






Melanoma Recognition in Dermoscopy Images via Aggregated Deep Convolutional Features

Zhen Yu , Xudong Jiang , Senior Member, IEEE, Feng Zhou , Jing Qin , Member, IEEE, Dong Ni, Member, IEEE, Siping Chen, Baiying Lei , Senior Member, IEEE, and Tianfu Wang

I. INTRODUCTION

Abstract—In this paper, we present a novel framework for dermoscopy image recognition via both a deep learning method and a local descriptor encoding strategy. Specifically, deep representations of a rescaled dermoscopy image are first extracted via a very deep residual neural network pretrained on a large natural image dataset. Then these local deep descriptors are aggregated by orderless visual statistic features based on Fisher vector (FV) encoding to build a global image representation. Finally, the FV encoded representations are used to classify melanoma images using a support vector machine with a Chi-squared kernel. Our proposed method is capable of generating more discriminative features to deal with large variations within melanoma classes, as well as small variations between melanoma and nonmelanoma classes with limited training data. Extensive experiments are performed to demonstrate the effectiveness of our proposed method. Comparisons with state-of-the-art methods show the superiority of our method using the publicly available ISBI 2016 Skin lesion challenge dataset.

Index Terms—Dermoscopy image, melanoma recognition, residual network, fisher vector, deep learning.

MELANOMA skin cancer is one of the most rapidly increasing and deadliest cancers in the world, which accounts for 75% of skin cancer deaths [1]–[3]. Early diagnosis is of great importance for treating this disease as it can be cured easily at early stages [1]–[4]. To improve the diagnosis of this disease, dermoscopy has been introduced to assist dermatologists in clinical examination since it is a non-invasive skin imaging technique that provides clinicians high-quality visual perception of skin lesion. Compared with the conventional macroscopic (clinical) images, fewer surface reflection, more sufficient deep layers' details, and lower screening errors make dermoscopy images achieve much better visibility and recognition accuracy [3], [5]. Since melanoma is more deadly than non-melanoma skin cancer, discrimination between cancer and non-cancerous melanoma dermoscopy images has attracted considerable interest [1], [2], [4]. Clinically, several heuristic approaches, such as “ABCD” rule [6], Menzies method [7] and “CASH” [8], have been developed to enhance clinicians' ability to distinguish melanomas from benign nevi. However, the correct diagnosis of a skin lesion is not trivial even for experienced professionals. Furthermore, dermoscopic diagnosis made by human visual inspection is often laborious, time-consuming and subjective. Hence, unsatisfactory accuracy and poor reproducibility are still issues for diagnosing this disease.

To tackle these issues, numerous algorithms were proposed for automatic dermoscopic image analysis. Interested readers can refer to [3], [9], [10] for a comprehensive summary of related work over the past decades. By and large, the pipeline of these computer-aided analysis models usually includes the following four steps: i) image preprocessing such as hair removal [11]–[13] and image enhancement [14], [15]; ii) border detection or segmentation [2], [16]–[18]; iii) feature extraction (i.e., color, texture, border gradient, shape related descriptors) [2], [18]–[20]; iv) classification (k-nearest neighbor (KNN) [18], support vector machine (SVM) [2], AdaBoost [20], neural network [19], etc.). Most of the existing studies have mainly focused on feature engineering and classification, either implicitly or explicitly, assuming that the input image contains a lesion object in well-condition [19]. However, dermoscopy images may not always capture entire lesions, or lesion object occupies only a small part of an image, as shown in Fig. 1. Several studies proposed to adopt the bag-of-features (BoF) model with local features to cope with these complex situations [19].

Manuscript received January 16, 2018; revised May 9, 2018 and July 24, 2018; accepted August 10, 2018. Date of publication August 20, 2018; date of current version March 19, 2019. This work was supported in part by the National Natural Science Foundation of China under Grants 81571758, 61871274, 61801305, 61501305 and 81771922; in part by the National Key Research and Development Program (2016YFC0104703); in part by the National Natural Science Foundation of Guangdong Province under Grants 2017A030313377 and 2016A030313047; in part by the Shenzhen Peacock Plan (KQTD2016053112051497); in part by the Shenzhen Key Basic Research Project (JCYJ20170818142347251 and JCYJ20170818094109846); in part by the Hong Kong RGC General Research Fund (PolyU152035/17E); and in part by the National Taipei University of Technology-Shenzhen University Joint Research Program (2018006). (Corresponding author: Baiying Lei.)

Z. Yu, D. Ni, S. Chen, and T. Wang are with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University.

X. Jiang is with the School of Electrical and Electronic Engineering, Nanyang Technological University.

F. Zhou is with the Department of Industrial and Manufacturing, Systems Engineering, The University of Michigan.

J. Qin is with Centre for Smart Health, School of Nursing, The Hong Kong Polytechnic University.

B. Lei is with the National-Regional Key Technology Engineering Laboratory for Medical Ultrasound, Guangdong Key Laboratory for Biomedical Measurements and Ultrasound Imaging, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen 518060, China (e-mail: leiby@szu.edu.cn).

Digital Object Identifier 10.1109/TBME.2018.2866166

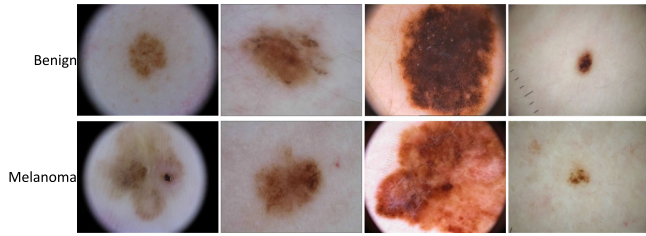


Fig. 1. Example dermoscopy images of skin lesions. There are low inter-class and high intra-class variations between the melanoma and non-melanoma (benign). The diagnosis of melanoma is non-trivial even for experienced clinicians.

Although feature encoding in BoF models, such as basic histogram (histogram of visual word) [21], [22], vector of locally aggregated descriptor (VLAD) [23] and fisher vector (FV) [24], have been widely used in various classification tasks [25], [26], the diagnostic performance delivered by these hand-crafted features is still unsatisfactory due to the high intra-class and low inter-class variations between melanoma and non-melanoma (benign) (examples shown in Fig. 1). In addition, most of these methods contain complex and tedious procedures, which lead to poor generality and inapplicability in clinical practice.

Different from approaches that rely on the hand-crafted features, deep learning methods such as deep convolutional neural networks (CNNs) have established an overwhelming presence in image recognition tasks in the past few years [27]–[30]. The main advantage of CNN is that it is endowed with an impressive visual representation capability for the recognition or detection task depending on the given training dataset [31]. State-of-the-art performance is achieved in numerous applications [32], [33]. Studies in [34], [35] demonstrated that transferred deep convolutional features can be utilized as generic visual representations, and CNN architectures pre-trained on large ImageNet dataset [36] also delivered promising results for other image recognition tasks even without retraining. For this reason, transferred CNN features have been also applied in dermoscopy image classification in recent years [4], [37], [38]. By default, deep convolutional features are extracted from fully connected (FC) layers of a CNN model. Although high-level CNN features have a good generalizability of representing images, these deep descriptors suffer from the paucity of descriptions of local patterns and are sensitive or vulnerable to geometric variations [39]–[41]. For images with dramatic variations in viewpoint and resolution, it would be a great challenge to perform classification directly using CNN features. Rescaling and data augmentation (crop, flip or rotate) strategy are commonly used solutions [34], [35], [42]. However, some transformations of the data might decrease performance. For instance, random cropped images may only capture a non-interested part of the object in the original image, or a background region without the object, in which nuisance representations are exposed to the classifier. Hence, the improvement in performance is highly limited. The situation gets worse when harnessing CNN in medical applications.

