples or do not come with pixel-level annotation (see Section 3). To facilitate the learning and evaluation of pedestrian color naming, we introduce a large-scale dataset with careful manual segmentation and region-wise color annotation. The dataset contains 14,213 images in total, which is the largest color naming dataset that we aware to our knowledge. All the images are collected under challenging surveillance scenarios (Market-1501 dataset [44]), with large variations in illumination, highlights, shadow changes, different pedestrian poses and view angles. We show that the dataset is essential for pre-training a color naming deep network for a number of pedestrian-related applications, including person re-identification and cloth retrieval.

– *End-to-end color naming* - We propose a Pedestrian Color Naming Convolutional Neural Network (PCN-CNN) to learn pixel-level color naming. In contrast to existing studies [24, 38] that require independent components for feature extraction and color mapping, our CNN-based model is capable of extracting strong features and regressing for color label for each pixel in an end-to-end framework. Conditional random field (CRF) is further adopted to smooth the pixel-wise color predictions. Our network is specially designed to handle images with low resolution, and hence it is well-suited for processing pedestrian images captured from low-resolution surveillance cameras.

Extensive results on the Market-1501 [44] and Colorful-Fashion [24] datasets show the superiority of our approach over existing color naming methods [4, 24, 26, 38]. We further show the applicability of PCN-CNN in complementing existing visual descriptors for the task of person re-identification (Re-ID). In particular, we demonstrate consistent improvement using the PCN-CNN features in conjunction with different existing Re-ID approaches. In addition, we also highlight an interesting application for outfit matching retrieval. In particular, in the absence of imagery or keyword query, we show that it is possible to retrieve desired fashion images from a gallery through just a simple and convenient 'color sketch'. An example is depicted in Figure 1(b). Such a color-driven query provides rich region-wise color description and can be used in conjunction with visual attribute-driven query [21, 28] for 'zero-shot' retrieval.

2 Related Work

Color naming. Benavente et al. [4] proposed a pixel-wise color naming model based on lightness and chromaticity distribution, which did not consider cross-pixel relations and intrinsic consistency. Serra et al. [35] and Liu et al. [26] improve the region consistency of color names based on this pixel-wise color naming results. In particular, Serra et al. [35] applied CRF to infer the color intrinsic components from images. They extracted the intrinsic information according to the segmentation results of Ridge Analysis of color Distribution (RAD) [39], and assigned the same color label to pixels connected by a ridge. However, the RAD method only described the RGB histogram distribution and may fail to handle the complicated color distribution. Besides, only with ridge information, the method cannot reliably predict the correct color label from a region if a big portion of the surface's pixels are affected by shadows or highlights. Liu et al. [26] applied the similar CRF model and built a label propagation model where the color labels of pixels in normal region will be propagated to those shadowed and highlighted regions in the same objects' surface. However, their model relied on the detection results of highlights and shadows [16, 20, 31, 37, 39] with mainly the intensive and reflectance information, which do not suit for complicated color distribution cases, especially for the challenging pedestrians under real-world settings.

Van de Weijer et al. [38] used LAB histogram features as 'words' and applied them into a Probabilistic Latent Semantic Analysis (PLSA) model to learn for 'topic' color naming. Liu et al. [24] designed a concatenated feature by RGB, LAB color spaces and SIFT, HOG features. Mojsilovic [32] built a multi-level color description model and estimated color naming combined with segmentation. However, this work did not address the issues of shadowed and highlighted regions. All of these hand-crafted features lack robustness to dramatic illumination changes.

Pedestrian descriptors. Person re-identification [13] aims at recognizing the same individual under different camera views. To tackle the challenging appearance changes by varying viewpoints, illumination and poses, many researchers have proposed different pedestrian descriptors. Gray et al. [15] introduced an ensemble of localized features (ELF) consisting of colors and textures for viewpoint robustness. Layne et al. [21] proposed to use mid-level semantic attributes, fused with low-level features in ELF to obtain improved results. Bazzani et al. [3] exploited three complementary aspects of the human appearance: the overall chromatic content, the spatial arrangement of colors into stable regions, and the presence of recurrent local motifs with high entropy. More recently, a 'mirror representation' [9] is proposed to explicitly model the relation between different view-specific transformations. Chen et al. [7] proposed a Spatially Constrained Similarity function on polynomial feature map and achieved a new state of the art results. Recent studies have explored the illuminant-invariant color distribution descriptors for Re-ID. Kviatkovsky et al. [19] introduced log-chromaticity color space to identify persons under varying scenes. To complement the traditional color information, color naming has been applied to recent studies [18, 41, 25] and achieved improvements over the state-of-the-art models. Kuo et al. [18] employed the semantic color names learned by [38]. Yang et al. [41] employed the salient color names according to RGB values. However, these applied color naming models did not show region consistency and had limited robustness to dramatic illumination changes.

3 Pedestrian Color Naming Dataset

A well-segmented region-level color naming dataset is essential for both model training and evaluation. A dataset collected from realistic scenes, with diverse illumination, highlights and shadows, and varying view angles, counts heavily to the success of pedestrian color naming.

Table 1: The train/test distribution of 11 basic colors at region level (arranged in alphabetical order) of the Pedestrian Color Naming (PCN) dataset.

	black	blue	brown	grey	green	orange	pink	purple	red	white	yellow
train	8040	2192	722	3256	1576	318	1138	669	1972	5013	1651
test	1264	275	113	491	234	37	139	109	249	710	202

There is no public large-scale color naming dataset with pixel-level labels. The Google Color Name [38] and Google-512 datasets [34] contain 1100 and 5632 images, respectively, but both of them are weakly labeled with only image-level color annotations. The Object dataset [26] and Ebay dataset [38] include 350 and 528 images with region-level color annotation. These datasets are far from enough for learning and testing a CNN-based color naming model.

