DeepBag: Recognizing Handbag Models

Yan Wang, *Student Member, IEEE*, Sheng Li, and Alex C. Kot, *Fellow, IEEE*

***Abstract*—In this paper, we address the problem of branded handbag recognition. It is a challenging problem due to the non-rigid deformation, illumination changes, and inter-class similarity. We propose a novel framework based on deep convolutional neural network (CNN). Concretely, we propose a new CNN model, called feature selective joint classification-regression CNN (FSCR-CNN). Its advantages lie in two folds: 1) it alleviates the illumination changes by a feature selection strategy to focus on the color-nondiscriminative features in the network learning, and 2) rather than only targeting on the hard label (i.e., the handbag model), it also incorporates a soft label (i.e., a distribution measuring the similarity between the ground truth model and all the models to be trained) to construct the loss function for training CNN, which leads to a better classifier for handbags with large inter-class similarity. We evaluate the performance of our framework on a newly built branded handbag dataset. The results show that it performs favorably for recognizing handbags with 94.48% in accuracy. We also apply the proposed FSCR-CNN model in recognizing other fine-grained objects with state-of-the-art CNN architectures, which is able to achieve over 5% improvement in accuracy.**

***Index Terms*—Convolutional neural networks, feature selection, handbag recognition, soft label.**

I. Introduction

VARIETIES of branded handbags have become the necessities of human beings in today's fashion world. When consumers are attracted to a handbag, they always ask more information regarding its specific model, pricing, etc. Recognizing the model of the handbag from its image can help those consumers to search for more information or even make a key-word based purchase. In addition, getting consumers' feedback from the social network can help the handbag marketing agencies in branding purpose. While consumers usually prefer to upload photos of their precious handbags without specific model description, it will result in difficulty for marketing agencies to retrieve consumers' comments based on the handbag model. By recognizing the handbags, descriptions or comments attached with the handbag photo can be easily collected. Therefore, to build up a convenient and useful multimedia system to help users retrieve more information about the query handbag, image-based branded handbag recognition is a key step.

The function of a typical multimedia system is to get information from multiple media sources e.g. text, graphics drawings or images and use it for various applications [1], [2]. A lot of researches have been done on multimedia retrieval [3]–[5], fashion search [6], fashion recommendation [7], [8], fashion parsing [9], and label or landmark recognition [10]. As a fashion recognition problem, handbag recognition has a large demand. In this paper, we consider studying this recognition problem from a visual standpoint: recognize the handbag model for a given input handbag image. The model information, recognized from still images, can be potentially employed to other multimedia applications, e.g., to be combined with other media sources, such as numerical number (ratings from the consumer, pricing) and text (comments and preferences), to build a powerful information mining system or an e-commerce recommender system.

Handbag recognition belongs to the category of fine-grained object recognition, which is a challenging problem even for human beings [11]. The main challenge of fine-grained object recognition is to differentiate fine-details among sub-categories of the same object class (e.g., birds, dogs, flowers or handbags). As for handbag recognition, the main difficulties lie in three folds: 1) non-rigid deformation: for those handbags with soft materials, their shapes or patterns might be heavily distorted [see Fig. 1(a)]; 2) illumination changes: some handbags differ with each other only by color (color sensitive), however, the illumination changes enlarge the intra-class color variance [see Fig. 1(b)]; and 3) inter-class similarity: the appearances of some handbags may be very similar [see Fig. 1(c)], which makes it difficult to learn a proper classifier to differentiate these visually similar models.

A growing literature corpus has proposed various techniques for fine-grained object recognition. For example, the deformable part descriptor-based methods [12] deal with large pose variation. Handbags, however, do not have prototypical regions or annotated parts, which brings difficulties when seeking for a solution by using those methods. The detection and segmentation based methods [13] decrease the impact of background, while the segmentation accuracy cannot be guaranteed. Human-interactive methods [14] incorporate human intelligence to assist the recognition. It is a burden for users who likely cannot understand the internal workings of the algorithm. Deep Convolutional Neural Network (CNN) based methods [15] transfer the CNN models trained on large labeled datasets (e.g., ImageNet [16]) to specific visual recognition

Manuscript received March 15, 2015; revised August 27, 2015; accepted September 13, 2015. Date of publication September 18, 2015; date of current version October 20, 2015. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Benoit Huet.

Y. Wang and S. Li are with the Rapid-Rich Object Search (ROSE) Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 637553 (e-mail: wang0696@e.ntu.edu.sg; lisheng@ntu.edu.sg; eackot@ntu.edu.sg).

A. C. Kot is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: eackot@ntu.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2015.2480228

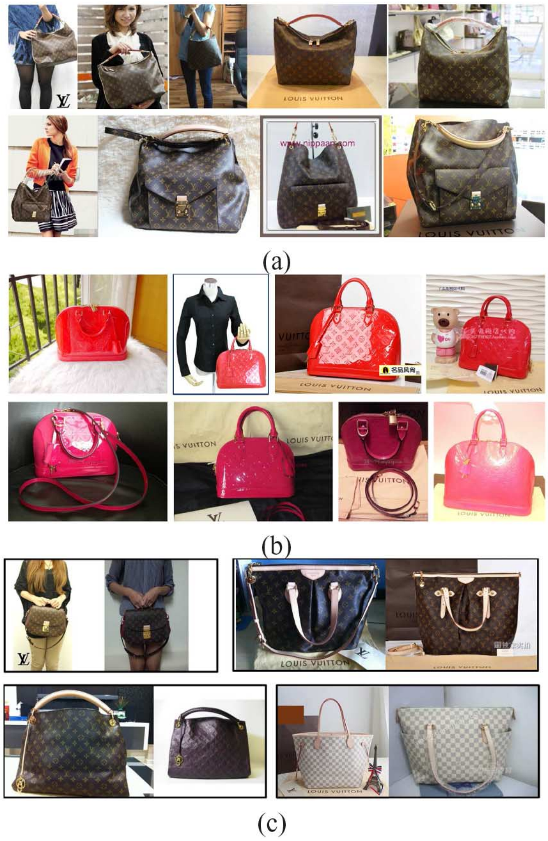


Fig. 1. Illustrations of the main difficulties in handbag recognition due to (a) non-rigid deformation, (b) illumination changes, and (c) inter-class similarity. The models of handbags in each row are the same in (a) and (b), while similar handbags are enclosed in the same box in (c). All the figures in this paper are best viewed in color.

tasks. But none of the methods train networks by considering the challenges in handbag recognition.

The traditional approaches to address this problem usually employ shallow architectures, which consist of hand-crafted features [17]–[19] followed by trainable classifiers [20], [21]. Some approaches have an additional feature selection or data representation procedure [22], [23]. In our previous works [24], [25], we first attempt to address handbag recognition by proposing a hierarchical structure or a complementary feature. However, the lack of discriminability of hand-crafted features becomes a bottleneck for training a better classifier. Recently, CNN has been shown to be effective in both feature extraction and classifier learning [26], [27]. However, previous CNN models do not embed discriminative color information during training, and only use the ground truth class label for recognition tasks. To handle the difficulties in handbag recognition, our focus here is learning more powerful discriminative information based on CNN. We explicitly consider (1) incorporating discriminative color information to classify color sensitive objects, and (2) assigning a distribution measuring how similar this class is to other classes to differentiate visually similar classes.

Concretely, we propose a novel end-to-end framework to recognize the model for a given input handbag image. This framework aims to address the difficulties mentioned above, where we explore CNN for feature and classifier learning. In the training phase, two kinds of CNN models are learned: a CNN detection model (Section III-B) and a Feature Selective joint Classification-Regression CNN (FSCR-CNN) classification model (Section III-C). We propose the following two innovations for FSCR-CNN.