Instead of directly using CNN features as general image representations, several studies devoted to combining deep features with local descriptors encoding methods [39], [40], [43]–[46] to

enhance the discrimination capability of these representations. Although impressive improvement is achieved in some benchmarks, these methods [39], [40], [43] are highly computational intensive due to the adoption of sliding-windows to generate deep descriptors from local regions in original images [40], the utilization of multi-scale pyramid pooling strategy to construct FV representations [39], or end-to-end training a CNN architecture integrated with an encoding layer. In our previous study [47], we exploited automatic melanoma classification using FV aggregated CNN features of local patches, which are randomly generated from a dermoscopy image. However, the feature extraction of subimages is time-consuming and computationally intensive. As studied in the literature [41], [59], [71], each CNN activation within a feature map can be traced back to a certain local region (receptive field) of the input image, and reflect the characteristics of that region. Hence, in [44], a more compact and efficient solution is proposed, which is based on densely aggregating local descriptors from convolutional layers within a CNN. Similar to [44], in the medical image analysis field, the studies of [45], [46] harness FV to encode CNN convolutional layer activations to detect interstitial lung disease patterns, and classify microscopy image, respectively. Nevertheless, these kinds of aggregated deep representations have not yet been exploited in dermoscopy images for melanoma recognition. Inspired by these methods [39], [41], [44], [46], we investigate the application of FV aggregated deep features to recognize a melanoma. Our study mainly focuses on these issues: How image feature and feature scale affect the performance of encoded representations? How feature encoding parameters setting affects the performance? How the performance of aggregated CNN activations from different depth networks?

Overall, in this study, we present a framework to address the challenges for automatic and accurate melanoma recognition in dermoscopy images. Our contributions of this study are two-fold. First, we propose a compact, and efficient framework based on very deep CNN and feature encoding strategy (FV encoding) to generate more representative features for more accurate melanoma recognition under limited training data. Using a very deep residual neural network [28] under limited dermoscopic data, activations of the intermediate convolutional layers are extracted as local descriptors, and further encoded by FV into a holistic image representation for each lesion image which is more discriminative than hand-designed descriptors or general CNN features. By elegantly aggregating features generated from the CNN with FV, we take full advantage of the learning capability of deep CNN, exploit the potential of the limited training data and avoid the paucity of representativeness of local patterns when directly using the features extracted by the FC layer, our preliminary result is summarized in [47]. Also, compared with our previous solution [48], the network only applied once to each input image, which indicates more computational efficiency. Second, we systematically investigate various factors in our framework on their impact on performance, including network architectures, number of layers and dimensionality of the FV representation. Further, extensive experiments are conducted to compare our approach with state-of-the-art CNN based methods using the public ISBI 2016 skin lesion data [5]. Our

experimental results demonstrate the effectiveness of our proposed approach. To the best of our knowledge, we do not aware of any previous work that employs FV encoding to integrate deep convolutional features generated from a very deep neural network under limited data to solve this problem.

II. RELATED WORK

A. Hand-Crafted Feature Based Methods

“ABCD” rule has become the standard in dermoscopy for classifying pigmented skin lesions into benign or melanoma [5]. Numerous automatic classification methods based on this rule have been developed [3], [9]. For instance, in [18], Ganster *et al.* adopted the combination of hand-designed features (shape, border gradient, and color descriptors), feature optimization framework and KNN to differentiate melanoma from benign lesions. Celebi *et al.* [2] extracted a series of features from dermoscopy image, including shape features, color, and texture related descriptors. Combine with various feature selection algorithms, a non-linear SVM classifier is trained for the classification. Capdehourat *et al.* [20] proposed to characterize each candidate lesion region by a set of descriptors containing shape, color and texture information, which were utilized to train an AdaBoost classifier. Xie *et al.* [19] presented an ensemble model for melanoma classification. In their study, a self-generating neural network was first adopted to generate lesion regions and a neural network ensemble model is trained for the classification, after the extraction of features (tumor color, texture, and border). Bi *et al.* [11] developed an automatic melanoma detection approach using multi-scale lesion-biased representation and joint reverse classification. For local feature-based method, Situ *et al.* [49] extracted local features (color, scale invariant feature transform (SIFT) [50], etc.) from small 16×16 patches of a dermoscopy image, then aggregated these local descriptors into final representations via a BoF model. Similarly, in [51], Barata *et al.* applied a BoF model to encode texture and color related feature for the classification of lesions.

B. Deep CNN Based Methods

CNN model contains multiple processing layers to learn different levels of representations. Hence, combining these hierarchical features preserves extremely discriminative and effective deep representations [4]. There are mainly two ways of applying CNN to dermoscopy image recognition. One approach is directly training or fine-tuning a deep model in an end-to-end fashion. Demyanov *et al.* [52] developed a five-layer CNN architecture to differentiate two types of skin lesion data. Yu *et al.* [51] proposed a multi-stage scheme based on fine-tuning very deep residual network for automated melanoma recognition in dermoscopy images. Very recently, Esteva *et al.* [53] utilized a single deep network for automatic skin cancer classification. The model was based on architecture of GoogleNet Inception v3, and trained with 129450 clinical images. Menegola *et al.* [38] investigated the impact of knowledge transfer of deep learning in the dermoscopy image recognition. In their study, several source dataset are exploited, includes Atlas, ISIC, Retinopathy,

and ImageNet. However, due to the highly dependence of large training data and computational resources for training an entire CNN, using deep features from pre-trained CNN is also a popular method in medical image domain. For dermoscopy image analysis, Codella *et al.* [4] adopted ImageNet pre-trained CNN to extract high level feature representations to differentiate melanoma and non-melanoma images. Kawahara *et al.* [37] used pre-trained CNN as a feature extractor, and combined with subimage features pooling for 10-classes lesion classification. Codella *et al.* [54] investigated a deep learning ensemble method for melanoma recognition. In their solution, latest deep residual network [28], U-Net architecture [33] and Caffe-Net [55] were combined for feature extraction. Simultaneously, traditional low-level visual descriptors and sparse coding were also adopted for the ensemble learning. In our previous study [48], we proposed an automatic melanoma recognition by aggregating CNN activations of subimages randomly extracted from a dermoscopy image.

III. METHODOLOGY

In this section, we present our proposed framework in details. We first introduce the deep residual neural network applied in our method, followed by the extraction of local dense activations as deep convolutional features in our framework. Then we elaborate how FV encoding strategy is utilized to aggregate these deep features for more discriminative and robust representations. Finally, the classification method of the FV representations is presented. The flowchart of the whole framework is illustrated in Fig. 2.

A. Deep Residual Neural Network

The deep hierarchy architecture of CNN models is of crucial importance for its powerful learning capability [28], [31]. In this work, we adopt the latest generation of the convolutional neural network (deep residual neural network, ResNet) introduced by He *et al.* [28], which is ranked number one in ImageNet large-scale visual recognition challenge 2016 (ILSVRC 2016) for feature extraction. Compared with typical CNN architectures, the main characteristic of ResNet lies in the adaptation of residual connection which is capable of addressing the degradation problem [28] when training a very deep network. It has been demonstrated that the residual links can speed up the convergence of deep network and maintain accuracy gains achieved by substantially increasing the network depth. Generally, a deep residual network consists of a set of residual blocks, and each block is composed of several stacked convolutional layers (we regard rectified linear unit layer and batch normalization layers as an appendage of the convolutional layer). A residual block with identity mapping can be formulated as:

$$h_{l+1} = Relu(h_l + \mathcal{F}(h_l, w_l)), \quad (1)$$

where h_l and h_{l+1} are input and output of the l -th residual block, respectively; $Relu(\blacksquare)$ is rectified linear unit function; \mathcal{F} denotes the residual mapping function and w_l are the parameters of the block. Specifically, when the channels (dimensions) of $\mathcal{F}(h_l, w_l)$ and h_l are unequal, a linear projection ϕ is