To facilitate the learning of evaluation of pedestrian color naming, we build a new large-scale dataset, named *Pedestrian Color Naming* (PCN) dataset, which contains 14,213 images, each of which hand-labeled with color label for each pixel. The dataset and the annotations can be downloaded at http://mmlab.ie.cuhk.edu .hk/projects/PCN.html.

Image collection. All images in the PCN dataset are obtained from the Market-1501 dataset [44]. The original Market-1501 dataset consists of pedestrian images of 1,501 identities, captured from a total of six surveillance cameras. Each identity has multiple images with varying scene settings and poses under multiple camera views. These images contain strong highlights and shadows with various illumination conditions and view angles. We carefully select a subset of 14,213 images which have good visibility of the full body and diverse color distribution. We consequently divide the dataset into a training set of 10,913 images, a validation set of 1,500 images and the remaining 1,800 images for testing. Table 1 summarizes the distribution of the different color labels in both the training and test subsets. Note that there may be multiple colors co-exist in the same image. Some colors, namely purple and orange, are relatively lower in numbers since pedestrians tend to wear clothes with more common colors such as black and white.

Super-pixel-driven annotation. Pixel-by-pixel labeling of color labels is a tedious task. We attempted this possibility but found it not scalable. To overcome this problem, we first oversegment each image into 100 super-pixels through the popular SLIC superpixel segmentation method [1]. We found that the super-pixels align well with the object contours most of the time. We then carefully identify the color label for each super-pixel following the 11 color names defined by Berlin and Kay [5], excluding the background, human skin, and hair areas. Note that some super-pixels are originated from the same region (*e.g.* different regions of a pair of jeans). We manually group these super-pixels together to form a single region. Eventually, each coherent region shares the same color label, and the labels for all regions collectively form a label map with the same resolution as of the associated image. Figure 2 depicts some example images and their corresponding pixel-level color label maps.

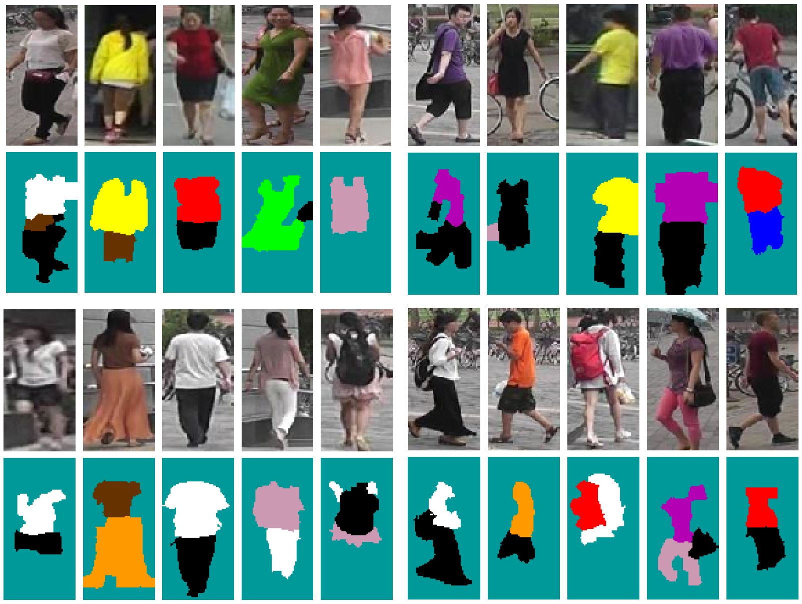


Fig. 2: Some examples of labeled images in the Pedestrian Color Naming dataset. The images in the first row are the original images and those in the second row are the color label maps where each region is visualized using the corresponding basic colors (black, blue, brown, grey, green, orange, pink, purple, red, white, and yellow). The background region is shown in dark cyan.

4 Pedestrian Color Naming Convolutional Network

Problem formulation. Given a pedestrian image **I**, our goal is to assign each pixel of **I** with a specific color name. Specifically, we define a binary latent variable $y_c^i \in \{0, 1\}$, indicating whether an i-th pixel should be named with a color name c, where $\forall c \in \mathcal{C} = \{1, 2, \ldots, 11\}$, representing the 11 basic color names [5].

We approach this problem in a general CRF [12] framework with the unary potentials generated by a deep convolutional network. The energy function of CRF is written as

$$E(\mathbf{y}) = \sum_{\forall i \in \mathcal{V}} U(y_c^i) + \sum_{\forall i,j \in \mathcal{E}} \pi(y_c^i, y_d^j), \tag{1}$$

where $\mathbf{y}$, $\mathcal{V}$, and $\mathcal{E}$, represent a set of latent variables, nodes, and edges in an undirected graph. Here, each node represents a pixel in image **I** and each edge captures the relation between pixels. The $U(y_c^i)$ measures the unary cost of assigning a label c to the i-th pixel, and $\pi(y_c^i, y_d^j)$ is the pairwise term that quantifies the penalty of assigning labels c, d to pixels i, j respectively. We define the unary term in Eq. (1) as

$$U(y_c^i) = -\ln p(y_c^i = 1 | \mathbf{I}), \tag{2}$$

where $p(y_c^i = 1|\mathbf{I})$ represents the probability of assigning label c to i-th pixel. In this study, we model the probability using PCN-CNN, which will be described next.