1) Feature Selective CNN (FS-CNN): we learn a regression function to map the first fully-connected feature to the color feature via random forest. Then, by measuring the color-discriminability of each element of the fully-connected feature based on the regression function, only those color-nondiscriminative ones will be forwarded and back-propagated.
2) Joint Classification-Regression CNN (CR-CNN): we propose a novel loss function by considering both the hard label and the soft label of the training data for CNN fine-tuning. For each training sample, the hard label means its ground truth class label, while the soft label refers to a distribution that measures the similarities between its ground truth class and all the classes.

During the testing stage, we first propose to localize and extract a set of handbag regions (proposals) by exploring the symmetry property of the handbag (Section III-A). These extracted proposals are then fed into the CNN detection model and FSCR-CNN classification model separately. The detection scores and classification scores are combined by employing a conditional probability model (Section III-D). Eventually, we recognize the query handbag image according to the highest combined score.

The major contributions of this work can be summarized as follows.

1) We propose a feature selection strategy to improve the discriminability of the learned CNN by optimizing the color-nondiscriminative features in handbag recognition.
2) We propose a novel loss function for CNN by taking both the hard label and the soft label into consideration, so as to facilitate the classifier modeling for visually similar objects. Such a loss function can be adopted on different CNN architectures, with over 7% improvement in accuracy for handbag recognition.
3) The proposed FSCR-CNN can be generalized to other image-based fine-grained object recognition problems.

II. Related Work

In this section, we briefly review the related works, including CNN learning based image classification and fine-grained object recognition.

A. Deep Convolutional Neural Network For Image Classification

Recently, deep learning shows promising results in many computer vision applications such as image classification, image understanding, etc. [28], [29]. CNN architectures [26], [30], [31] achieve state-of-the-art results, surpassing methods incorporating hand-crafted feature representations or traditional classifiers obtained through years of domain-specific expertise. Several CNN structures are proposed to learn discriminative features from raw image inputs and exhibit hierarchical semantic information along their deep architecture [31]–[33].

The popular "AlexNet" ImageNet model [26] is built on an eight-layer architecture, where the first five layers are convolutional layers, followed by three fully connected layers, with an N-way softmax as the output, where N is the number of categories. Its performance is more than 10% better when compared with traditional methods in the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC-2012) competition. An integrated framework OverFeat [34] was proposed later to train a convolutional network which can simultaneously do the task of detection, localization and object classification. This work shows how a multiscale and sliding window approach can be efficiently implemented within a Convolutional Network. OverFeat wins the localization task of ILSVRC-2013. The CNN-S architecture proposed by Chatfield *et al.* [27] is related to the structure of the OverFeat network and achieves better performances in image classification. To address the object recognition problem, Girshick *et al.* [35] proposed regions with CNN features (R-CNN) to localize objects by applying CNN to bottom-up region proposals when labeled training data is scarce. Recently, a new 22 layers deep network GoogLeNet [31] has been proposed to increase the depth and width of the network with the computational budget unchanged. It achieves the state-of-the-art performance for detection and classification in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14).

In this paper, we propose a handbag recognition framework. Our framework consists of a training phase and a testing phase. A CNN detection model and a proposed CNN classification model are trained. In the testing phase, we first incorporate the symmetry property of the handbag for handbag proposals extraction. The extracted top proposals are fed into the detection model and classification model to yield detection scores and classification scores, respectively. A conditional probability model is further employed to combine the detection and classification scores. Our handbag proposals extraction is related to the R-CNN [35] framework for object detection. We also study various CNN architectures for training the classifier. We find that the previous CNN models [26], [34], [27], [31] do not provide discriminative color information during training. Moreover, CNN models only consider the hard label (i.e., the ground truth class label) to train a multi-class classifier. This is not sufficient especially for visually similar classes. In order to train a better CNN for classification, we propose a Feature Selective joint Classification-Regression CNN (FSCR-CNN) model, which consists of Feature Selective CNN (FS-CNN) and joint Classification-Regression CNN (CR-CNN). FS-CNN incorporates a feature selection strategy after the first fully connected layer such that the feature elements not describing color well are forwarded and back-propagated. CR-CNN introduces the usage of the soft label, to complement the hard label for classification.

B. Fine-Grained Datasets and Classification Strategies

Fine-grained object recognition aims at classifying visual data in a subordinate level, e.g., to differentiate blackbird from crow or to tell dandie dinmont from maltese. Several competitive benchmarks have been built for the research of fine-grained object recognition such as the Caltech-UCSD bird [36], the Stanford Dogs [37], the Oxford Flower 102 [38], and Wang's large-scale car datasets [39].

One way to tackle fine-grained recognition is to seek for the localization of the discriminative parts [40]–[42], [12]. The motivation is driven by the observations that some semantic parts have isolated subtle appearance differences among fine-grained subcategories. Thus, keeping the discrimination among visually similar categories facilitates fine-grained categorization. Segmentation/detection with classification methods [13], [39] show that segmenting out the background distracters is beneficial. It helps classification in several ways, such as localizing the object. Human-interactive assistance [43], [14] requires user's input to assist the recognition process. However, these methods are not directly applicable to handbag recognition problem. Specifically speaking, handbags do not have semantic parts; segmentation based recognition methods would not perform well if the foreground region were not segmented correctly; and human-interactive methods require human labor.

Recently, CNN is applied to fine-grained object recognition. Branson *et al.* [12] normalize the pose of bird species and feed each region into a CNN, where features are extracted from multiple layers. Xiao *et al.* [44] combine object-level attention and part-level attention to train domain-specific deep nets. Azizpour *et al.* [15] investigate the transferability of ConvNet representation for a particular target task from aspects such as network width, network depth, and dimension. Our approach also falls into the CNN learning category, while in order to deal with difficulties mentioned in Section I, our work differs in the network configuration of color selection and label distribution learning.

III. Handbag Recognition

Fig. 2 shows our proposed handbag recognition framework. Given an input handbag image, we localize a set of handbag regions by exploring the symmetry property of handbags. Two different deep CNN models are trained, the CNN detection model and the FSCR-CNN classification model, to predict the foreground detection score and classification score of each proposal, respectively. Finally, conditioned on the scores of the CNN detection model, a probability model is employed to refine the classification scores of FSCR-CNN model. The model (class) of the input handbag is predicted based on the refined classification scores.

A. Symmetry-Based Proposals Localization

Object proposal indicates a candidate bounding box covering an object in the image [45]. Using object proposals increases the computational efficiency for object detection. Recent works include objectness cues [46], selective search [47], BING with high computation [48] and edge box [45]. Edge box method [45] computes how likely a bounding box contains an object by calculating the number of contours wholly contained in the box, which is suitable for localizing handbag proposals. The shape of handbags is rectangular-like, which satisfies the assumption of the edge boxes, i.e., their contours are more likely to be wholly enclosed or fitted by a box.

However, edge box method is designed for general objects, which does not consider the specific property of handbags.

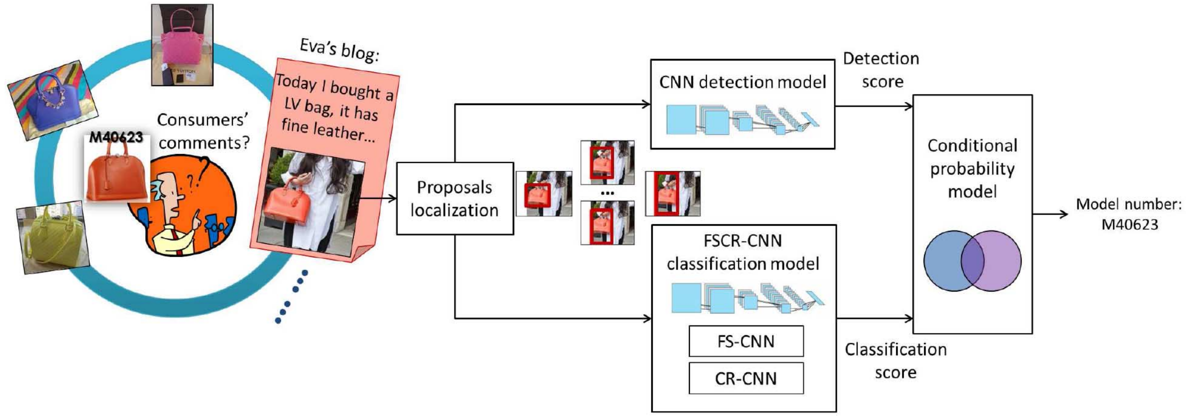


Fig. 2. Overview of the proposed handbag recognition framework. Given an input handbag image, a set of proposals are localized and extracted, which are further fed into the CNN detection model and the FSCR-CNN classification model. Eventually, the conditional probability model recognizes the handbag model by combining the classification scores and the detection scores.