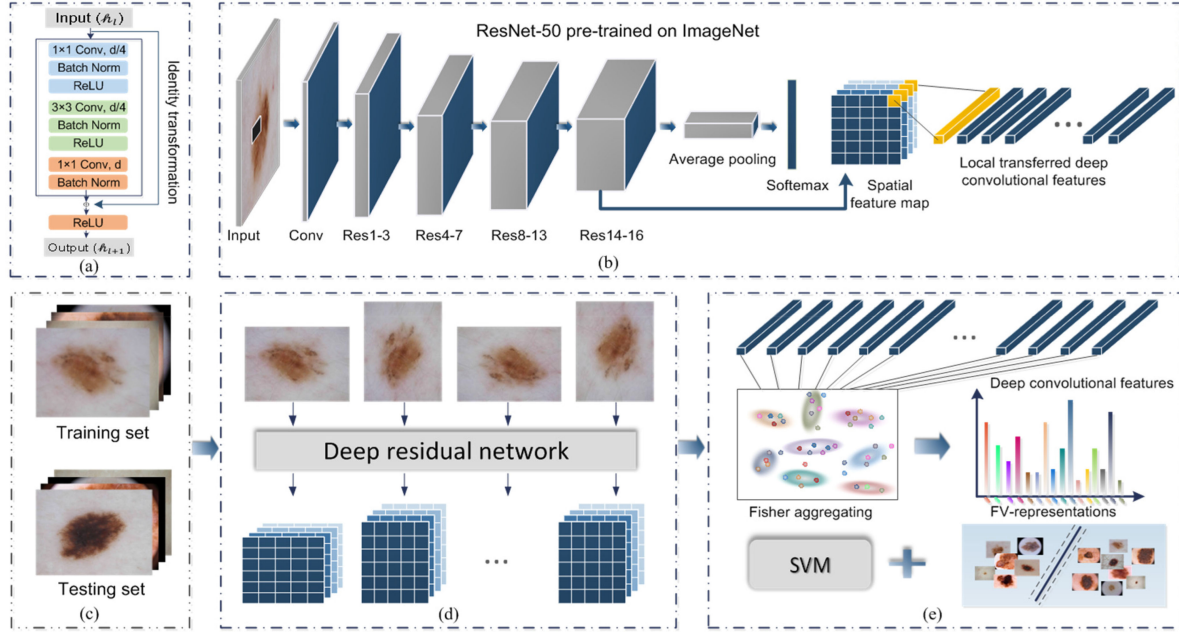


Fig. 2. Flowchart of the proposed framework for melanoma recognition. (a) Illustration of the residual block, which is abbreviated as “Res”, d represents the depth of feature maps. (b) Local feature extraction, outputs of last Residual block are extracted as deep feature vectors. (c) Dataset of dermoscopic lesion images. (d) Data augmentation and feature extraction. (e) FV encoding and classification.

TABLE I

THE ARCHITECTURE OF RESNET-50 (RESNET-101) USED IN THIS STUDY

Layer/Residual block		Parameter (kernels, channels)	
Conv1		7×7, 64	
Max pooling, 3×3, 64			
Conv2_x	ResBlock1-3	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix}$	×3
Conv3_x	ResBlock4-7	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix}$	×4
Conv4_x	ResBlock8-13(8-30)	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix}$	×6(23)
Conv5_x	ResBlock14-16(31-33)	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix}$	×3
Average pooling, 7×7, 2048			
FC, softmax output, 1×1, 1000			

usually applied to match the dimensions, then Eq. (1) can be further converted to:

$$h_{l+1} = \text{Relu}(\phi(h_l) + \mathcal{F}(h_l, w_l)). \quad (2)$$

A basic residual block is illustrated in Fig. 2(a) and readers can refer to [28] for more detail.

In our study, two ResNet models with different depth are exploited (i.e., ResNet-50 and ResNet-101) and both the residual networks are pre-trained on ImageNet. The detailed information of the architectures is shown in Table I, and the convolutional layer is denoted as Conv for simplicity. We also investigate several different CNN architectures (i.e., AlexNet [29], VGGNet [30]) for comparison.

B. Image Preprocessing and Data Augmentation

- 1) *Image size*: Since there is a huge variation in image resolutions of the skin lesion dermoscopy dataset provided by ISBI 2016 challenge [5], ranging from the largest scale (4288×2848) to the smallest scale (722×542). Resizing and cropping these images directly into required size introduces object distortion and substantial information loss [41], [43], [56]. In this study, we take relatively large (compared with 224×224) as inputs. For the skin lesion dataset, we resize these images along the shortest side to a uniform scale (hereafter, we denote this scale as S for simplicity) while maintaining the aspect ratio. We also investigate the recognition performance with various values of S , and the results are presented in Section IV.
- 2) *Image normalization and augmentation*: Typically, before processing by CNN, images are normalized by subtracting the mean pixel value, which is calculated over the entire training dataset. As a result, the RGB values are centered at zero (denoted as all-img-mean). However, the lighting, skin tone and viewpoint of the skin lesion images vary greatly across the dataset, subtracting a uniform mean value does not well normalize the illumination of individual images. A recent study [16] has illustrated this effect as well.

To address this issue, we normalize each skin image by subtracting channel-wise average intensity values calculated over the individual image (denoted as per-img-mean). In this work, we also investigate the influence of the two different normalization approaches. The results are provided in Section IV. In addition, simple rotation and translation based augmentation is adopted to further improve the performance. Specifically, we

rotate each resized image by four degrees ($0^\circ, 90^\circ, 180^\circ$, and 270°), and then pixel translation (with a shift between -10 and 10 pixels) is randomly added over the rotated images. The deep features of these augmented images are aggregated into a single FV representation. Details are presented in Section III-D.

C. Extraction of Local Convolutional Features

Given a pre-trained network, an input skin lesion image \mathcal{X}_i is first processed by the above mentioned operations. Supposing the shorter side of resized images is fixed as 448 (i.e., $S = 448$), we obtain four augmented images $\mathbb{X}_i = \{\mathcal{X}_{i1}, \mathcal{X}_{i2}, \mathcal{X}_{i3}, \mathcal{X}_{i4}\}$ with size of $448 \times N$ (or $N \times 448$) for each skin image. These images are passed through the CNN model in a forward pass. In the l -th convolutional layer \mathcal{L}_l , we obtain $w_{ia}^l \times h_{ia}^l \times d^l$ spatial feature maps \mathcal{M}_{ia}^l ($a = 1, \dots, 4$), where w_{ia}^l and h_{ia}^l denote the width and height, respectively, d^l is the depth or channels of the current feature map. For brevity, we denote $\mathcal{N}_{ia}^l = w_{ia}^l \times h_{ia}^l$. It is worth noting that, for input images with different sizes, the size of the resulting feature maps can be different. Similar to [41], for activations at each location $c = (c_x, c_y)$, $1 \leq c_x \leq w_{ia}^l$ and $1 \leq c_y \leq h_{ia}^l$ in the feature map \mathcal{M}_{ia}^l , we obtain d^l -dimensional vector $f_{ia,c}^l \in \mathbb{R}^{d^l}$ which is considered as feature vector (local deep feature) [41] in our study. Thus, \mathcal{N}_{ia}^l local deep feature vectors are obtained for each augmented image \mathcal{X}_{ia} , denoted as:

$$\mathcal{F}_{ia}^l = \{f_{ia,(1,1)}^l, \dots, f_{ia,c}^l\} \in \mathbb{R}^{\mathcal{N}_{ia}^l \times d^l}. \quad (3)$$

At this point, for i -th original skin lesion image \mathcal{X}_i , at convolutional layer \mathcal{L}_l of the network, we obtain a set of deep features

$$\mathbb{F}_i^l = \{f_{i1,(1,1)}^l, \dots, f_{i4,(w_{i4}^l, h_{i4}^l)}^l\} \in \mathbb{R}^{\sum_{a=1}^4 \mathcal{N}_{ia}^l \times d^l} \quad (4)$$

These features are encoded by FV into a global image representation for the final classification.

Conventionally, deep image representations are extracted at the final layer of a pre-trained CNN model after removing the softmax layer (classifier layer). These high-level features are robust to geometric variance and other image transformations to some extent [40], [56]. However, features from higher layers are more domain-specific and fail to capture characteristics of local patterns [41], [43], [57], [58]. In this study, the features extracted from various convolutional layers of deep neural networks are explored. Also, the fusion results are reported by averaging scores of different layers in the proposed method.