For the pairwise term, we let $\pi(y_c^i, y_d^j) = \mu(u,v)D(i,j)$, where $\mu(u,v)$ represents a prior color co-occurrence. Although this prior can be learned from data, to simplify the problem we make a mild assumption that $\mu(u,v) = 1$ for any arbitrary pair of color labels. The $D(i,j)$ measures the distances between pixels,

$$D(i,j) = w_1||f(\mathbf{I}_i) - f(\mathbf{I}_j)||^2 + w_2||(x_i, y_i), (x_i, y_j)||^2, \tag{3}$$

where f is a function that extracts features from the i-th pixel, *e.g.*, RGB values, while (x,y) denote the coordinates of a pixel, and w_1, w_2 are constant weights. The pairwise term encourages pixels that are close and similar to each other to share the same color label.

Network architecture. Deep convolutional network has shown immense success for various image recognition tasks. Different from existing problems, we need to cope with a few unique challenges. Firstly, we need to deal with the background clutter, which is detrimental to the foreground color prediction. Secondly, our problem requires special care in designing the architecture since pedestrian images are typically low in resolution, *e.g.* 128 × 64 in the Market-1501 dataset [44]. This challenge is especially crippling since most off-the-shelf deep networks contain pooling layers that could significantly reduce the effective size of the input images. We cannot afford this information loss.

Consequently, we based our solution on the VGG_{16} network [36] but with the following modifications. To handle the background clutter, we additionally consider background as a label and train the network to jointly estimate for both foreground-background segmentation and color naming, resulting in 12-category output. That is, the network output has 11 color names and a background indicator $b \in \{0,1\}$ to indicate the presence of background at a pixel.

To handle the small input resolution issue, we need to modify the VGG_{16} network. We still initialize the filters in our network with all the learned parameters to make full use of VGG_{16} pre-trained by ImageNet. Nevertheless, for the pixel-to-pixel prediction of low-resolution input, more information should be preserved. Table 2 compares the hyper-parameters of the VGG_{16} network and our network. We use ai and bi to denote the i-th group in Table 2(a) and 2(b). Our network contains 13 convolutional layers, two max-pooling layers, and the last three layers act as the fully convolutional layers and de-convolutional layers, which generates the final labeling results. As summarized in Table 2, we increase the resolution of convolved data by removing three max-pooling layers from VGG_{16}. As a result, the smallest size of feature map in our model is 32×16 (based on input-size of 128×64), keeping more information compared with VGG_{16}.

Filters of $b6$ are initialized with the filers of $a7$, where each filter in $a7$ should be convolved with $a5$ on a stride (the stride length is 2). Since the max-pooling layer $a6$ has been removed, the 3×3 receptive filed is padded into 5×5 with zeros every other parameter in the filter, to keep the resolution identical to one-stride

Table 2: The comparisons between VGG_{16} and our PCN-CNN, as shown in (a) and (b) respectively. The 'fs', '#cha', 'act' and 'size' represent the filter stride size, number of output feature maps, activation function, and size of output feature maps, respectively. And 'conv', 'max', 'dconv', and 'fc' represent the convolution layer, max-pooling layer, deconvolution layer, and fully-connected layer, respectively. The 'relu', 'idn' and 'soft' represent the rectified linear unit, identity and softmax activation functions.

(a) VGG_{16}: 224×224×3 input image; 1×1000 output labels.

	1	2	3	4	5	6	7	8	9	10	11	12
layer	2×**conv**	**max**	2×**conv**	**max**	3×**conv**	**max**	3×**conv**	**max**	3×**conv**	**max**	2×**fc**	**fc**
fs	3-1	2-2	3-1	2-2	3-1	2-2	3-1	2-2	3-1	2-2	-	-
#cha	64	64	128	128	256	256	512	512	512	512	1	1
act	relu	idn	relu	idn	relu	idn	relu	idn	relu	idn	relu	soft
size	224	112	112	56	56	28	28	14	14	7	4096	1000

(b) Our PCN-CNN: 128×64×3 input image; 128×64×12 output label maps.

	1	2	3	4	5	6	7	8	9	10
layer	2×**conv**	**max**	2×**conv**	**max**	3×**conv**	3×**conv**	3×**conv**	**conv**	**conv**	**dconv**
fs	3-1	2-2	3-1	2-2	3-1	5-1	9-1	17-1	1-1	1-1
#cha	64	64	128	128	256	512	512	4096	4096	12
act	relu	idn	relu	idn	relu	relu	relu	relu	relu	soft
size	128×64	64×32	64×32	32×16	32×16	32×16	32×16	32×16	32×16	128×64

convolution. The following convolutional layers are padded in the similar way. For the fully convolutional layer $b8$, if all the 7×7 parameters are to be applied for initialization, a padded 49×49 receptive filed is needed in the similar way, which needs more padding to the input feature map to keep the output size after up-sampling. Since large zero padding can affect the performance, we down-sample the parameters of receptive field [8] from 7×7 to 3×3 before applying them for initialization. In this way, the padded 17×17 with zeros from 3×3 is applied in $b8$ as the fully convolutional layer. Finally, the $b10$ layer up-samples the feature maps to 128×64 by bilinear interpolation, and generates the 12-dimensional prediction for each pixel (11 color + background labels).

It is worth pointing out that deep convolutional network has been widely used for image segmentation [8, 27, 29]. Differs from these prior studies, our work is the first attempt to use CNN for color naming. In terms of network architecture, our network shares some similarity to the Deep Parsing Network (DPN) [27]. Unlike DPN that accepts input image of resolution 512 × 512, we design our network to accommodate for small pedestrian images and remove pooling layers to avoid information loss. We attempted to enlarge pedestrian images to fit DPN's requirement but the performance of this alternative is inferior to that achieved by our final design.