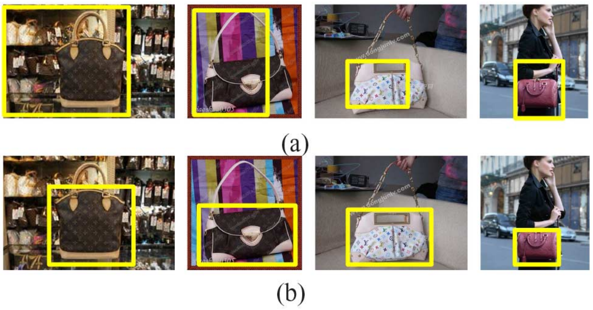


Fig. 3. Top ranked handbag proposals (enclosed in yellow boxes) by (a) the edge box method [45] and (b) the proposed method.

Thus, the returned top proposals by edge box method sometimes cannot accurately enclose a handbag region, which cover parts of the region, or parts of the background instead, as shown in Fig. 3(a). This problem can be alleviated by utilizing the observation that handbags are often shown in symmetry.

We follow the notation in [45]. For the computed edge map of an input image I, any of the pixel is defined as e, represented as a complex number, with magnitude m_e and orientation θ_e. Then some candidate bounding boxes are computed on the edge map based on a sliding window search. Each bounding box b localizes a corresponding object proposal, the score of which is calculated by [45]

$$h_b = h_{1b} - h_{2b} \tag{1}$$

where the first term h_{1b} computes the score of whether a set of edge groups is wholly contained in box b, the second term h_{2b} computes the edge magnitudes from a smaller box centered in b, and the subtraction is because those edges in the center of the box are less important.

Next, we propose to compute a symmetry score for the proposal enclosed by box b, which is to measure how symmetric the proposal is. The feature extraction procedure below for computing symmetry score is shown in Fig. 4:

1) quantize m_e and θ_e into 10 and 6 bins that are uniform in space, respectively;

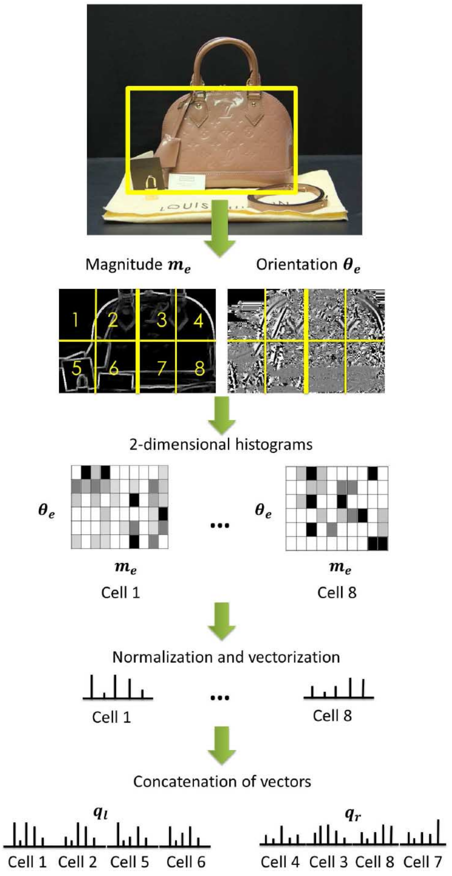


Fig. 4. Feature extraction procedure for computing the symmetry score for a proposal. Each block in the 2-dimensional histogram indicates the frequency of occurrence of edge pixels in a cell with corresponding quantized magnitude and orientation, darker means higher frequency.

2) decompose b into 2×4 spatial cells (4 cells on the right and 4 cells on the left). Each cell is represented by a 6×10-bin 2-dimensional histogram;

3) normalize the 2-dimensional histogram by the number of edge pixels inside each cell, which is further pulled into a vector to represent the cell;

4) use the same sequence to concatenate the vector of each cell on the right, denoted as q_r and left, denoted as q_l, respectively.

The symmetry score s_b of the proposal is computed by

$$s_b = -\chi^2(q_r, q_l) \tag{2}$$

where $\chi^2(q_r, q_l)$ indicates the Chi-square distance between q_r and q_l. The final proposal score of b is

$$h_b^* = h_{1b} - h_{2b} + \gamma s_b \tag{3}$$

where $\gamma > 0$ is to balance the object proposal score and the symmetry score.

The proposals with P highest scores are selected for future recognition. Noted that we exclude the proposals which are too small to provide sufficient information for recognition. Only bounding boxes which satisfy the criteria that $b_w > \eta I_w$ and $b_h > \eta I_h$ will be considered, where $0 < \eta < 1$, b_w and b_h are the width and height of box b, I_w and I_h are the width and height of image I. Some returned top proposals by proposed method are shown in Fig. 3(b).

B. CNN Detection Model

Deep CNN model is shown to be a powerful image descriptor or classifier [26], [49]. However, for fine-grained datasets which only have limited resources, CNN suffers from over-fitting. Therefore, for all CNNs trained in our paper, we adopt the ImageNet pre-trained model, and fine-tune it accordingly. ImageNet [16] organizes objects according to the WordNet [50] hierarchy and each node is depicted by hundreds and thousands of images. It contains the subset of bags categorized by the shoulder bag, evening bag, clutch, reticule and etui.

In the CNN detection model, the deep CNN is trained as a binary classifier to distinguish the foreground handbag region from the background. Data preparation details will be discussed in the experiment.

C. FSCR-CNN Classification Model

Feature Selective CNN Architecture (FS-CNN): In deep CNNs, after the first convolutional layer, RGB channels are all mixed to be fed into the consecutive layers. Such CNN models may not be good at dealing with illumination changes [see Fig. 1(b)]. To address this problem, we introduce a feature selection strategy into CNN to help learn features which can better describe color information. The proposed FS-CNN is shown in Fig. 5. Based on the proposed feature selection, the color-discriminative features of fc_1 remain unchanged and the color-nondiscriminative features participate in the forward pass and the backpropagation.