D. Fisher Vector Encoding Strategy

Each local deep convolutional feature f_n^l extracted from layer \mathcal{L}_l , refers to a small region (receptive field) in the input image, and reflects the local distinction of that region [41], [59]. This is similar to traditional local descriptors (i.e., SIFT [50]). Since each image contains a set of deep features, we propose to aggregate these local deep representations into a single image representation (FV representation) using FV encoding method, which can be regarded as a variant of BoF model. The FV

encoding derived from fisher kernel is effective for encoding local features and has demonstrated excellent performance in image recognition [60], [61].

- 1) *Gaussian mixture model*: To implement FV encoding, a probability distribution (generative model) of deep features \mathbf{f} (i.e., $P(\mathbf{f}|\lambda)$) needs to be specified. To achieve this, the popular Gaussian mixture model (GMM), which can well approximate arbitrary continuous distribution, is adopted to model the probability distribution (generation process) of deep features. GMM is a parametric estimation model of probability density function and has been regarded as “probabilistic visual word vocabulary” [24]. For more details, interested readers can refer to [62].
- 2) *Fisher vector encoding*: For a set of local deep features $\{\mathcal{F}_{i1}^l, \mathcal{F}_{i2}^l, \mathcal{F}_{i3}^l, \mathcal{F}_{i4}^l\}$ extracted from the augmented images of the i -th original skin lesion image, the first and second order differences of the GMM clusters are given by:

$$\mathbf{u}_k = \frac{1}{N\sqrt{\pi_k}} \sum_{n=1}^N q_{kn} \left(\frac{f_n^l - \mu_k}{\Sigma_k^{1/2}} \right), \quad (5)$$

$$\mathbf{v}_k = \frac{1}{N\sqrt{2\pi_k}} \sum_{n=1}^N q_{kn} \left[\frac{(f_n^l - \mu_k)^2}{\Sigma_k} - 1 \right], \quad (6)$$

$k = 1, 2, \dots, K,$

where $N = \sum_{a=1}^4 \mathcal{N}_{ia}^l \times d^l$ represents the number of local deep descriptors of a skin image, q_{kn} denote the soft assignment of feature vector f_n^l to cluster k . By concatenating \mathbf{u}_k and \mathbf{v}_k for all K components, we obtain the final FV representation Φ_i :

$$\Phi_i = [\mathbf{u}_1^T, \mathbf{v}_1^T, \dots, \mathbf{u}_K^T, \mathbf{v}_K^T]^T. \quad (7)$$

It is noteworthy that the dimensionality of the deep feature vector is reduced by principal component analysis (PCA) before FV encoding because more Gaussian components are needed to capture the distribution of higher dimensional feature vector. Accordingly, the dimensionality of image representation is reduced [63]. For each FV representation, we further compute the improved FV by applying L2 and power normalization the same as that in [24], [26].

E. Kernel-based Classification

For the classification of the FV representations, we train an SVM classifier with Chi-squared (chi2) kernel. Although linear kernels are efficient for the classification, non-linear kernels tend to yield better performance and empirical studies have demonstrated the superiority of the chi2 kernel for image classification [64], [65]. The homogeneity degree of the kernel is set to 1 in our experiment. Note that the FV representations are L2 normalized. In our experiment, we adopt the standard hinge loss in the objective function and the parameter C use to scale the loss is fixed as 1. During SVM training, the stochastic dual coordinate ascent algorithm (SDCA solver) is employed to minimize the regularized loss because of its efficiency and fast convergence rate.

IV. EXPERIMENTAL SETTING AND RESULTS

A. Experimental Setting

- 1) *Dataset*: We validate our proposed method using ISBI 2016 challenge dataset of dermoscopic lesion images [5]. This dataset is based on the international skin imaging collaboration (ISIC) archive,¹ which is the largest publicly available collection of quality controlled dermoscopic images of skin lesions. The dataset released in the challenge contains 1279 dermoscopic lesion images with corresponding class labels pre-partitioned into a training set of 900 images and a testing set of 379 images. There are two lesion categories in the dataset: melanoma and benign (non-melanoma). Approximately 20% of the dataset is melanoma (173 images in training set, 75 images in testing set). In our experiment, the hyper-parameters of SVM classifiers are obtained by using cross-validation strategy on the training data.
- 2) *Evaluation and implementation*: For performance metrics, we adopt the mean average precision (mAP), accuracy (Acc), area under receive operation curve (AUC). The detailed formulation of these measures is described in [5]. The implementation of the proposed framework is mainly based on open source libraries including MatConvNet [66] for deep feature extraction and VLFeat [67] for FV encoding. All experiments are conducted on a computer with CPU Inter Xeon E5-2680 @ 2.70 GHz, GPU NVIDIA Quadro K4000, and 128G of RAM.

B. Experiments on Image Preprocessing and Augmentation

The objective of this experiment is to demonstrate the importance of image preprocessing and augmentation for the deep feature extraction and classification. The preprocessing adopted in our study mainly includes image resize and normalization. Each original dermoscopy image is resized along its shortest side into a fixed scale S , while maintaining the aspect ratio of the image (described in Section III-B). We first extract deep features of the dermoscopy images from the last convolutional layers of pre-trained ResNet-50 and then obtain the FV representations under various preprocessing settings. The classification result is shown in Table II and Fig. 3. As seen from Fig. 3(a), in the case of adopting the normalization strategy of per-img-mean, the classification performance improves gradually as the scale increases and remains stable after the scale reaches 448 (double size of the required network input). For normalization with all-img-mean, the classification performance first increases as the increasing S . It reaches a peak at $S = 384$ followed by a steep fall. The detailed results are summarized in Table II. Our tentative explanation of this phenomenon is that a large S value will increase complexity and variation between images intensity values. Simply subtracting a uniform mean pixel value will decrease the performance of the deep representations. Although a large input is capable to improve the result by preserving more

TABLE II

IMPACT OF IMAGE PREPROCESSING ON THE CLASSIFICATION RESULTS (%)

Scale (S)	Image-norm	mAP	Acc	AUC
224	per-img-mean	53.95	82.32	77.78
	all-img-mean	58.89	82.85	82.29
256	per-img-mean	58.74	83.64	79.13
	all-img-mean	59.05	84.17	82.19
384	per-img-mean	60.18	84.17	80.96
	all-img-mean	62.44	85.49	82.97
448	per-img-mean	65.08	86.54	81.49
	all-img-mean	59.02	84.17	80.30
512	per-img-mean	64.74	84.70	82.19
	all-img-mean	59.51	83.91	80.78

information, the improvement is insufficient to compensate for the performance loss. By balancing the memory consumption and efficiency, we fix S as 448 in the following experiments. We also explore the impact by maintaining the aspect ratio of the lesion images. The result is illustrated in Fig. 3(b), where the images with the aspect ratio unchanged denoted as “non-square-img”, otherwise “square-img”. It can be seen that maintaining the aspect ratio is beneficial when the image scale becomes large. At $S = 448$, an improvement of $\sim 2\%$ in mAP is achieved. To verify the effectiveness of rotation and translation based augmentation adopted in our framework, the experiments with and without augmentation are conducted. Fig. 3(c) shows the comparison results with various scales. By augmentation, a significant improvement is achieved except for the size of 224.