Training details are given as follows. We start with an initial learning rate of 0.001, and reduce it by a factor of 10 at every 5K iterations. We use a momentum of 0.9, and mini-batches of 12 images.

5 Experiments

In this section, we first evaluate PCN-CNN's performance for color naming. We also examine the effectiveness of color names descriptor extracted from PCN-CNN for the task of person re-identification. Furthermore, we show an interesting application with PCN-CNN, using only simple sketches as probe to retrieve desired outfit matching of fashion images from a real-world image gallery.

5.1 Pedestrian Color Naming

In this experiment, we analyze PCN-CNN's performance for pedestrian color naming.

Datasets. We perform evaluations on the proposed PCN dataset (relabelled Market 1501 dataset [44]) and a cloth dataset, Colorful-Fashion [24], both of which have a test subset of 1,800 and 2,682 images, respectively. The PCN dataset is challenging due to its low image resolution (128 × 64) and large variations in terms of illumination and pedestrian pose. The Colorful-Fashion dataset contains images with a higher resolution (600 × 400), but the cloth patterns are more complex and colorful. Images in the Colorful-Fashion dataset comes with region-wise color labels. Note that the dataset also annotates hair pixels with color names, we therefore include the hair region estimation in our evaluation. For the PCN dataset, we label the color names based on the procedure described in Section. 3.

Evaluation metrics. To measure the performance of both the pixel-wise and region-wise accuracies, we apply two metrics for model evaluation:
(1) *Pixel Annotation Score* (PNS) - this score [38] measures the percentage of correctly predicted color names at pixel level. We average the PNS for all regions as the final score to measure the consistency of color naming.
(2) *Region Annotation Score* (RNS) - each region's color label is specified by its dominant color names prediction of pixels. We then calculate the averaged accuracy of prediction at the region level.

Results. We compare our PCN-CNN against with state-of-the-art methods, including PLSA [38], PFS [4], SVM-based color classifier [24] and DCLP [26]. Besides, we also adopt CRF to smooth PCN-CNN color names prediction and evaluate the performance. For a better foreground estimation on pedestrian images, the PCN-CNN is first pre-trained on the large-scale pedestrian parsing dataset PPSS [30], which encourages the network to generate binary map composed of pedestrian region and the background. The pre-trained parameters of PCN-CNN are then fine-tuned on the training partition of the PCN dataset and Colorful-Fashion dataset, respectively, for the respective tests on the two datasets. Likewise, all learning based methods, *e.g.* PLSA and SVM, are re-trained using the same training partition employed by PCN-CNN to ensure a fair comparison. It is worth pointing out that during the evaluation of PCN-CNN, we employ the foreground masks generated by itself before applying the evaluation metrics. For other baselines (PLSA, PFS, SVM, and DCLP), we use

Table 3: Performance over PCN and Colorful-Fashion test set. PNS and RNS denote the averaged pixel annotation score and region annotation score respectively. The smoothed color names prediction is denoted by PCN-CNN+CRF.

Method	PCN		Colorful-Fashion	
	PNS	RNS	PNS	RNS
PLSA [38]	63.1	68.4	57.4	71.4
PFS [4]	61.1	68.5	48.6	60.5
SVM [24]	62.8	62.2	43.5	45.4
DCLP [26]	56.8	62.0	47.8	54.8
PCN-CNN	74.1	80.3	70.2	81.8
PCN-CNN+CRF	**74.3**	**80.8**	**71.1**	**81.9**

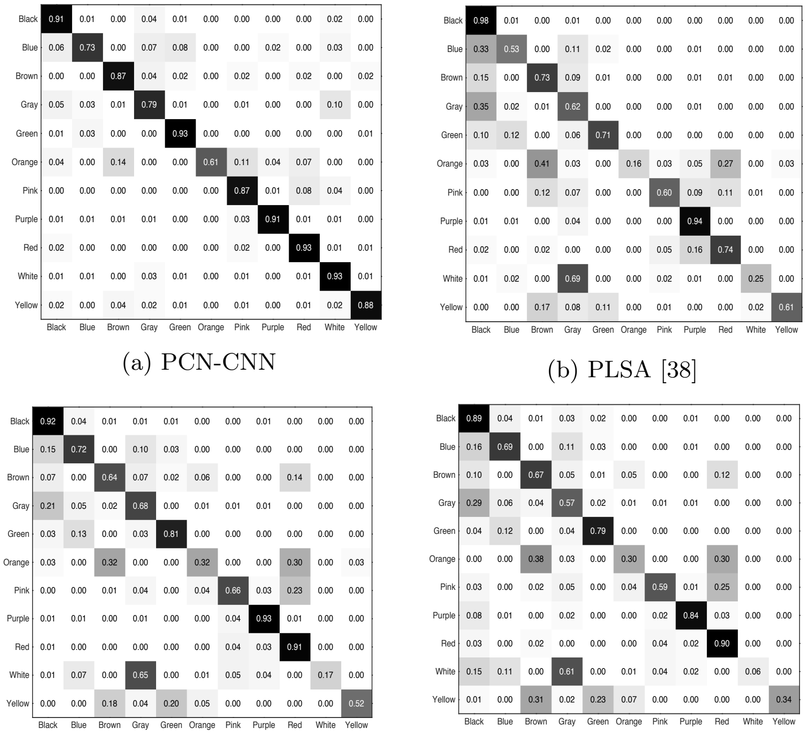
the ground-truth masks for this purpose. Given the more accurate masks compared to PCN-CNN generated ones, these baselines therefore gain additional advantages than PCN-CNN.