We choose color-nondiscriminative features to forward and back-propagate because these features are not informative for the color, which may result in an unsatisfactory classification result. Their associated neurons are required to be further optimized based on the corresponding classification error. In such a way, the whole network is more capable to do the classification. To select the color-nondiscriminative features, we propose a random forest [21] based feature selection procedure. Random forest is an ensemble of randomized decision trees, which is shown to perform well for multi-class classifications in many tasks [51]–[53]. Each random forest tree consists of several branch nodes and leaf nodes. Each branch node selects a feature dimension which is discriminative for the classification or regression. According to [21], a certain feature can regress another feature using random forest. Here we adopt such procedure to regress the color feature of each training image using CNN fc_1 feature. In our implementation, the color feature is computed by employing color naming method [54]. It learns color from real-world images which is more robust to the illumination variance. During such regression process, each node of the random forest tree will select the most color discriminative dimension of fc_1 feature. We propose to measure the color discriminability of the ith dimension of fc_1 feature by

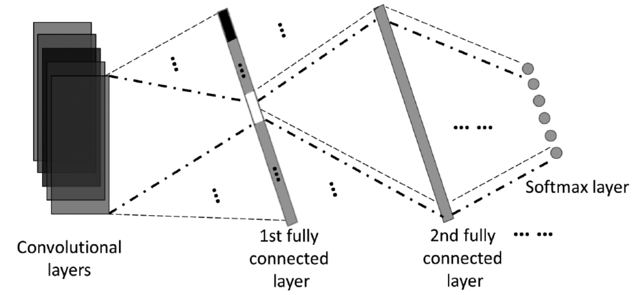


Fig. 5. Illustration of the FS-CNN. The feature selection is applied on the first fully connected layer, where the black part indicates the color-discriminative feature elements, and the white part indicates the color-nondiscriminative feature elements. The dashed line indicates forward pass and dotted dashed line indicates the backpropagation.

$$d(i) = \sum_{k=1}^{Y} \Phi(k, i) \tag{4}$$

where Y is the number of the branch nodes for all trees and

$$\Phi(k, i) = \begin{cases} 1 & \text{if } \vartheta(k) = i \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $\vartheta(k)$ refers to the index of the feature element chosen in the kth node. Among H-dimensional fc_1 feature elements, βH $(0 < \beta < 1)$ most non-discriminative feature elements are selected according to the color discriminability.

Joint Classification-Regression CNN Model (CR-CNN): Handbag recognition is a multi-class classification problem. A straightforward way to address it is to train a hard label multi-class classifier, such as the softmax classifier. However, some handbags are extremely similar, as shown in Fig. 1(c), even human beings have difficulties to distinguish them. In other words, given a single hard label of a handbag class, it's difficult to train a reliable classifier to distinguish it from its visually similar classes. This is because the penalties for misclassifications to its visually similar classes and dissimilar ones are equal. Therefore, a better way is to assign an additional soft label to each handbag class, which is a distribution measuring how similar this class is to all the classes. By using the soft label, the penalties for misclassifying a handbag to its visually similar classes are less than those to the dissimilar classes. Researches have been done to show that soft label is helpful

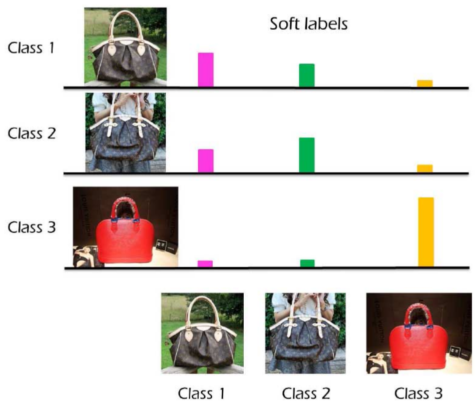


Fig. 6. Examples of soft labels in a three-class dataset.

for some computer vision tasks [55], [56]. Here we propose to take advantages of both hard label and soft label for the CNN training procedure. Fig. 6 illustrates the soft labels of a 3-class dataset, each of which can be represented as a 3-dimensional vector.

To assign an additional soft label to each handbag class, we can adopt a confusion matrix based on the classification performance on a validation set as indicated in [57], [58]. However, the confusion matrix measures how easy it is to discriminate between different classes, which may not have a good measure for the similarities between classes. Therefore, in order to establish the soft labels, we propose to use learned CNN features of the training data to measure the similarities between classes. As these features are learned together with classifiers, they are descriptive and distinctive. The first fully-connected layer (i.e., fc_1) features of the training data are extracted and processed as follows:

1) compute the mean of fc_1 features among all the training samples within the same class (each class has a corresponding mean fc_1 feature);
2) obtain a distance matrix $\mathbf{D}$ with each element as

$$D(i,j) = \chi^2(\bar{\mathbf{v}}_i, \bar{\mathbf{v}}_j) \tag{6}$$

where $\bar{\mathbf{v}}_i$ and $\bar{\mathbf{v}}_j$ indicate the mean fc_1 features of class i and j, respectively, $\chi^2(,)$ refers to the Chi-square distance of the mean features;
3) normalize $D(i,j)$ to $D^*(i,j) \in [0,1]$;
4) compute a matrix $\mathbf{F}$ with entry

$$F(i,j) = \frac{1 - D^*(i,j)}{\sum_j (1 - D^*(i,j))} \tag{7}$$

which measures the similarity between classes i and j;
5) the soft label for class i can therefore be assigned as $\mathbf{s}(i) = (F(y^{(i)},1), \ldots F(y^{(i)},K))$, where K is the number of classes being trained.

Now with the learned soft labels, we propose a joint classification-regression loss to learn the network. Given a handbag training dataset with N foreground handbag images $x^{(i)}$ (belonging to K handbag class) with the labels $y^{(i)} \in \{1,2,\ldots,K\}$, where $i = 1,\ldots,N$, and let $a_j^{(i)}(j = 1,\ldots,K)$ be the output of the last inner-product layer for $x^{(i)}$, the proposed joint loss function is defined as

$$J = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\Bigg(\mathbf{1}(y^{(i)} = j)\log p_j^{(i)} - \lambda\Big(\log F(y^{(i)},j) - \log p_j^{(i)}\Big)^2\Bigg) \tag{8}$$

where

$$p_j^{(i)} = \frac{\exp a_j^{(i)}}{\sum_{l=1}^{K}\exp a_l^{(i)}}. \tag{9}$$

$\mathbf{1}(\cdot)$ is the indicator function, and λ is a tradeoff parameter which balances the two loss terms. The first term is the standard softmax loss which penalizes the classification error for each class equally. The second term is the regression squared loss term, which penalizes the difference between the predicted scores of $x^{(i)}$ (i.e., $p_j^{(i)}$) and the soft labels $\mathbf{s}(y^{(i)})$. The second term is learned jointly with the first term, which also acts as a regularizer for the first term.

In order to run stochastic gradient descend (SGD) on the proposed loss function, we apply the back-propagation based on the partial derivatives of the new loss with respect to the output of the last inner product layer $a_j^{(i)}$. The partial derivatives are given as follows:

$$\frac{\partial J}{\partial a_j^{(i)}} = -\frac{1}{N}\Bigg[1(y^{(i)} = j) - p_j^{(i)} - 2\lambda\Bigg(\sum_{l=1}^{K}(\log F(y^{(i)},l) - \log p_l^{(i)})p_j^{(i)} - \log F(y^{(i)},j) + \log p_j^{(i)}\Bigg)\Bigg]. \tag{10}$$

D. Conditional Probability Model

Given a test image, its P top-ranked proposals are fed into both CNN detection model and FSCR-CNN classification model. For each of the top P proposals $r_i, i = 1,\ldots,P$, we compute the probability that r_i belongs to class j ($j = 1,\ldots,K$) as

$$c(j, f_i|r_i) = c(j|f_i, r_i)c(f_i|r_i) \tag{11}$$

where $c(j|f_i, r_i)$ indicates the classification score of the FSCR-CNN model for r_i with the assumption that it belongs to the foreground region, and $c(f_i|r_i)$ denotes the foreground detection score of the CNN detection model for r_i. The query image I is then predicted as class j^*, where

$$j^* = \arg\max_{\substack{j=1,\ldots,K,\\ i=1,\ldots,P}} (c(j, f_i|r_i)). \tag{12}$$

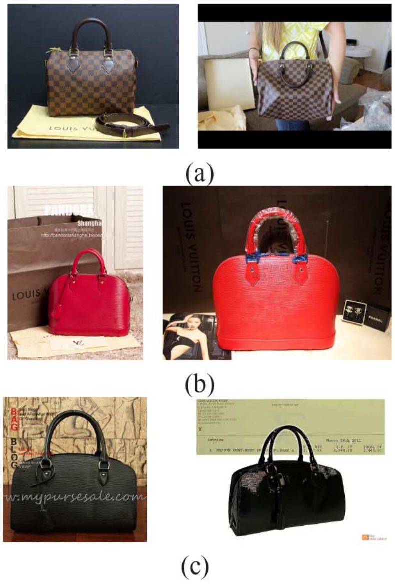


Fig. 7. Examples of visually indistinguishable handbags. Handbags with the same appearance but with (a) different sizes, (b) indistinguishable colors, and (c) different materials.