C. Experiments on Network Types and Convolutional Features

- 1) *Effectiveness of different network architectures*: Apart from the 50-layers ResNet (ResNet-50), we explore several other CNN models with different depth including 8-layers AlexNet, 16-layers VGGNet (VGG-16), and much deeper 101-layers ResNet (ResNet-101) for performance comparison. The total number of parameter (para for short) in these networks is shown in Table III. All the models are pre-trained on ImageNet. The same image preprocessing and FV encoding parameters setting (i.e., GMM numbers, PCA dimensions) are adopted for fair comparison. Table III shows the experimental results. It is observed that the network architectures have a great impact on the performance, and ResNet-50 and ResNet-101 achieve better performance than the other two networks. Note that features extracted from shallower AlexNet outperform those from the VGG-16. There is a margin of $\sim 4\%$, $\sim 0.4\%$, and $\sim 1\%$ in mAP, Acc, and AUC, respectively, which indicates that lower error rates on ImageNet do not always lead to better performance in other tasks. The average running time of each network for the deep feature extraction and encoding over single skin lesion image is provided in Table III. Although ResNet models are deeper and more complex than VGG-16 and AlexNet, the computational time margin is quite small. Replacing AlexNet with ResNet-50 and ResNet-101, there is only an

¹<https://isic-archive.com/>

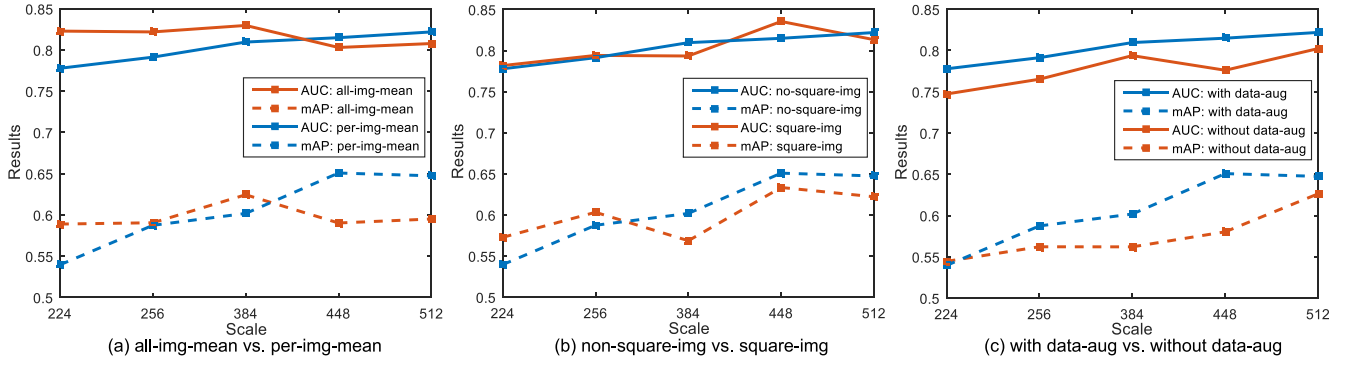


Fig. 3. Evaluation of proposed method on image preprocessing and augmentation. Due to the extreme imbalance of the lesion dataset, mAP is considered as the main metric, and AUC measured as a reference in this figure.

TABLE III
IMPACT OF NETWORK ARCHITECTURES ON THE
CLASSIFICATION RESULTS (%)

Network	Para	Layers	mAP	Acc	AUC	Time
AlexNet	61M	Conv5	61.37	84.70	82.08	0.94s
VGG-16	138M	Conv5_3	57.66	84.43	81.18	2.72s
ResNet-50	25.6M	Conv5_9	65.08	86.54	81.49	1.33s
ResNet-101	44.5M	Conv5_9	66.57	86.81	82.48	2.09s

increase ~ 0.4 s and ~ 1.0 s in computational time, which demonstrates the efficiency of our method. Considering the small performance margin between ResNet-50 and ResNet-101, the former is adopted in our next experiment for efficiency.

- 2) *Effectiveness of convolutional features in different levels*: Features from different convolutional layers within a network vary in their receptive field sizes and semantic levels. To investigate the performance of the convolutional features in different levels, we adopt FV to encode deep features extracted from various convolutional layers of different networks. Two different scales over each network (one is 448 and the other is 224) are utilized. In ResNet-50, we consider residual blocks with the same output feature map size as a group, and the last convolutional layer in the group is regarded as representative and adopted for feature extraction. In VGG-16, similar to ResNet-50, we take convolutional layers with equal output size as a group, and the last layer of a group is considered as representative. In AlexNet, all the five convolutional layers are adopted.

Fig. 4 shows the performance of different level features for three networks. By and large, there is a relatively clear trend in the results of the three networks. Features from the first few layers are inferior. This is expected since bottom layer features hold smaller receptive field which captures low-level visual patterns, hence generally lack of invariance and discrimination. These features to some extent resemble the visual descriptors in the BoF models with SIFT [50]. However, there is no significant improvement observed in the performance after reaches a certain intermediate layer and the metrics even gradually drop. Table IV

shows the detailed results of the intermediate layers (Mid. layer) and last convolutional layers (Last layer) for three models on the scale of 448. Furthermore, the fusion results by averaging the scores of two different layers are further computed for the three networks. Comparing the best single layer results in Table IV, there are significant improvements in mAP of $\sim 2\%$, $\sim 4\%$, $\sim 3\%$ for AlexNet, VGG-16, and ResNet-50, respectively. Comparing these results, we can conclude that the highest layer features are not always optimal, though these high-level representations demonstrate more invariant information to achieve the impressive performance in many other classification tasks. In transfer problem, there is a significant gap between original dataset and target dataset. In addition, layers perform quite differently in different network architectures. Hence, it is not trivial to give an indication of which layers are superior. In our proposed method, the performance can be improved by fusing results of FV representations encoded from different levels convolutional features within a network.

D. Experiments on Feature Encoding Strategy

We conducted an experiment to gain insight of how parameters of FV encoding affect the classification performance. Apart from the size of GMM codebook, we also investigated the influence of dimensionality of the deep feature. Fig. 5 shows the performance of our method in various settings of the number of Gaussian components and the feature dimensionality. It can be observed that increasing the dimensions of the deep feature, yields significant improvements of the performance metrics initially.

As the dimensionality becomes higher, the performance gradually drops in all the GMM components setting, which indicates that the current number of Gaussian components and the number of training samples are insufficient to model the distribution of higher dimensional features. In addition, we can see that the performance of a larger number of Gaussian components (i.e., GMM Num = 100) outperforms the other fewer Gaussians in the high dimensionality of 500, which suggests that increasing the number of GMM components can improve the performance. However, for a larger number of GMM components, more training data is needed to estimate the GMM parameters [26], [62]. Furthermore, increasing Gaussians numbers under the case of

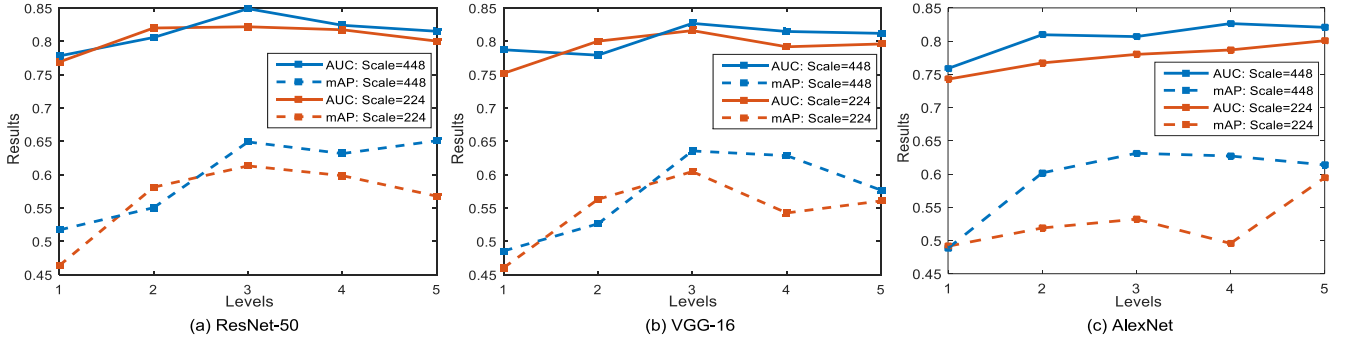


Fig. 4. Evaluation of proposed method on different network architectures and layers. We mainly consider major layers within these networks. For AlexNet and VGG-16, fully connected layers are discarded due to incompatible size.