Table 3 and Figure 3 show the performance comparison and confusion matrix (based on RNS), respectively, among different methods. Qualitative results are provided in Figure 4. As shown in the experimental results, our model achieves superior performance in both PNS and RNS metrics, with outstanding robustness to shadows and highlights, creases and folds. Adding CRF to PCN-CNN further boosts its performance.

5.2 Color Naming for Person Re-identification

Pedestrian color naming provides a powerful feature for person re-identification, even in low-resolution images, due to its robustness to varying illumination and view angles. A robust color naming model with good consistency helps to describe people more accurately by ignoring the minor change in RGB values. In this section, we combine the region-level color names generated by PCN-CNN with several existing visual descriptors for the task of person re-identification, and test the performance on the widely used VIPeR dataset [14].

Feature representation. Similar to [23], we first partition a pedestrian image into six equal-size horizontal stripes, represented as $H = [h_1, ..., h_6]^\mathsf{T}$. For the i-th part h_i, we use a histogram of color names as the feature representation, resulting into a 66-dimensional descriptor for all the parts. The c-th bin of a histogram h_i denotes the probability of all pixels in the corresponding part being assigned to color name c. To minimize the influence of background clutter, we only extract the color distribution of the foreground region. The estimated feature is called pedestrian color naming (PCN) descriptor in the following session. We concatenate the PCN descriptor with several representative visual descriptors for person re-identification. These include one of the most widely used features called ensemble of localized features (ELF) [15, 21]; a pure color-based features, named salient color names based color descriptor (SCNCD) [41]; a recent advanced features known as mirror representation [9]. The original ELF, SCNCD and 'mirror representation' descriptors and those concatenated with PCN descriptor are fed



(a) PCN-CNN

	Black	Blue	Brown	Gray	Green	Orange	Pink	Purple	Red	White	Yellow
Black	0.91	0.01	0.00	0.04	0.01	0.00	0.00	0.00	0.00	0.02	0.00
Blue	0.06	0.73	0.00	0.07	0.08	0.00	0.00	0.02	0.00	0.03	0.00
Brown	0.00	0.00	0.87	0.04	0.02	0.00	0.02	0.00	0.02	0.00	0.02
Gray	0.05	0.03	0.01	0.79	0.01	0.00	0.01	0.00	0.00	0.10	0.00
Green	0.01	0.03	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.00	0.01
Orange	0.04	0.00	0.14	0.00	0.00	0.61	0.11	0.04	0.07	0.00	0.00
Pink	0.00	0.00	0.00	0.00	0.00	0.00	0.87	0.01	0.08	0.04	0.00
Purple	0.01	0.01	0.01	0.01	0.00	0.00	0.03	0.91	0.01	0.01	0.00
Red	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.93	0.01	0.01
White	0.01	0.01	0.00	0.03	0.01	0.00	0.01	0.00	0.01	0.93	0.01
Yellow	0.02	0.00	0.04	0.02	0.01	0.00	0.01	0.00	0.01	0.02	0.88

(b) PLSA [38]

	Black	Blue	Brown	Gray	Green	Orange	Pink	Purple	Red	White	Yellow
Black	0.98	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Blue	0.33	0.53	0.00	0.11	0.02	0.00	0.00	0.01	0.00	0.00	0.00
Brown	0.15	0.00	0.73	0.09	0.01	0.00	0.00	0.01	0.01	0.00	0.00
Gray	0.35	0.02	0.01	0.62	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Green	0.10	0.12	0.00	0.06	0.71	0.00	0.00	0.00	0.00	0.00	0.00
Orange	0.03	0.00	0.41	0.03	0.00	0.16	0.03	0.05	0.27	0.00	0.03
Pink	0.00	0.00	0.12	0.07	0.00	0.00	0.60	0.09	0.11	0.01	0.00
Purple	0.01	0.01	0.00	0.04	0.00	0.00	0.00	0.94	0.00	0.00	0.00
Red	0.02	0.00	0.02	0.00	0.00	0.00	0.05	0.16	0.74	0.00	0.00
White	0.01	0.02	0.00	0.69	0.00	0.00	0.02	0.01	0.00	0.25	0.00
Yellow	0.00	0.00	0.17	0.08	0.11	0.00	0.01	0.00	0.00	0.02	0.61

(c) PFS [4]

	Black	Blue	Brown	Gray	Green	Orange	Pink	Purple	Red	White	Yellow
Black	0.92	0.04	0.01	0.01	0.01	0.00	0.00	0.01	0.00	0.00	0.00
Blue	0.15	0.72	0.00	0.10	0.03	0.00	0.00	0.00	0.00	0.00	0.00
Brown	0.07	0.00	0.64	0.07	0.02	0.06	0.00	0.00	0.14	0.00	0.00
Gray	0.21	0.05	0.02	0.68	0.00	0.01	0.01	0.01	0.00	0.00	0.00
Green	0.03	0.13	0.00	0.03	0.81	0.00	0.00	0.00	0.00	0.00	0.00
Orange	0.03	0.00	0.32	0.00	0.00	0.32	0.00	0.00	0.30	0.00	0.03
Pink	0.00	0.00	0.01	0.04	0.00	0.04	0.66	0.03	0.23	0.00	0.00
Purple	0.01	0.01	0.00	0.01	0.00	0.00	0.04	0.93	0.01	0.00	0.00
Red	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.03	0.91	0.00	0.00
White	0.01	0.07	0.00	0.65	0.00	0.01	0.05	0.04	0.00	0.17	0.00
Yellow	0.00	0.00	0.18	0.04	0.20	0.05	0.00	0.00	0.00	0.00	0.52

(d) DCLP [26]