IV. Dataset Construction

As no existing benchmark is available for branded handbag recognition, we construct a handbag dataset, covering 220 Louis Vuitton handbag models downloaded from Google, Flickr or some shopping websites. Building such a dataset costs a lot of human labor due to the following reasons: (1) the image resources are limited for most of the handbag models and (2) slight texture or color changes of handbags will lead to different handbag models. Handbag images of the same model might appear differently due to serious distortions or variations of illumination.

The dataset construction procedure consists of four key steps: 1) list the target handbag models to collect; 2) merge handbag models which are visually undistinguishable, such as handbags with the same appearance but different sizes, indistinguishable colors, or different materials (as shown in Fig. 7); 3) for each handbag model, search for handbag images from public websites (Google, Flickr or online shops) by key-words such as the model name, which costs human labor because many attached annotations do not match with the images; and 4) remove handbag images which are noisy, duplicated, heavily occluded, with low quality or in wrong viewpoints, and retain handbag models containing at least 10 images. Eventually, the dataset contains 5545 images of 220 handbag models. Each handbag image in our dataset is annotated manually with a bounding box, which is a rectangular region outside the handbag surface without strap, as shown in Fig. 8. To the best of our knowledge, this is the first dataset created for handbag recognition.

V. Experiments and Discussions

A. Experimental Setup

Data Preparation: We randomly split our dataset into 5 images per model for training and the rest for testing. Each image in our dataset has a ground truth bounding box. We use the cropped images as the input to train the framework, which is guided by the bounding boxes.

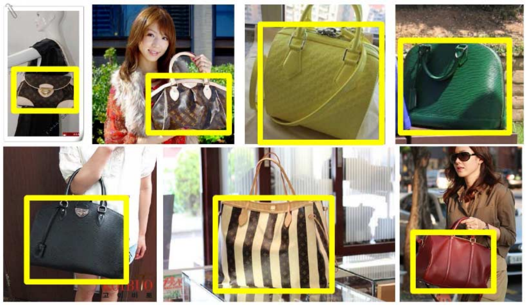


Fig. 8. Examples of handbag images with the associated bounding boxes (marked with yellow rectangles) in our dataset.

For training FSCR-CNN classification model, we augment the input data in two ways:

1) sample random crops around the bounding box regions from both original images and their flipped versions. The cropping is done such that $T > 0.8$, where T is defined as the ratio between the intersection and the union of the crop and the bounding box;
2) apply our proposal localization method on the training images. For each training image, we select the first 20 proposals (i.e., crops) which satisfy $T > 0.8$.

Eventually we generate 17,774 images for training 220 handbag models.

For CNN detection model, we regard the previous generated crops as positive data for training. To create the negative data, we randomly crop image patches from the background of the training images, which also contains two sets, including the crops with $T < 0.6$ and the first 20 proposals (extracted by the proposed proposal localization method) for each image with $T < 0.6$. Thus, we obtain 17,774 positive and 18,102 negative training data.

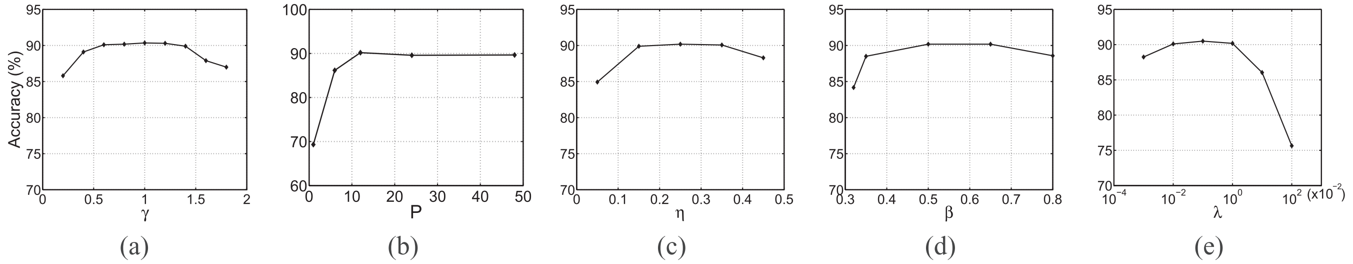
CNN Model: For all the CNN models we train, we start the training with a fixed learning rate and decrease it by a factor of 10 after the training error stops reducing. In our implementation, we use the MatConvNet toolbox [59], which provides different CNNs for computer vision applications. The CNN model proposed in [26] is incorporated in our proposed handbag recognition framework.

B. Evaluation of the Proposed Framework

In this section, we evaluate the performance of the proposed handbag recognition framework. Several parameters are needed to be set in advance. Similar to the parameter tuning in [60], [61], we tune our parameters with the help of cross-validation on the training data. Based on our observation, the top 10 proposals are sufficient to cover the handbag region. With this observation, we set the number of selected proposals $P = 10$ as the initial value. Percentage of non-discriminative feature elements β is within the range of [0,1]. To disable feature selection during classifier training first, we set $\beta = 1$ as the initial value. Following [60] and [61], we initialize γ, η and λ to be zero. We then sequentially learn one after another by applying cross validation (e.g., search for the best value of γ first, and then fix γ,

TABLE I
COMPARISONS OF HANDBAG RECOGNITION ACCURACIES ON THE HANDBAG DATASET

Method	Accuracy (%)
EdgeBox + CNN classification (baseline)	65.63
Symmetry-based EdgeBox + CNN classification	73.83
Symmetry-based EdgeBox + CNN detection + CNN classification	81.96
Symmetry-based EdgeBox + CNN detection + FS-CNN classification	85.02
Symmetry-based EdgeBox + CNN detection + CR-CNN classification	89.32
Symmetry-based EdgeBox + CNN detection + FSCR-CNN classification	**90.18**



(a) (b) (c) (d) (e)

Fig. 9. Handbag recognition accuracies of symmetry-based EdgeBox +CNN detection +FSCR − CNN classification when using different parameters: (a) tradeoff between the object proposal score and the symmetry score γ, (b) number of selected proposals P, (c) scale ratio for the selected proposals η, (d) percentage of non-discriminative feature elements β, and (e) tradeoff between classification loss and regression loss λ.

and search for the best value of P). Unless otherwise specified, we use the default setting for these parameters.

In our framework, we use the foreground images to do the training. Thus, similar to the work in [35], we regard the object proposals +CNN classification as our baseline method, but we adopt a more recent edge box method [45] for extracting the object proposal. For each image, we extract the top P proposals and take the classification result of the object proposal which gives the maximum classification response. The performance of the baseline is given in Table I, which is over 5% higher compared with using CNN only (59.48%).

Performance of the Symmetry-Based Proposal Localization: We evaluate our symmetry-based edge box method, which provides better proposals compared with the baseline. It is shown from the third row in Table I that we are able to obtain an improvement of over 8% in accuracy compared with the baseline.

Performance of the CNN Detection Model and Conditional Probability Model: We recognize handbags with the CNN classification scores of their proposals conditioned on the CNN detection scores. The result of such handbag recognition is shown as Symmetry-based EdgeBox +CNN detection +CNN classification in Table I. The combination of detection scores with classification scores by the conditional probability model provides better results with 8% improvement over Symmetry-based EdgeBox +CNN classification.