TABLE IV
PERFORMANCE OF DIFFERENT LEVEL FEATURE FOR THREE NETWORKS ON A SCALE OF 448 (%)

Network	Layer	mAP	Acc	AUC
AlexNet	Mid. layer	63.12	85.49	80.68
	Last layer	61.37	84.70	82.08
	Fusion	65.26	86.02	82.41
VGG-16	Mid. layer	63.54	85.49	82.68
	Last layer	57.66	84.43	81.18
	Fusion	67.12	85.22	83.98
ResNet-50	Mid. layer	64.90	85.75	84.88
	Last layer	65.08	86.54	81.49
	Fusion	68.49	86.81	85.20

high dimensional features leads to very high memory consumption and computational complexity [63]. Next, the testing results of replacing the FV encoding adopted in our proposed method by VLAD encoding denoted as DCNN-VLAD are recorded in Table V. In order to obtain comparable dimensional representations, the vocabulary size is set to 100. It can be seen that this replacement account for a drop of around 3% in mAP and $\sim 1.3\%$ in Acc (DCNN-VLAD vs. DCNN-FV).

E. Comparison of the Classification with Other Methods

- 1) *Comparison with traditional hand-crafted feature:* SIFT descriptor is a good representative of hand-crafted features, which has been widely adopted in a myriad of applications [26], [63], [68]. We compare the BoF models based on densely sampled SIFT (DSIFT) descriptors to our framework. The comparison result is presented in Table V. In all three BoF models, local DSIFT descriptors are extracted equally with a stride of three pixels for each skin lesion image. DSIFT-VQ denotes basic vector quantization. The dictionary size of visual words is set to 1024. To be fair, SPM algorithm is adopted which leads to 8192-dimensional representation for each skin lesion image. In DSIFT-VLAD model, the vocabulary size is set to 100, which result in 12800-dimensional representation. In DSIFT-FV, the number of GMM components is correspondingly set to 64 to keep it the same as our method, and the 128-dimensional DSIFT descriptor is reduced to

the dimensionality of 100 by PCA. The dimensionality of final representation is 12800. It is noteworthy that we keep the same setting in the process of classification whenever possible.

We can see that both DSIFT-VLAD and DSIFT-VQ are inferior to DSIFT-FV model, and there are performance gaps of $\sim 4\%$ and $\sim 10\%$ in mAP, respectively, which demonstrate the superiority of FV for aggregating local descriptors. In Table V, we denote our method as DCNN-FV. It can be observed that about 10% improvements in mAP are achieved when replacing DSIFT with deep convolutional features (CNN-FV vs. DSIFT-FV and CNN-VLAD vs. DSIFT-VLAD), which indicates that the deep CNN features possess significantly more powerful representation ability than low-level hand-crafted descriptors.

- 2) *Comparison with existing CNN based methods:* We perform a set of experiments to compare the proposed framework with other CNN based methods. Table VI lists the comparison results. For CNN-SVM, we first rescale each skin lesion image into required squared size, then high dimensional CNN features are extracted and further classified by SVM. For AlexNet and VGG-16, we extract the activations of the first fully-connected layer as image representations. For ResNet-50, 2048 dimensional outputs of the penultimate layer are computed. Note that we use chi-square kernel as well in SVM for a fair comparison. Similar to [34], in CNNAug-SVM, we augment each skin image with randomly rotation, flipping, and cropping which result in 32 subimages. The scores of these subimages are averaged for final results.

As shown in Table VI, both models benefit from data augmentation, and there is a small but significant improvement of 1%~2% margin in mAP. Same as in Section III-C, deep features from AlexNet outperform VGG-16, and ResNet-50 achieves the best performance among three models with mAP of 59.93%. This result, nevertheless, is inferior to our method by a large margin of $\sim 5\%$ in mAP. This shows the limitation of directly using CNN features for the recognition due to the huge variations in viewpoint and resolution of lesion images, which cannot be well represented by features learned from limited training data. Our method achieved significant improvement even without adaptation of transferred features across domains, demonstrating that aggregating the features produced by CNN

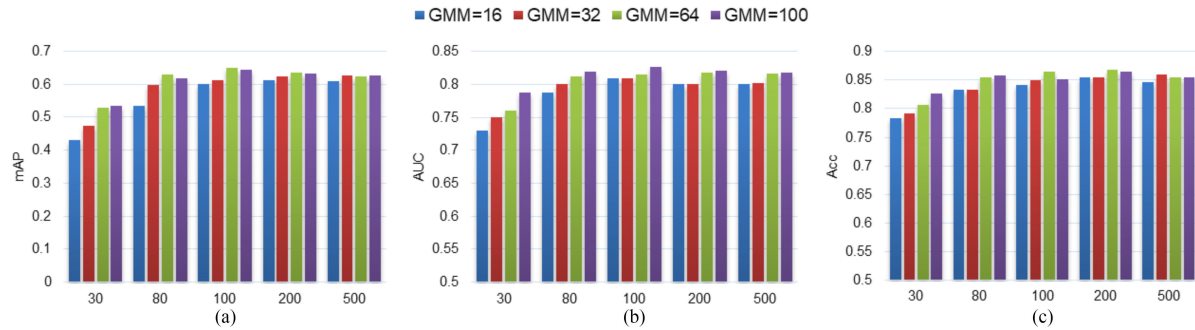


Fig. 5. Performance of the proposed method with varying number of Gaussian components and dimensionality of deep feature (ranges from 30 to 500). (a), (b), and (c) denote the evaluation results of mAP, AUC, and Acc respectively.

TABLE V
PERFORMANCE OF THE PROPOSED APPROACH WITH
HAND-CRAFTED FEATURE BASED METHODS (%)

Method	Codebook	Dimension	mAP	Acc	AUC
DSIFT-VQ	1024	8192	46.94	81.53	76.97
DSIFT-VLAD	100	12800	51.51	82.06	78.53
DSIFT-FV	64	12800	55.63	83.11	78.01
DCNN-VLAD	100	12800	62.23	85.22	82.57
DCNN-FV	64	12800	65.08	86.54	81.49

using effective statistical models can generate more discriminative representations for better recognition performance. We further fine-tune ResNet-50 for comparison. To circumvent the risk of overfitting caused by limited training data, a real-time data augmentation technique is adopted. Namely, we augment the data during training, while the network is being trained on a chunk of data on GPU, and the next chunk would be generated on CPU in multiple processes. Table VI demonstrates that extracting deep features from the fully-connected layer of a pre-trained network or directly fine-tuning an entire deep network may not be optimal for the skin lesion images due to the paucity of representative training samples.

Finally, we compare our result with the top-ranked method [70] in the challenge and methods reported in the recently published literature [54], [69]. For Fisher-GL [69], we randomly sample 400 windows from each skin lesion image and adopt pre-trained AlexNet to compute features for each window. Then these features are aggregated into a single representation by FV, which is similar to our method. As seen in Table VI, our proposed framework outperforms the three methods in mAP by margins of $\sim 1.3\%$, $\sim 0.5\%$, and $\sim 6\%$, respectively (in no fusion case). Further, a simple fusion by averaging scores of two different layers in our method leads to $\sim 4\%$ advancement in mAP comparing to the best-performing approach in the literature [54]. It is noteworthy that these methods have either high process complexity or high computational complexity. For example, in [70], the utilized multi-stage scheme involves additional segmentation preprocessing and network fine-tuning. Also, in [54], the method heavily relies on the ensemble strategy, and it combines three pre-trained networks, sparse coding and a set of hand-crafted features. In contrast, our proposed method

TABLE VI
COMPARISON OF THE PROPOSED APPROACH WITH OTHER
CNN BASED METHODS (%)

Method	Network	mAP	Acc	AUC
CNN-SVM [34]	AlexNet	56.59	83.38	80.00
	VGG-16	57.70	84.70	78.61
	ResNet-50	58.42	84.43	81.82
CNNaug-SVM [34]	AlexNet	58.35	83.64	80.13
	VGGNet	57.72	83.38	81.08
	ResNet-50	59.93	83.19	81.73
Fisher-GL [69]	AlexNet	58.95	82.85	81.21
Fine-tuned CNN [28]	ResNet-50	63.36	84.96	81.58
CUMED [70]	ResNet-50	63.70	85.50	80.40
Codella [54]	Hybrid Net	64.50	80.50	83.80
DCNN-FV	ResNet-50	65.08	86.54	81.49
DCNN-FV (fusion)	ResNet-50	68.49	86.81	85.20

is more compact and efficient, which can be easily generalized to other application as well.