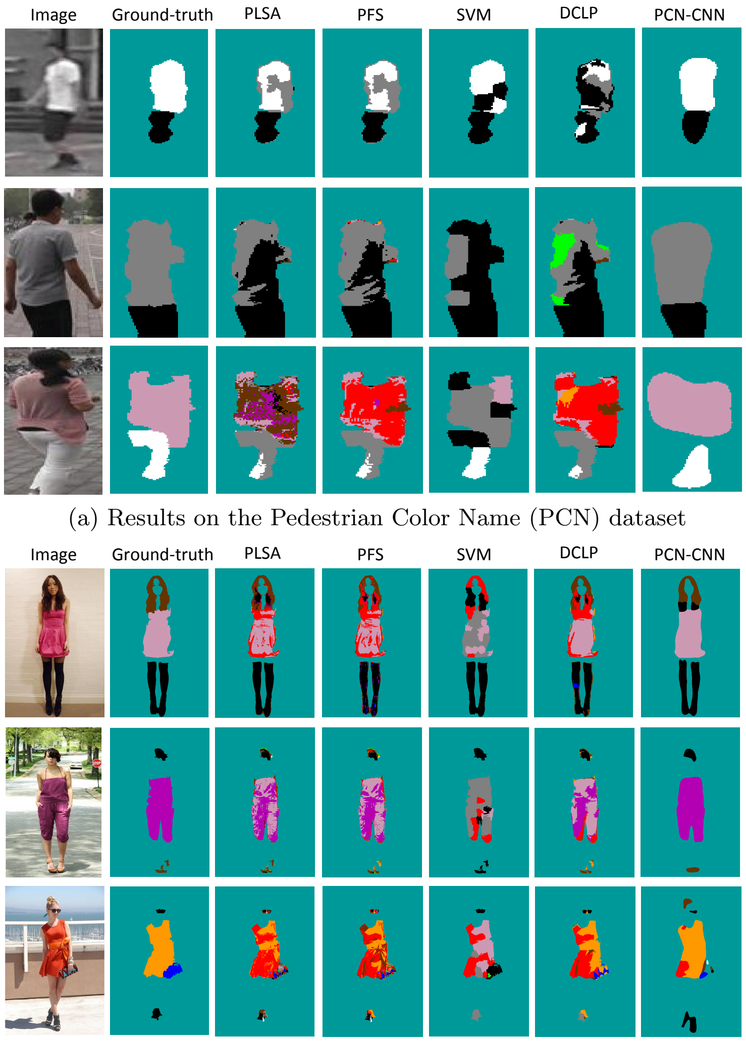
	Black	Blue	Brown	Gray	Green	Orange	Pink	Purple	Red	White	Yellow
Black	0.89	0.04	0.01	0.03	0.02	0.00	0.00	0.01	0.00	0.00	0.00
Blue	0.16	0.69	0.00	0.11	0.03	0.00	0.00	0.01	0.00	0.00	0.00
Brown	0.10	0.00	0.67	0.05	0.01	0.05	0.00	0.00	0.12	0.00	0.00
Gray	0.29	0.06	0.04	0.57	0.02	0.01	0.01	0.01	0.00	0.00	0.00
Green	0.04	0.12	0.00	0.04	0.79	0.00	0.00	0.00	0.00	0.00	0.00
Orange	0.00	0.00	0.38	0.03	0.00	0.30	0.00	0.00	0.30	0.00	0.00
Pink	0.03	0.00	0.02	0.05	0.00	0.04	0.59	0.01	0.25	0.00	0.00
Purple	0.08	0.01	0.00	0.02	0.00	0.00	0.02	0.84	0.03	0.00	0.00
Red	0.03	0.00	0.02	0.00	0.00	0.00	0.04	0.02	0.90	0.00	0.00
White	0.15	0.11	0.00	0.61	0.00	0.01	0.04	0.02	0.00	0.06	0.00
Yellow	0.01	0.00	0.31	0.02	0.23	0.07	0.00	0.00	0.00	0.00	0.34

Fig. 3: Confusion matrix of color naming (regional level) on the Pedestrian Color Naming dataset.

into the KMFA metric learning method [40] for matching. Moreover, the PCN feature is also fed into a recent outperforming similarity learning method with spatial constraints (SCSP) [7], fused with other originally used visual cues.

Experiment settings. The VIPeR dataset contain 632 pedestrian image pairs, with varying illumination conditions and view angles. Each pedestrian has two images per camera view. All the images are normalized to 128×48 pixels. We randomly choose half of the image pairs for training and the others for testing. This procedure is repeated for 10 evaluation trials. Averaged performance is measured over the trials by using the typical cumulative matching characteristic (CMC) curve. In particular, we report the rank k matching rate, which refers to the percentage of probe images that are correctly matched with the true positives in the gallery set in the top k rank.

Results. As can be observed from Table 4, the PCN descriptor is capable of improving the performance of a wide range of existing Re-ID visual descriptors, from ensemble of color/texture features (ELF), pure color based features (SCNCD), as well as the more elaborated mirror representation. Moreover, a new state-of-the-art accuracy can be achieved by SCSP learning method, when concatenating with the PCN descriptor. It is interesting to see that PCN descriptors



(a) Results on the Pedestrian Color Name (PCN) dataset

(b) Results on Colorful-fashion dataset

Fig. 4: Qualitative results on the PCN and Colorful-Fashion datasets. The background is indicated by dark cyan. PCN-CNN generates color labels over its own learned foreground region while other methods use ground-truth foreground mask.

Table 4: Comparative results between original person re-identification descriptors vs. descriptors enhanced with PCN descriptor. Results are reported on the VIPeR dataset.