Performance of the Proposed FSCR-CNN Classification Model: We have made two contributions in this model, which are FS-CNN and CR-CNN. We report the handbag recognition accuracies after replacing existing CNN classification with our proposed classification models in Table I. Replacing CNN classification with FS-CNN classification and CR-CNN classification in our framework both lead to better results, with around 4% and 8% improvement in the accuracy respectively. We further combine FS-CNN and CR-CNN (i.e., FSCR-CNN) for the classification. With FSCR-CNN, our framework achieves 24.55% better than the baseline (EdgeBox +CNN classification).

We also evaluate the performance of applying only CNN and FSCR-CNN directly for handbag recognition (without foreground detection). As mentioned earlier in this section, if we only used CNN, the recognition accuracy would be 59.48%. FS-CNN and CR-CNN help to boost the performance to 62.66% and 69.10% respectively. FSCR-CNN achieves the best accuracy, which is 71.98%. Therefore, it is beneficial to embed CNN models into our proposed framework for handbag recognition.

Parameter Analysis: We vary the parameters γ, P, η, β and λ in the following ranges while keeping the others as the default values.

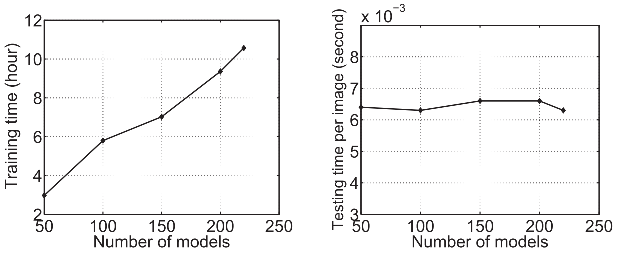
- $\gamma \in \{0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8\}$;
- $P \in \{1, 6, 12, 24, 48\}$;
- $\eta \in \{0.05, 0.15, 0.25, 0.35, 0.45\}$;
- $\beta \in \{0.32, 0.35, 0.5, 0.65, 0.8\}$;
- $\lambda \in \{0.001, 0.01, 0.1, 1, 10, 100\}$.

We evaluate the sensitivity of each parameter for our proposed framework (i.e., Symmetry-based EdgeBox +CNN detection +FSCR − CNN classification) in Fig. 9. We observe that the performance is not sensitive to these parameters within certain ranges. λ is the most significant parameter in our method, as it balances the classification loss and regression loss. Accordingly, when λ is large (i.e., $\lambda > 1$), the soft label plays a more important role than the hard label for classification. When λ is small (i.e., $\lambda < 10^{-5}$), the effect of regression loss can be almost neglected.

Computational Complexity: Our framework consists of three parts: Symmetry-based EdgeBox, CNN detection and FSCR-CNN classification, where Symmetry-based EdgeBox has only testing phase, while the other two have both training and testing phases. We report the training time for 220 handbag classes (in hour) and testing time per image (in second) in

TABLE II
TRAINING TIME (IN HOUR) OF HANDBAG DATASET AND TESTING TIME (IN SECOND) PER HANDBAG IMAGE OF DIFFERENT METHODS

Method	Training time (h)	Testing time (s/image)
EdgeBox	–	0.0465
Symmetry-based EdgeBox	–	0.0565
CNN detection	6.2384	0.0066
CNN classification	0.8390	0.0064
FS-CNN classification	6.7794	0.0061
CR-CNN classification	1.0786	0.0063
FSCR-CNN classification	10.5676	0.0063



(a) (b)

Fig. 10. Training and testing time based on different number of handbag models for FSCR-CNN. (a) Averall training time (hours). (b) Testing time for each image (seconds).

Table II. The experiment is conducted on MATLAB R2013a, in a workstation of E5-2630 CPU, 96 GB RAM, and a GPU Tesla K40. Note that for all CNN methods, we exclude the image loading time and set the batch size to 200 for each epoch. We also evaluate the complexity of FS-CNN and CR-CNN separately, and compare them with CNN. We observe that CR-CNN takes longer training time because it requires several more epochs to converge (25 to 30 for training handbags) than CNN (normally around 25 epochs). FS-CNN converges around 25 epochs. Besides, it takes longer time than CNN or CR-CNN for each epoch since the random forest implementation is time consuming. The training time of FSCR-CNN is longer, because it converges around 35 to 40 epochs. However, the training process can be applied off-line. To speed up, feature selection by random forest can be parallelized. In addition, with more GPUs, those CNNs can be also trained in parallel. For testing, the time costs of all CNNs are about the same.

In addition, we report the scalability of FSCR-CNN in terms of computational complexity and recognition accuracy. Fig. 10(a) shows the increase in the total training time vs. the increase in the number of handbag models. Fig. 10(b) plots the testing time for each image. We observe that with the increase in the number of training classes, the training time is somewhat linear. Noted that this is due to the increasing number of batches required at each epoch. The class number does not heavily influence the testing time per image. Based on the different number of handbag classes, Fig. 11 illustrates the handbag recognition accuracies of the proposed framework. With the increase of class number, the accuracy decreases slightly in a certain range. While the performance does not suffer from a significant drop when the number of classes grows.

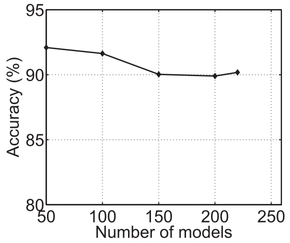


Fig. 11. Handbag recognition accuracies of the proposed framework based on different number of handbag models.

TABLE III
TOP-1 AND TOP-5 ACCURACY (%) OF CNN BASED ARCHITECTURES ON THE OXFORD FLOWERS DATASET [38], THE STANFORD DOGS DATASET [37], AND THE UCSD-BIRDS DATASET [36]

Method	Oxford Flowers		Stanford Dogs		UCSD-Birds	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CNN	77.04	91.95	52.41	81.12	59.94	83.34
FS-CNN	79.69	93.71	51.40	80.73	64.58	87.04
CR-CNN	81.83	93.53	**64.79**	**88.73**	67.38	88.07
FSCR-CNN	**82.21**	**94.29**	–	–	**69.78**	**89.54**

C. Evaluation of the Generality of Proposed FSCR-CNN Model

To verify the generality and superiority of our proposed FS-CNN and CR-CNN model over the CNN model, we also apply them on other fine-grained datasets: the Oxford Flowers [38], the Stanford Dogs [37] and the Caltech-UCSD Birds [36].

Oxford Flowers dataset consists of 102 different flower categories covering 40 to 258 images per category. Follow the data augmentation method provided by [49], we build 16 representatives for each image without segmentation (original image, 5 crops, 2 rotation and their mirrors). The top-1 and top-5 accuracies reported for CNN, FS-CNN, CR-CNN and FSCR-CNN are listed in the second and third column of Table III. The existence of background with green grasses affects the color of flowers. Nevertheless, color component is still of a certain importance for classifying flowers. FS-CNN and CR-CNN both perform better than CNN, and the FSCR-CNN achieves the best.

Stanford Dogs dataset contains over 20,000 annotated images of 120 breeds of dogs. We apply the bounding box annotations for both training and testing procedure as indicated in [11]. For all CNN networks, we randomly crop around the bounding box regions and keep the crops with $T > 0.8$. Eventually, the data augmentation is done by making an average of 9 representations for each training image. The fourth and fifth column of Table III shows the comparisons of the top-1 and top-5 accuracies for different CNN models. The proposed CR-CNN outperforms CNN with at least 12% improvement, which even surpasses the previously published results (see Table IV). However, FS-CNN is not helpful in improving the performance. This is due to the reason that the color is not a sensitive feature for dogs. In this dataset, some dogs even wear clothes or heavily occluded. Therefore, in the following experiments for dogs, we will not evaluate the FS-CNN and FSCR-CNN models unless otherwise specified.

