Short-term Lesion Change Detection for Melanoma Screening with Novel Siamese Neural Network

Boyan Zhang, Zhiyong Wang, *Member, IEEE,* Junbin Gao, Chantal Rutjes, Kaitlin Nufer, Dacheng Tao, *Fellow, IEEE,* David Dagan Feng, *Fellow, IEEE,* and Scott W. Menzies

***Abstract*— Short-term monitoring of lesion changes has been a widely accepted clinical guideline for melanoma screening. When there is a significant change of a melanocytic lesion at three months, the lesion will be excised to exclude melanoma. However, the decision on change or no-change heavily depends on the experience and bias of individual clinicians, which is subjective. For the first time, a novel deep learning based method is developed in this paper for automatically detecting short-term lesion changes in melanoma screening. The lesion change detection is formulated as a task measuring the similarity between two dermoscopy images taken for a lesion in a short time-frame, and a novel Siamese structure based deep network is proposed to produce the decision: changed (i.e. not similar) or unchanged (i.e. similar enough). Under the Siamese framework, a novel structure, namely Tensorial Regression Process, is proposed to extract the global features of lesion images, in addition to deep convolutional features. In order to mimic the decision-making process of clinicians who often focus more on regions with specific patterns when comparing a pair of lesion images, a segmentation loss (SegLoss) is further devised and incorporated into the proposed network as a regularization term. To evaluate the proposed method, an in-house dataset with 1,000 pairs of lesion images taken in a short time-frame at a clinical melanoma centre was established. Experimental results on this first-of-a-kind large dataset indicate that the proposed model is promising in detecting the short-term lesion change for objective melanoma screening.**

***Index Terms*— Melanoma screening, short-term lesion change detection, Siamese neural network, tensorial regression process, deep learning**

This work was supported in part by the Australian Research Council (ARC) grant #DP170104304.

B. Zhang, Z. Wang, D. Tao, and D. D. Feng are with the School of Computer Science, The University of Sydney, NSW 2006, Australia (e-mail: bzha8220@uni.sydney.edu.au;{zhiyong.wang, dacheng.tao, dagan.feng}@sydney.edu.au).

J. Gao is with Discipline of Business Analytics, The University of Sydney Business School, The University of Sydney, NSW 2006, Australia (e-mail: junbin.gao@sydney.edu.au).

C. Rutjes and K. Nufer are with Diamantina Institute, The University of Queensland, QLD 4102, Australia (e-mail: {c.rutjes, k.nufer}@uq.edu.au).

S. W. Menzies is with Discipline of Dermatology, Sydney Medical School, The University of Sydney, NSW 2006, Australia, and with Sydney Melanoma Diagnostic Centre, Royal Prince Alfred Hospital, Camperdown, NSW 2050, Australia (e-mail: scott.menzies@sydney.edu.au).

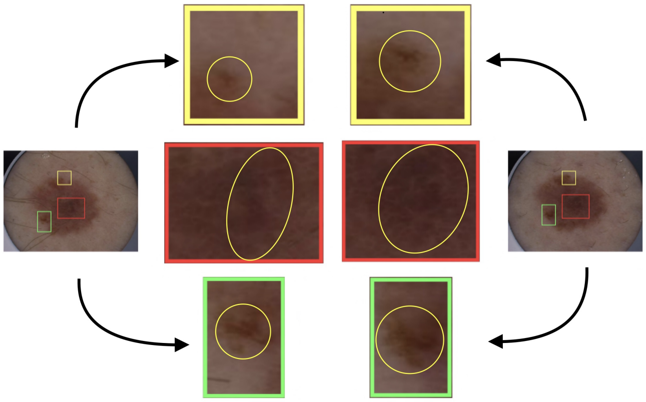


Fig. 1: Illustration of the subtle changes between a pair of lesion images taken in a short time-frame. The three changed regions are highlighted in different colours in the image pair with manually segmented masks.

I. Introduction

MELANOMA is one of the most malignant types of skin cancer that develops from melanocytes, the cells making the brown pigment that gives skin its color [1]. According to Australian Institute of Health and Welfare (AIHW) [2], about 15,229 new melanomas were diagnosed in 2019, with a mortality of 1,725 persons. Early detection of melanoma is critical to the survival of melanoma patients. Among all the people with melanoma of the skin, from the time of initial diagnosis, the 5-year survival is 91%. In contrast, if melanoma has spread to other distant parts of the body, the survival rate decreases significantly to about 23% [3]. Therefore, timely melanoma screening is essential for identifying suspicious lesions at an early stage.