Rank k	1	5	10	20
ELF [15]	23.77	51.17	64.62	78.89
ELF [15]+PCN	**36.36**	**68.92**	**82.69**	**92.63**
SCNCD [41]	21.33	38.86	49.02	59.91
SCNCD [41] +PCN	**28.45**	**52.09**	**64.11**	**75.03**
Mirror-KMFA [9]	42.97	75.82	87.28	94.84
Mirror-KMFA [9]+PCN	**45.03**	**77.56**	**89.05**	**96.04**
SCSP [7]	53.54	82.59	**91.49**	96.65
SCSP [7]+PCN	**54.24**	**82.78**	91.36	**99.08**

yields large improvement to the SCNCD method, which is also based on color names. The results suggest the robustness of our approach in complementing existing pedestrian descriptors.

5.3 Color Naming for Zero-Shot Cloth Retrieval

One may relatively often has to do with combining different colors of shirts, pants, and shoes together. Or one might want to purchase a particular piece of garment in mind but do not know how to describe its combination of colors and patterns. Instead of elucidating a long textual description of it, one could just draw a sketch! A recent paper [42] has applied this idea for fine-grained shoe retrieval using monochrome sketches. In this section, we show the possibility to 'retrieve with colors'.

Specifically, one simply needs to paint with a few strokes the desired color on a sketch with specific combinations and patterns. The sketch can then serve as a query for cloth/fashion retrieval. This is possible through the following steps: we process a color sketch using PCN-CNN to transform it into a map with 11 color names, and further convert it into a PCN histogram (see Section 5.2). We assume all the gallery images have been processed in the same way. We then apply histogram intersection to measure the similarity of features for retrieval.

Experiment settings. A total of 80 images with rich color and complex patterns are selected from the Colorful-Fashion dataset[24], and we ask volunteers to draw for the corresponding color sketches. The task is to use the sketch as query and correctly retrieve the true image among the 2,682 test images of Colorful-Fashion dataset. The top-k retrieval accuracy is adopted as the metric.

Results. Table 5 shows the cloth retrieval results with different color naming models. Thanks to the robustness of PCN-CNN, our method achieves an impressive top-1 retrieval rate of 42.86%, surpassing other baselines. Some qualitative results are shown in Figure 5, in which we compare the retrieved results and generated color map of our PCN-CNN and PLSA. With poorer region-level consistency compared to PCN-CNN, which is critical for the cloth retrieval task, PLSA can easily fail to retrieve the ground-truth matching clothes with strong highlights and shadows.

Table 5: Top-k retrieval accuracy on Colorful-Fashion dataset (a subset is selected, see text for details) using color sketch as query.

Rank k	1	5	10	20
PLSA [38]	6.49	12.99	16.99	23.38
PFS [4]	6.48	12.98	18.18	20.78
PCN-CNN	**42.86**	**68.83**	**72.73**	**83.12**

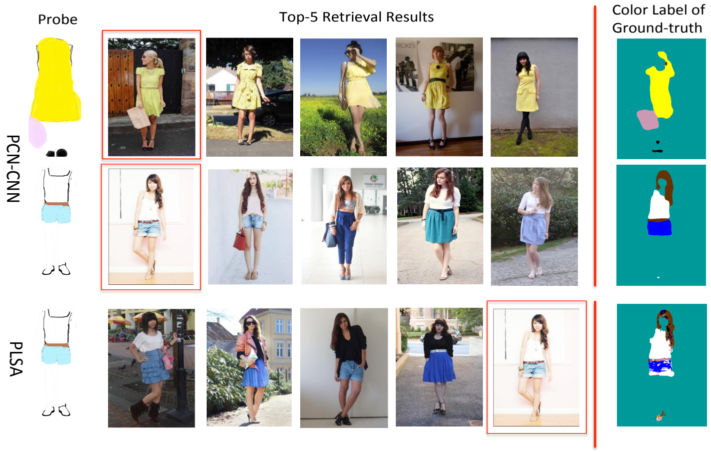


Fig. 5: We show the top-5 retrieval results with sketches as probes, using PCN-CNN and PLSA, respectively. The retrieved images highlighted with red boundary represent the ground-truth matching cloth images.

6 Conclusion

We have presented an end-to-end, pixel-to-pixel convolutional neural network for pedestrian color naming, named PCN-CNN. To facilitate model training and evaluation, we have introduced a large-scale pedestrian color naming dataset, containing 14,213 images with carefully labeled pixel-level color names. Extensive experiments show that the PCN-CNN is capable of generating consistent color name to clothing surfaces regardless of large variations in clothing material and illumination. The PCN descriptor extracted from the model is not only useful for complementing existing pedestrian descriptors, but also generalizable for sketch-to-image retrieval.