TABLE IV
COMPARISONS WITH OTHER LEADING FINE-GRAINED OBJECT RECOGNITION APPROACHES ON THE STANFORD DOGS DATASET [37]

Method	Accuracy (%)	Mean accuracy (%)
Yang et al. [11]	38.00	–
Pu et al. [62]	39.30	–
Chai et al. [63]	–	45.60
Gavves et al. [40]	–	50.10
CNN	52.41	51.08
CR-CNN	**64.79**	**63.23**

TABLE V
TOP-1 AND TOP-5 ACCURACY (%) OF CNN-S BASED ARCHITECTURES ON THE OXFORD FLOWERS DATASET [38], THE STANFORD DOGS DATASET [37], AND THE UCSD-BIRDS DATASET [36]

Method	Oxford Flowers		Stanford Dogs		UCSD-Birds	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CNN-S	81.09	94.02	71.27	93.62	65.79	88.25
FS-CNN-S	82.05	94.49	–	–	67.10	88.82
CR-CNN-S	85.15	95.10	**76.64**	**94.92**	71.73	89.63
FSCR-CNN-S	**86.21**	**95.38**	–	–	**73.90**	**91.23**

TABLE VI
COMPARISONS OF CNN-G AND FSCR-CNN-G ON THE OXFORD FLOWERS DATASET [38], THE STANFORD DOGS DATASET [37], AND THE UCSD-BIRDS DATASET [36]

Method	Oxford Flowers		Stanford Dogs		UCSD-Birds	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CNN-G	85.46	95.77	73.91	94.26	75.73	92.66
FSCR-CNN-G	**90.43**	**97.06**	**78.91**	**94.98**	**81.74**	**94.93**

TABLE VII
COMPARISONS OF HANDBAG RECOGNITION ACCURACIES ON DIFFERENT FRAMEWORKS OF CNN-S AND CNN-G-BASED ARCHITECTURE

Framework	Accuracy (%)
EdgeBox + CNN-S classification	78.92
Symmetry-based EdgeBox + CNN-S detection + FSCR-CNN-S classification	92.79
EdgeBox + CNN-G classification	83.40
Symmetry-based EdgeBox + CNN-G detection + FSCR-CNN-G classification	**94.48**

Caltech-UCSD Birds-200-2011 dataset contains 11,788 annotated images of 200 bird species. A total of 5994 images are used for training and the rest 5794 images are used for evaluation. Like the Oxford Flowers dataset, we follow the data augmentation method provided by [49]. Results are shown in the last two columns of Table III. Like flowers, color is an important component for classifying birds. Therefore, FS-CNN is also helpful for bird recognition. Compared with CNN, FSCR-CNN improves the recognition accuracy significantly by 10% for Top-1 accuracy.

Nowadays different CNN architectures have been designed [32], [27], [31]. We also embed our proposed Feature Selection or joint Classification-Regression on a well performed architecture CNN-S [27]. Similarly, the corresponding networks are denoted as FS-CNN-S, CR-CNN-S and FSCR-CNN-S. CNN-S is similar to the OverFeat structure [34], while unlike OverFeat network, less filters are applied in the 5th convolutional layer and a local response normalization layer is added after the 1st convolutional layer rather than contrast normalization. In order to update our proposed FSCR-CNN on recent networks, we also adopt the GoogLeNet [31] (the latest released pre-trained model in MatConvNet toolbox [59]). For GoogLeNet, we denote the original network as CNN-G, and the proposed corresponding network is FSCR-CNN-G.

Table V shows the performance comparisons among FS-CNN-S, CR-CNN-S, FSCR-CNN-S and CNN-S on the Oxford Flowers dataset and the UCSD-Birds dataset, as well as the comparison between CR-CNN-S and CNN-S on the Stanford Dogs dataset. Again, the CR-CNN-S model performs better on these fine-grained object datasets, with over 4% increase in accuracy compared with CNN-S. FS-CNN-S and FSCR-CNN-S are able to achieve better performances on the flower and bird dataset. Compared with CNN-S, FSCR-CNN-S achieves over 5% improvement in accuracy. The comparisons of CNN-G with FSCR-CNN-G for the three datasets are summarized in Table VI. We observe that the improvements in classification accuracy are not limited to CNN structures, as the accuracies are further boosted by over 5%.

CNN-S and CNN-G could also be incorporated into our proposed framework (to replace CNN) for handbag recognition. Table VII compares our proposed framework (i.e., Symmetry-based EdgeBox + CNN-S (or CNN-G) detection + FSCR-CNN-S (or FSCR-CNN-G) classification) with the baseline structure (i.e., EdgeBox + CNN-S (or CNN-G) classification). It can be seen that our proposed framework can boost the handbag recognition performance by more than 10%.

VI. CONCLUSION

In this paper, we design a novel framework to recognize handbag models. In this framework, we propose to incorporate the symmetry property of the handbag for extracting handbag proposals. Then each proposal is fed into a CNN detection model and a proposed FSCR-CNN classification model. FSCR-CNN model attenuates the illumination changes and inter-class similarity among handbags. Proposal detection scores and classification scores are eventually combined by a conditional probability model to further improve the performance of handbag recognition. Extensive experiments on our newly constructed handbag dataset verify the advantages of each component of our framework and shows that it achieves 94.48% in accuracy for recognizing handbags. In addition, FS-CNN is shown to be helpful at recognizing color sensitive fine-grained objects (3% improvement on the handbag dataset, 2% improvement on the Oxford Flowers dataset and 5% improvement on the UCSD-Birds dataset) and CR-CNN performs fairly well on fine-grained object recognition tasks, with 8% improvement for the handbag dataset, 4% improvement for the Oxford Flowers dataset, 12% improvement for the Stanford Dogs dataset and 7% improvement for the UCSD-Birds dataset.

ACKNOWLEDGMENT

This research was carried out at the Rapid-Rich Object Search (ROSE) Laboratory, Nanyang Technological University, Singapore. The ROSE Laboratory is supported by the National

Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme. The authors gratefully acknowledge the support of the NVIDIA Corporation for their donation of a Tesla K40 GPU used for research at the ROSE Laboratory.

References

[1] B. Safadi, M. Sahuguet, and B. Huet, "When textual and visual information join forces for multimedia retrieval," in *Proc. Int. Conf. Multimedia Retrieval*, 2014, pp. 265–272.

[2] W. Yin, T. Mei, C. W. Chen, and S. Li, "Socialized mobile photography: Learning to photograph with social context via mobile devices," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 184–200, Jan. 2014.

[3] Y.-H. Kuo, W.-H. Cheng, H.-T. Lin, and W. Hsu, "Unsupervised semantic feature discovery for image object retrieval and tag refinement," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 1079–1090, Aug. 2012.

[4] Q. You, J. Yuan, J. Wang, P. Guo, and J. Luo, "Snap n' shop: Visual search-based mobile shopping made a breeze by machine and crowd intelligence," in *IEEE Int. Conf. Semantic Comput.*, Feb. 2015, pp. 173–180.

[5] R. Ji, L.-Y. Duan, J. Chen, L. Xie, H. Yao, and W. Gao, "Learning to distribute vocabulary indexing for scalable visual search," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 153–166, Jan. 2013.

[6] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 3330–3337.

[7] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan, "Hi, magic closet, tell me what to wear!," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 619–628.

[8] L. Liu, J. Xing, S. Liu, H. Xu, X. Zhou, and S. Yan, "Wow! You are so beautiful today!," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 11, no. 1s, pp. 20:1–20:22, 2014.

[9] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan, "Fashion parsing with weak color-category labels," *IEEE Trans. Multimedia*, vol. 16, no. 1, pp. 253–265, Jan. 2014.