It is well recognised that the clinical and pathological diagnosis of pigmented lesions can be extremely difficult. Misdiagnosis of melanoma can have tragic results for individuals, and overdiagnosis can result in excessive direct and indirect health-care costs [4]. Dermoscopy is a technique that uses a hand-held magnifying device combined with either the application of a liquid between the transparent plate of the device and the skin, or the use of cross-polarized light. This allows the visualization of diagnostic features of skin lesions

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2020.3037761, IEEE Transactions on Medical Imaging

that are not seen with the naked eye [5]. Sequential digital dermoscopy imaging or dermoscopy monitoring involves the capture and assessment of successive dermoscopic images, separated by an interval of time, of one or many melanocytic lesions to detect suspicious changes. With short-term monitoring, any morphological change over a 3-month interval leads to excision of the lesion so as to exclude melanoma [6]. When monitoring clinically atypical melanocytic lesions and detecting any change, the sensitivity for the diagnosis of melanoma is 94% (all melanoma sub-types excluding lentigo maligna) and the specificity 84%. That is, 94% of melanoma will change at 3 months but only 16% of atypical benign moles [6]. Furthermore, with changing lesions in short-term monitoring, there are no specific types of morphological change that predict a lesion is melanoma or benign [7]. Hence all changed lesions require excision, with the majority of melanomas detected having no specific diagnostic features of melanoma (the concept of "featureless" melanoma) [8]. This technique has been shown to increase the sensitivity (detection) for the diagnosis of melanoma while reducing unnecessary excisions of benign nevi (moles) in both specialist and primary care physicians [8]. However, the manual screening process heavily relies on the naked eyes and experiences of a clinician, which is clearly subjective. As shown in Fig. 1, the changes between a pair of lesion images are very subtle, which also makes short-term lesion change detection very challenging.

In recent years, due to the groundbreaking success of deep learning techniques, there have been significant progresses [9]–[12] on computer aided melanoma detection which aims to classify a lesion as melanoma or not. For example, as reported in [13], deep learning algorithms were able to achieve dermatologist level performance on melanoma detection. In the ISIC 2018 challenge [14], a detection accuracy 96.80% was reported. However, few studies using deep learning techniques have been conducted to provide computer aided solutions for short-term melanoma screening, except several studies using traditional keypoint matching based methods [15].

Therefore, in this paper, we for the first time propose a novel deep learning based method for automatically detecting short-term dermoscopy change. We formulate lesion change detection as a task of measuring the similarity between two dermoscopy images taken for a lesion in a short time-frame and utilize a Siamese network to learn the decisions of change or no-change. Under the Siamese framework, we propose a novel structure, namely Tensorial Regression Process, to extract the global features of lesion images, while also utilizing deep convolutional features which focus more on local patterns of lesion images. That is, our model contains conventional convolutional layers for extracting local features and new Tensorial Regression Layers (TRL) to extract global features. In order to mimic the decision-making process of clinicians who often focus more on regions with specific patterns when comparing a pair of lesion images, a segmentation loss function, SegLoss, is further devised and incorporated into our proposed network as a regularization term. Our extensive experiments also demonstrate that segmentation in the intermediate feature maps will improve the performance of lesion change detection.

In summary, the key contributions of our work are as follows:

1) We conduct the first deep learning based study of short-term melanoma change detection. In particular, we formulate lesion change detection as a task measuring the similarity between two sequential dermoscopy images, and propose a novel Siamese structure based deep network to learn the decisions of change (i.e. not similar) or no-change (i.e. similar enough).
2) We also propose a Tensorial Regression Process in order to extract the global features of lesion images, while conventional convolutional neural networks focus more on learning local features.
3) We further devise a segmentation loss function, SegLoss, as a regularization term so that the decision making process could be more sensitive to the regions with specific patterns.
4) We construct the first-of-a-kind large dataset for studying short-term melanoma screening with computer aided change detection. Our experimental results indicate that it is promising to develop deep learning algorithms for automatic short-term lesion change detection.

The remaining part of this paper is organized as follows. In Section II, we review the related works of melanoma detection, change detection in remote sensing, and near-duplicate image detection. In Section III, we describe the details of our proposed model, including the overall network architecture and individual components. In Section IV, we provide the experimental results on the first-of-a-kind melanoma dataset we purposely built, as well as the qualitative analysis of TRL and SegLoss. At last, we conclude our study in Section V.

II. Related Work

In this section, we first review the studies on melanoma detection which is also used for melanoma screening by classifying a given dermoscopy image as Malignant or Benign. Then we review two categories of relevant studies outside of the medical domain: change detection in remote sensing and near-duplicate image detection, which are both relevant to our study in principle. The former category is for observing environmental changes through aerial images (e.g., changes over different seasons [16], [17]), while the latter category is for identifying image pairs which are nearly identical, and is different with general image matching [18], [19].