V. DISCUSSIONS

We have presented a novel and efficient method for automatic and accurate recognition of melanoma from dermoscopy images and performed extensive experiments to investigate its effectiveness. Beyond the impressive results, there are several factors we should pay attention to. First, input image size has a great effect on the recognition performance. Using large image is demonstrated beneficial for the performance improvement. However, with the scale gets larger, careful image preprocessing and normalization should be taken into account. Furthermore, choosing proper CNN architectures of deep convolutional features is very critical. In addition, the FV encoding settings should be in line with the limitation of training data. It is also noteworthy that, in our framework, the network only applied once to each input image and then deep descriptors are extracted from the dense activation maps. Thus, comparing with existing similar hybrid methods which generate deep features using sliding-windows [40] or construct FV representations using multi-scale pyramid pooling strategy [39], the proposed framework is more computationally efficient.

Although the proposed method achieves quite promising results, there are still some limitations. First, for transfer problem, the original training data have a significant impact on the target tasks [38], [57]. For example, network pre-trained on ImageNet

may yield different results when compared with a counterpart that was trained on other datasets [71]. A recent study [38] also exploited this issue on dermoscopy images. However, we have only investigated the networks that are pre-trained on ImageNet. Obtaining more insight on the influence of other datasets is surely beneficial. Second, during the encoding step, we heavily compress the high dimensional deep features due to the constraint on the number of Gaussian components of GMM caused by the limited number of training samples. This limits the discriminative information captured by the GMM. Our future work will explore the effectiveness of exploiting it in our framework. Finally, the proposed framework is merely tested on the ISBI skin lesion dataset. It is interesting to see how our proposed method performs on other data, which is surely another future work in our group.

VI. CONCLUSION

In this paper, we propose a compact framework for dermoscopy image classification. It utilizes the state-of-the-art local descriptors encoding method (FV) to encode local convolutional features extracted from a very deep residual network into more sophisticated representations. Our method is motivated by the fact that discriminative characteristics of an image are generally well-captured by the feature maps of a CNN architecture. For a convolutional layer, usually, hundreds or thousands of kernels are equipped to generate numbers of activation maps to capture various aspects of the image. Therefore, we can aggregate the dense local activations of these feature maps to construct more sophisticated representations than both the hand-designed descriptors and the direct CNN features. Comparing with the existing methods based on end-to-end fully training or fine-tuning a network, our framework only utilizes pre-trained CNN as a feature extractor, thus the complexity in training process is circumvented, which is more practicable under limited training samples.

Systematical and extensive experiments are performed to investigate a range of key elements that could affect the performance of our method, including image preprocessing, data augmentation, network architectures, levels of convolutional features and parameters setting of FV encoding. Also, we compare the proposed framework with the existing well-established methods for further validating its superiority. Our work shows that deep convolutional features can be aggregated by FV encoding effectively and efficiently, and the encoded deep representations are more discriminative than hand-crafted descriptors and CNN features. Results are reported on the publicly available ISBI 2016 challenge skin lesion dataset. Future investigations include evaluating our method on more datasets and promoting its application in clinical practice.

REFERENCES

- [1] R. Kasmi and K. Mokrani, "Classification of malignant melanoma and benign skin lesions: Implementation of automatic ABCD rule," *IET Image Process.*, vol. 10, pp. 448–455, 2016.
- [2] K. H. M. Celebi et al., "A methodological approach to the classification of dermoscopy images," *Comput. Med. Imag. Grap.*, vol. 31, pp. 362–373, 2007.
- [3] A. R. A. Ali and T. M. Deserno, "A systematic review of automated melanoma detection in dermoscopic images and its ground truth data," *Proc. SPIE*, vol. 8318, 2012, Art. no. 831811.
- [4] N. Codella et al., "Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images," presented at the Proc. Med. Image Comput. Comput.-Assisted Intervention, Quebec, QC, Canada, 2015.
- [5] D. Gutman et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, 2018, pp. 168–172.
- [6] W. Stolz et al., "ABCD rule of dermoscopy: A new practical method for early recognition of malignant-melanoma," *Eur. J. Dermatol.*, vol. 4, pp. 521–527, 1994.
- [7] S. W. Menzies et al., "Frequency and morphologic characteristics of invasive melanomas lacking specific surface microscopic features," *Arch. Dermatol.*, vol. 132, pp. 1178–1182, 1996.
- [8] J. S. Henning et al., "The CASH (color, architecture, symmetry, and homogeneity) algorithm for dermoscopy," *J. Amer. Acad. Dermatol.*, vol. 56, pp. 45–52, 2007.
- [9] K. Korotkov and R. Garcia, "Computerized analysis of pigmented skin lesions: A review," *Artif. Intell. Med.*, vol. 56, pp. 69–90, 2012.
- [10] U. Jamil and S. Khalid, "Comparative study of classification techniques used in skin lesion detection systems," presented at the IEEE Int. Multi-topic Conf., Karachi, Pakistan, 2014.
- [11] L. Bi et al., "Automatic melanoma detection via multi-scale lesion-biased representation and joint reverse classification," presented at the IEEE 13th Int. Symp. Biomed. Imag., Prague, Czech Republic, 2016.
- [12] T. Lee et al., "Dullrazor: A software approach to hair removal from images," *Comput. Biol. Med.*, vol. 27, pp. 533–543, 1997.
- [13] A. Sáez et al., "Model-based classification methods of global patterns in dermoscopic images," *IEEE Trans. Med. Imag.*, vol. 33, no. 5, pp. 1137–1147, May 2014.
- [14] H. Iyatomi et al., "Automated color calibration method for dermoscopy images," *Comput. Med. Imag. Graph.*, vol. 35, pp. 89–98, 2011.
- [15] G. Schaefer et al., "Colour and contrast enhancement for improved skin lesion segmentation," *Comput. Med. Imag. Graph.*, vol. 35, pp. 99–104, 2011.
- [16] M. Rastgoo et al., "Automatic differentiation of melanoma from dysplastic nevi," *Comput. Med. Imag. Graph.*, vol. 43, pp. 44–52, 2015.
- [17] K. Shimizu et al., "Four-class classification of skin lesions with task decomposition strategy," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 1, pp. 274–283, Jan. 2015.
- [18] H. Ganster et al., "Automated melanoma recognition," *IEEE Trans. Med. Imag.*, vol. 20, no. 3, pp. 233–239, Mar. 2001.
- [19] F. Xie et al., "Melanoma classification on dermoscopy images using a neural network ensemble model," *IEEE Trans. Med. Imag.*, vol. 36, no. 3, pp. 849–858, Mar. 2017.
- [20] G. Capdehourat et al., "Toward a combined tool to assist dermatologists in melanoma detection from dermoscopic images of pigmented skin lesions," *Pattern Recognit. Lett.*, vol. 32, pp. 2187–2196, 2011.
- [21] S. Lazebnik et al., "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," presented at the IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., New York, NY, USA, 2006.
- [22] G. Csurka et al., "Visual categorization with bags of keypoints," presented at the 8th Eur. Conf. Comput. Vis./Workshop Statist. Learn. Comput. Vis., Prague, Czech Republic, 2004.
- [23] H. Jegou et al., "Aggregating local image descriptors into compact codes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1704–1716, Sep. 2012.
- [24] J. Sánchez et al., "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, pp. 222–245, 2013.
- [25] K. Chatfield et al., "The devil is in the details: An evaluation of recent feature encoding methods," presented at the 22nd Brit. Mach. Vis. Conf., Dundee, U.K., 2011.
- [26] F. Perronnin et al., "Improving the Fisher kernel for large-scale image classification," presented at the 11th Eur. Conf. Comput. Vis., Berlin, Germany, 2010.
- [27] Y. Song et al., "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 10, pp. 2421–2433, Oct. 2015.
- [28] K. He et al., "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 770–778.
- [29] A. Krizhevsky et al., "Imagenet classification with deep convolutional neural networks," presented at the Proc. Int. Conf. Neural Inf. Process. Syst., Lake Tahoe, NV, USA, 2012.