[10] T. Chen, K.-H. Yap, and D. Zhang, "Discriminative soft bag-of-visual phrase for mobile landmark recognition," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 612–622, Apr. 2014.

[11] S. Yang, L. Bo, J. Wang, and L. G. Shapiro, "Unsupervised template learning for fine-grained object recognition," in *Proc. Adv. Neural Inf. Process. Syst. 25*, 2012, pp. 3131–3139.

[12] S. Branson, G. V. Horn, P. Perona, and S. J. Belongie, "Improved bird species recognition using pose normalized deep convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2014.

[13] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 811–818.

[14] J. Deng, J. Krause, and L. Fei-Fei, "Fine-grained crowdsourcing for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 580–587.

[15] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, to be published.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2009, pp. 248–255.

[17] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[18] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.

[19] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[21] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[22] Z. Li, J. Liu, J. Tang, and H. Lu, "Robust structured subspace learning for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2085–2098, Oct. 1, 2015.

[23] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2138–2150, Sep. 2014.

[24] Y. Wang, S. Li, and A. Kot, "Category-separating strategy for branded handbag recognition," in *Proc. 6th Int. Symp. Commun., Control Signal Process.*, 2014, pp. 61–64.

[25] Y. Wang, S. Li, and A. C. Kot, "Complementary feature extraction for branded handbag recognition," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 5896–5900.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. 25*, 2012, pp. 1097–1105.

[27] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014.

[28] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.

[29] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, to be published.

[30] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *CoRR*, 2013 [Online]. Available: http://arxiv.org/abs/1311.2901

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, to be published.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2014, vol. 8691, pp. 346–361.

[33] C. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. 18th Int. Conf. Artificial Intell. Statist.*, 2015.

[34] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learn. Representations*, Apr. 2014.

[35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 580–587.

[36] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The Caltech-UCSD Birds-200-2011 dataset," California Inst. of Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.

[37] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei, "Novel dataset for fine-grained image categorization," in *Proc. 1st Workshop Fine-Grained Vis.Categorization, IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011.

[38] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.

[39] X. Wang, T. Yang, G. Chen, and Y. Lin, "Object-centric sampling for fine-grained image classification," *CoRR*, 2014 [Online]. Available: http://arxiv.org/abs/1412.3161

[40] E. Gavves, B. Fernando, C. G. M. Snoek, A. W. M. Smeulders, and T. Tuytelaars, "Fine-grained categorization by alignments," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1713–1720.

[41] J. Krause, T. Gebru, J. Deng, L. Li, and L. Fei-Fei, "Learning features and parts for fine-grained recognition," in *Proc. Int. Conf. Pattern Recog.*, Aug. 2014, pp. 26–33.

[42] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 834–849.

[43] A. Rejeb Sfar, N. Boujemaa, and D. Geman, "Confidence sets for fine-grained categorization and plant species identification," *Int. J. Comput. Vis.*, vol. 111, no. 3, pp. 255–275, 2014.

[44] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, to be published.

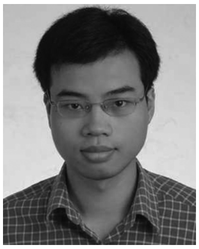
[45] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2014, pp. 391–405.

[46] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.

[47] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.
[48] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300 fps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 3286–3293.
[49] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2014, pp. 512–519.
[50] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Wordnet: An on-line lexical database," *Int. J. Lexicography*, vol. 3, pp. 235–244, 1990.
[51] B. Yao, A. Khosla, and L. Fei-Fei, "Combining randomization and discrimination for fine-grained image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 1577–1584.
[52] W. Shen, K. Deng, X. Bai, T. Leyvand, B. Guo, and Z. Tu, "Exemplar-based human action pose correction," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1053–1066, Jul. 2014.
[53] J. Uijlings, A. Smeulders, and R. Scha, "Real-time visual concept classification," *IEEE Trans. Multimedia*, vol. 12, no. 7, pp. 665–681, Nov. 2010.
[54] J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Trans. Image. Process.*, vol. 18, no. 7, pp. 1512–1523, Jul. 2009.
[55] X. Geng and R. Ji, "Label distribution learning," in *Proc. IEEE 13th Int. Conf. Data Mining Workshops*, Dec. 2013, pp. 377–383.
[56] A. Sironi, V. Lepetit, and P. Fua, "Multiscale centerline detection by learning a scale-space distance transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2697–2704.
[57] S. Godbole, S. Sarawagi, and S. Chakrabarti, "Scaling multi-class support vector machines using inter-class confusion," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2002, pp. 513–518.
[58] Z. Yan, V. Jagadeesh, D. DeCoste, W. Di, and R. Piramuthu, "HD-CNN: Hierarchical deep convolutional neural network for image classification," *CoRR*, 2014 [Online]. Available: http://arxiv.org/abs/1410.0736
[59] A. Vedaldi and K. Lenc, "MatConvNet–convolutional neural networks for MATLAB," *CoRR*, 2014 [Online]. Available: http://arxiv.org/abs/1412.4564
[60] Z. Zuo, G. Wang, B. Shuai, L. Zhao, Q. Yang, and X. Jiang, "Learning discriminative and shareable features for scene classification," in *Proc. ECCV*, 2014, vol. 8689, pp. 552–568.
[61] Z. Zuo, G. Wang, B. Shuai, L. Zhao, and Q. Yang, "Exemplar based deep discriminative and shareable feature learning for scene image classification," *Pattern Recog.*, vol. 48, pp. 3004–3015, 2015.
[62] J. Pu, Y.-G. Jiang, J. Wang, and X. Xue, "Which looks like which: Exploring inter-class relationships in fine-grained visual categorization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 425–440.
[63] Y. Chai, V. Lempitsky, and A. Zisserman, "Symbiotic segmentation and part localization for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 321–328.

Yan Wang (S'13) received the B.S. degree in electronics and information engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2011, and is currently working toward the Ph.D. degree in electrical and electronic engineering at Nanyang Technological University, Singapore.

She was an Exchange Student with the Tokyo Institute of Technology International Research Opportunities Program (TiROP), Tokyo, Japan, in 2012. Her research interests include computer vision, object recognition, and fashion search.



Sheng Li received the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2013.

He is currently a Research Fellow with the Rapid Rich Object Search (ROSE) Laboratory, Nanyang Technological University, Singapore. His research interests include biometric template protection, pattern recognition, multimedia forensics, and security.

Dr. Li was the recipient of the IEEE WIFS Best Student Paper Silver Award.

Alex C. Kot (S'85–M'89–SM'98–F'06) has been with the Nanyang Technological University, Singapore, since 1991. He was Head of the Division of Information Engineering, School of Electrical and Electronic Engineering for eight years, and also served as the Associate Chair/Research and Vice Dean Research for the School of Electrical and Electronic Engineering. He is currently Professor and Associate Dean for the College of Engineering and Director of the Rapid-Rich Object Search (ROSE) Laboratory. His research interests include signal processing for communication, biometrics, data-hiding, image forensics, information security, and image object retrieval and recognition.

Dr. Kot is a Fellow of IES and a Fellow of the Academy of Engineering, Singapore. He is the IEEE Distinguished Lecturer for the Signal Processing Society. He served as Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE SIGNAL PROCESSING LETTERS, *IEEE Signal Processing Magazine*, IEEE JOURNAL OF SPECIAL TOPICS IN SIGNAL PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: FUNDAMENTAL THEORY AND APPLICATIONS, and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—II: ANALOG AND DIGITAL SIGNAL PROCESSING. He has served the IEEE Signal Processing Society in various capacities, such as the General Co-Chair for the 2004 IEEE International Conference on Image Processing and the Vice President for the IEEE Signal Processing Society. He was the recipient of the Best Teacher of the Year Award and coauthored several Best Paper Awards, including for ICPR, IEEE WIFS, and IWDW.