A. Melanoma Detection

Early studies on melanoma detection mainly focus on designing morphological features by taking the clinical expertise into account. For example, Nachbar *et al.* [20] proposed the ABCD rule of dermoscopy criteria (i.e., Asymmetry, Border irregularity, Color patterns, and Diameter) and Argenziano *et al.* [21] proposed the seven-point checklist (i.e., atypical pigment network, gray-blue areas, atypical vascular pattern, and etc). Based on these clinical rules, many computer aided melanoma detection methods have been proposed. In [22], Jain *et al.* proposed to use various image processing techniques to segment target lesions and measure the ABCD parameters of

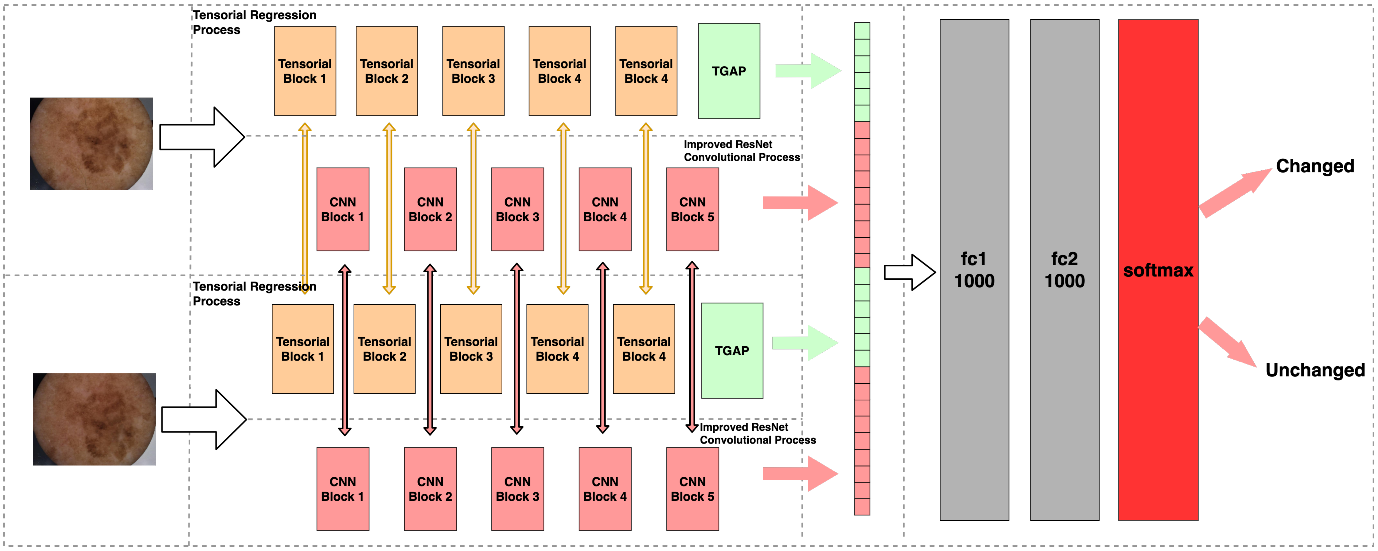


Fig. 2: Illustration of our proposed Siamese network. The network takes as input one pair of lesion images captured a short time-frame, and extracts global and local features with the tensorial regression process (Tensorial Block) and the improved ResNet convolutional processs (CNN Block), respectively. Finally, learned features will be concatenated for the final classification through two fully connected layers and one softmax layer. Note that the vertical arrows across different computation channels denote weight sharing.

those lesions for deciding whether the lesions are Melanoma or Benign.

However, the features inspired by clinical rules are sensitive to image processing techniques. In [23], a set of high-level intuitive features (HLIF) was proposed to characterize melanoma lesions in images captured by conventional cameras. With the advancements of various image feature extraction techniques, many methods have also been proposed to characterize different dermoscopic attributes, such as pigment networks [24], blue-white areas [25], and streaks [26]. However, it is very challenging for these hand-crafted features to cope with different image processing techniques and extensive variations of melanoma lesions from patient to patient. As a result, various classification methods have also been proposed for melanoma detection. Yao *et al.* [27] proposed a novel framework to extract the sparse representation of the interested points for melanoma detection. Xie *et al.* [28] also proposed an ensemble neural network structure for melanoma detection. However, both of them are limited to represent the global image structure, which further results in the limited performance.