Acknowledgement

We would like to show our gratitude to the authors of [9], for sharing their features and codes of matching procedure for the person re-identification experiments.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 34(11), 2274–2282 (2012)
2. Barron, J.T.: Convolutional color constancy. In: International Conference on Computer Vision (ICCV) (2015)
3. Bazzani, L., Cristani, M., Murino, V.: Symmetry-driven accumulation of local features for human characterization and re-identification 117(2), 130–144 (2013)
4. Benavente, R., Vanrell, M., Baldrich, R.: Parametric fuzzy sets for automatic color naming. JOSA A 25(10), 2582–2593 (2008)
5. Berlin, B., Kay, P.: Basic color terms: Their universality and evolution. Univ of California Press (1991)
6. Bianco, S., Cusano, C., Schettini, R.: Single and multiple illuminant estimation using convolutional neural networks. arXiv preprint arXiv:1508.00998 (2015)
7. Chen, D., Yuan, Z., Chen, B., Zheng, N.: Similarity learning with spatial constraints for person re-identification. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
8. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014)
9. Chen, Y.C., Zheng, W.S., Lai, J.: Mirror representation for modeling view-specific transform in person re-identification. In: International Joint Conference on Artificial Intelligence (IJCAI) (2015)
10. Cheng, D., Price, B., Cohen, S., Brown, M.S.: Effective learning-based illuminant estimation using simple features. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
11. Farenzena, M., Bazzani, L., Perina, A., Murino, V., Cristani, M.: Person reidentification by symmetry-driven accumulation of local features. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
12. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. International Journal of Computer Vision (IJCV) 40(1), 25–47 (2000)
13. Gong, S., Cristani, M., Yan, S., Loy, C.C.: Person re-identification. Springer (2014)
14. Gray, D., Brennan, S., Tao, H.: Evaluating appearance models for recognition, reacquisition, and tracking. In: International Workshop on Performance Evaluation for Tracking and Surveillance (2007)
15. Gray, D., Tao, H.: Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: European Conference on Computer Vision (ECCV), pp. 262–275. Springer (2008)
16. Guo, R., Dai, Q., Hoiem, D.: Single-image shadow detection and removal using paired regions. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2033–2040 (2011)
17. Hirzer, M., Roth, P.M., Köstinger, M., Bischof, H.: Relaxed pairwise learned metric for person re-identification. In: European Conference on Computer Vision (ECCV) (2012)
18. Kuo, C.H., Khamis, S., Shet, V.: Person re-identification using semantic color names and rankboost. In: Winter Conference on Applications of Computer Vision (WACV) (2013)
19. Kviatkovsky, I., Adam, A., Rivlin, E.: Color invariants for person reidentification. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(7), 1622–1634 (2013)

20. Lalonde, J.F., Efros, A.A., Narasimhan, S.G.: Detecting ground shadows in outdoor consumer photographs. In: Computer Vision–ECCV 2010, pp. 322–335. Springer (2010)
21. Layne, R., Hospedales, T.M., Gong, S., Mary, Q.: Person re-identification by attributes. In: British Machine Vision Conference (BMVC) (2012)
22. Liao, S., Hu, Y., Zhu, X., Li, S.Z.: Person re-identification by local maximal occurrence representation and metric learning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
23. Liu, C., Gong, S., Loy, C.C., Lin, X.: Person re-identification: What features are important? In: European Conference on Computer Vision Workshop (2012)
24. Liu, S., Feng, J., Domokos, C., Xu, H., Huang, J., Hu, Z., Yan, S.: Fashion parsing with weak color-category labels. IEEE Transactions on Multimedia 16(1), 253–265 (2014)
25. Liu, X., Wang, H., Wu, Y., Yang, J., Yang, M.H.: An ensemble color model for human re-identification. In: Winter Conference on Applications of Computer Vision (WACV) (2015)
26. Liu, Y., Yuan, Z., Chen, B., Xue, J., Zheng, N.: Illumination robust color naming via label propagation. In: International Conference on Computer Vision (ICCV) (2015)
27. Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X.: Semantic image segmentation via deep parsing network. In: International Conference on Computer Vision (ICCV) (2015)
28. Liu, Z., Luo, P., Qiu, S., Wang, X., Tang, X.: DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
29. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
30. Luo, P., Wang, X., Tang, X.: Pedestrian parsing via deep decompositional network. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2648–2655 (2013)
31. McHenry, K., Ponce, J., Forsyth, D.: Finding glass. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 2, pp. 973–979. IEEE (2005)
32. Mojsilovic, A.: A computational model for color naming and describing color composition of images. Image Processing, IEEE Transactions on 14(5), 690–699 (2005)
33. Roth, P.M., Hirzer, M., Köstinger, M., Beleznai, C., Bischof, H.: Mahalanobis distance learning for person re-identification. In: Person Re-Identification, pp. 247–267. Springer (2014)
34. Schauerte, B., Fink, G.A.: Web-based learning of naturalized color models for human-machine interaction. In: International Conference on Digital Image Computing: Techniques and Applications (2010)
35. Serra, M., Penacchio, O., Benavente, R., Vanrell, M.: Names and shades of color for intrinsic image estimation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 278–285 (2012)
36. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
37. Tan, R.T., Ikeuchi, K.: Separating reflection components of textured surfaces using a single image 27(2), 178–193 (2005)
38. Van De Weijer, J., Schmid, C., Verbeek, J., Larlus, D.: Learning color names for real-world applications. IEEE Transactions on Image Processing 18(7), 1512–1523 (2009)

39. Vazquez, E., Baldrich, R., Van de Weijer, J., Vanrell, M.: Describing reflectances for color segmentation robust to shadows, highlights, and textures. Pattern Analysis and Machine Intelligence, IEEE Transactions on 33(5), 917–930 (2011)
40. Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S.: Graph embedding and extensions: a general framework for dimensionality reduction. Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(1), 40–51 (2007)
41. Yang, Y., Yang, J., Yan, J., Liao, S., Yi, D., Li, S.Z.: Salient color names for person re-identification. In: European Conference on Computer Vision (ECCV) (2014)
42. Yu, Q., Liu, F., Song, Y., Xiang, T., Hospedales, T.M., Loy, C.C.: Sketch me that shoe. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
43. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: International Conference on Computer Vision (ICCV) (2013)
44. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: International Conference on Computer Vision (ICCV) (2015)
45. Zheng, W.S., Gong, S., Xiang, T.: Person re-identification by probabilistic relative distance comparison. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2011)