- [30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–14.
- [31] Y. LeCun *et al.*, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.
- [32] Y. Song *et al.*, "Segmentation, splitting, and classification of overlapping bacteria in microscope images for automatic bacterial vaginosis diagnosis," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 4, pp. 1095–1104, Jul. 2017.
- [33] O. Ronneberger *et al.*, "U-net: Convolutional networks for biomedical image segmentation," presented at the Conf. Med. Image Comput. Comput. Assisted Intervention, Munich, Germany, 2015.
- [34] A. Sharif Razavian *et al.*, "CNN features off-the-shelf: an astounding baseline for recognition," presented at the IEEE Conf. Comput. Vis. Pattern Recognit./DeepVis. Workshop, Columbus, OH, USA, 2014.
- [35] J. Donahue *et al.*, "Decaf: A deep convolutional activation feature for generic visual recognition," presented at the Proc. 31st Int. Conf. Int. Conf. Mach. Learn., Beijing, China, 2013.
- [36] J. Deng *et al.*, "ImageNet: A large-scale hierarchical image database," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Miami, FL, USA, 2009.
- [37] J. Kawahara *et al.*, "Deep features to classify skin lesions," presented at the IEEE 13th Int. Symp. Biomed. Imag., Prague, Czech Republic, 2016.
- [38] A. Menegola *et al.*, "Knowledge transfer for melanoma screening with deep learning," presented at the IEEE 14th Int. Symp. Biomed. Imag., Melbourne, Vic, Australia, 2017.
- [39] D. Yoo *et al.*, "Multi-scale pyramid pooling for deep convolutional representation," presented at the IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, Boston, MA, USA, 2015.
- [40] Y. Gong *et al.*, "Multi-scale orderless pooling of deep convolutional activation features," presented at the 13th Eur. Conf. Comput. Vis., Zurich, Switzerland, 2014.
- [41] J. Yue-Hei Ng *et al.*, "Exploiting local features from deep networks for image retrieval," presented at the IEEE Conf. Comput. Vis. Pattern Recognit. Workshop, Boston, MA, USA, 2015.
- [42] A. G. Howard, "Some improvements on deep convolutional neural network based image classification," arXiv: 1312.5402, 2013.
- [43] Z. Hang *et al.*, "Deep TEN: Texture encoding network," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Honolulu, HI, USA, 2017.
- [44] M. Cimpoi *et al.*, "Deep filter banks for texture recognition and segmentation," presented at the IEEE Conf. Comput. Vis. Pattern Recognit., Boston, MA, USA, 2015.
- [45] M. Gao *et al.*, "Multi-label deep regression and unordered pooling for holistic interstitial lung disease pattern detection," presented at the Int. Workshop Mach. Learn. Med. Imag., Athens, Greece, 2016.
- [46] Y. Song *et al.*, "Low dimensional representation of Fisher vectors for microscopy image classification," *IEEE Trans. Med. Imag.*, vol. 36, no. 8, pp. 1636–1649, Aug. 2017.
- [47] Z. Yu *et al.*, "Aggregating deep convolutional features for melanoma recognition in dermoscopy images," presented at the Int. Workshop Mach. Learn. Med. Imag., Quebec City, QC, Canada, 2017.
- [48] Z. Yu *et al.*, "Hybrid dermoscopy image classification framework based on deep convolutional neural network and Fisher vector," presented at the IEEE 14th Int. Symp. Biomed. Imag., Melbourne, Vic, Australia, 2017.
- [49] N. Situ *et al.*, "Malignant melanoma detection by Bag-of-Features classification," presented at the 30th Int. Conf. IEEE Eng. Med. Biol. Soc., Vancouver, BC, Canada, 2008.
- [50] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, pp. 91–110, 2004.
- [51] C. Barata *et al.*, "Two systems for the detection of melanomas in dermoscopy images using texture and color features," *IEEE Sys. J.*, vol. 8, no. 3, pp. 965–979, Sep. 2014.
- [52] S. Demyanov *et al.*, "Classification of dermoscopy patterns using deep convolutional neural networks," presented at the IEEE 13th Int. Symp. Biomed. Imag., Prague, Czech Republic, 2016.
- [53] A. Esteva *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.
- [54] N. Codella *et al.*, "Deep learning ensembles for melanoma recognition in dermoscopy images," *IBM J. Res. Develop.*, vol. 61, no. 4/5, pp. 5:1–5:15, Jul.–Sep. 2017.
- [55] J. Donahue *et al.*, "Caffe: Convolutional architecture for fast feature embedding," presented at the Proc. 22nd ACM Int. Conf. Multimedia, Orlando, FL, USA, 2014.
- [56] L. Zheng *et al.*, "Good practice in CNN feature transfer," arXiv: 1604.00133, 2016.
- [57] J. Yosinski *et al.*, "How transferable are features in deep neural networks?," presented at the Proc. 27th Int. Conf. Neural Inf. Process. Syst., Montreal, QC, Canada, 2014.
- [58] A. Mousavian and J. Kosecka, "Deep convolutional features for image based retrieval and scene categorization," arXiv: 1509.06033, 2015.
- [59] J. Long *et al.*, "Do convnets learn correspondence?" presented at the Proc. 27th Int. Conf. Neural Inf. Process. Syst., Montreal, QC, Canada, 2014.
- [60] B. Lei *et al.*, "Multi-modal and multi-layout discriminative learning for placental maturity staging," *Pattern Recognit.*, vol. 63, pp. 719–730, 2016.
- [61] M. Faraki *et al.*, "Fisher tensors for classifying human epithelial cells," *Pattern Recognit.*, vol. 47, pp. 2348–2359, 2014.
- [62] D. Reynolds, "Gaussian mixture models," in *Encyclopedia of Biometrics*. New York, NY, USA: Springer, 2009.
- [63] L. Liu *et al.*, "Encoding high dimensional local features by sparse coding based Fisher vectors," presented at the Proc. 27th Int. Conf. Neural Inf. Process. Syst., Montreal, QC, Canada, 2014.
- [64] A. Vedaldi and A. Zisserman, "Efficient additive kernels via explicit feature maps," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 480–492, Mar. 2012.
- [65] P. Li *et al.*, "Sign Cauchy projections and chi-square kernel," presented at the 26th Int. Conf. Neural Inf. Process. Syst., Lake Tahoe, NV, USA, 2013.
- [66] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional neural networks for Matlab," presented at the 23th Int. Conf. Multimedia, Brisbane, QLD, Australia, 2015.
- [67] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," presented at the Int. Conf. Multimedia, Firenze, Italy, 2010.
- [68] B. Lei *et al.*, "Discriminative learning for automatic staging of placental maturity via multi-layer Fisher vector," *Sci. Rep.*, vol. 5, 2015, Art. no. 12818.
- [69] T. Uricchio *et al.*, "Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging," presented at the IEEE Int. Conf. Comput. Vis. Workshop, Santiago, Chile, 2015.
- [70] L. Yu *et al.*, "Automated melanoma recognition in dermoscopy images via very deep residual networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 4, pp. 994–1004, Apr. 2017.
- [71] M. Everingham *et al.*, *PASCAL Vis. Object Classes Challenge*, 2012. [Online]. Available: <http://www.pascalnetwork.org/challenges/VOC/voc2012/workshop/index.html>