Due to the great success of deep learning techniques such as Convolutional Neural Networks (CNNs) on various vision tasks such as image classification in recent years, many deep learning based melanoma detection methods have been proposed. In 2015, Cai *et al.* [29] for the first time proposed a framework to integrate CNN based feature learning, sparse coding and SVM, offering quite reasonable performance for melanoma recognition task. Md. Zahangir *et al.* [30] obtained a U-Net-based structure to extract the local features of the lesion regions of interest, and achieved improved performance on the melanoma detection task. In [31], Gulshan *et al.* proposed a deep convolution neural network for diabetic retinopathy in retinal fundus photographs. Similarly, Esteva *et al.* [13] demonstrated that fine tuning of the pre-trained version of Google's Inception v3 architecture with 132,824 lesion images (including both 129,450 dermoscopy images and 3,374 conventional lesion images) was able to achieve (near) human expert performance. In 2019, Zhang *et al.* [9] incorporated the attention module into their system and proposed an attention residual learning convolutional neural network (ARL-CNN) model for skin lesion classification. To facilitate the research on melanoma detection, the ISIC melanoma detection challenge (https://challenge.isic-archive.com/) was established in 2016 with datasets publicly available to the research community.

B. Change Detection in Remote Sensing

Traditionally, there are two steps involved in detecting changes in remote sensing images: 1) obtaining a similarity-feature map (e.g., difference image) between a pair of images with various arithmetical operations (e.g., differencing and rotation), and 2) labeling the pixels of the similarity-feature map as changed or unchanged. Many change detection algorithms have been proposed by solving the pixel labeling task from different perspectives. For example, simple thresholding can be utilized to identify changed pixels or regions, which lacks robustness, where a change is detected if a difference value exceeds pre-defined thresholds [32]. Other advanced classification based methods formulate change detection as a classification task which assigns each pixel into one of the two classes: changed or unchanged, such as extreme learning machine (ELM) [33], principal component analysis [34], and Markov random field [35].

Recently, deep learning techniques have also been utilized for change detection in the remote sensing field. In 2017, Zhan *et al.* [36] proposed a Siamese convolutional network using the weighted contrastive loss, which enables the Siamese network to extract features directly from the image pairs. Zhao *et al.* [37] proposed a Siamese neural network model to analyze multi-temporal images of the same scene for identifying the changes that have occurred by combining unsupervised feature learning and supervised fine-tuning. Though this research achieved a good performance on several remote sensing datasets, the learned "high-level abstractions" did not obtain the task-specific features (i.e. border and texture) in remote sensing images, resulting in a bottleneck of the change detection performance. Similarly, Yang *et al.* [38] proposed a supervised deep learning framework for the remote sensing image change detection task, with a pre-training step to transfer a difference image to the target label domain. However, the proposed model is still a conventional end-to-end model without utilizing any domain knowledge of remote sensing, which also results in a bottleneck of the change detection performance. Though recent works utilized object detection [39], saliency detection [40] and Spectral-Spatial feature fusion [41] in their deep frameworks for various remote sensing image analysis tasks, none of them were designed for detecting the changes within a lesion image pair.

Note that similar techniques were also explored for detecting changes in MR images [18], [42]. However, these methods cannot be directly utilized for detecting changes of skin lesions, as aerial image pairs can be well aligned using geographical information for producing high quality similarity feature maps, while lesion images taken at different times will have different rotation and distortion caused by uncertain imaging conditions. Recently, the conventional Siamese neural network was utilized to investigate disease severity evaluation and change detection in two medical imaging domains: retinopathy of prematurity (ROP) in retinal photographs and osteoarthritis (OA) in knee radiographs [43].

C. Near Duplicate Image Detection

Near duplicate images generally refer to different versions of an image which are obtained by undertaking various editing steps such as color mapping, scaling, translation, and rotation on the image. Near duplicate image detection aims to identify near duplicate image pairs. Due to the highly similar visual content between near duplicate image pairs, local descriptors such as SIFT (Scale Invariant Feature Transform) features have been widely used for accurately matching two images.

Local descriptors are often prone to false positive matches, as they do not take into account spatial coherence. In order to reduce false alarms, various pruning techniques were proposed to improve specificity and scalability [44], whereas other methods focused on post-query verification [45]. Connor *et al.* proposed a method to evaluate the specificity of near duplicate detectors and chose the optimal distance threshold, based on Receiver Operating Curve (ROC) analysis [46].

For the second strategy, geometric verification has been widely used to filter false matches. Local matches are first obtained between images, and the geometric consistency among the matches is then utilized to filter false matches that are geometrically inconsistent. In [47], Jegou *et al.* proposed a weak geometric consistency (WGC) scheme by verifying the consistency of the angle and scale parameters for matched features. With an additional assumption that true matches also follow consistent translation transformation, Zhao *et al.* [48] improved WGC by adding translation information. To fully capture geometric relationships of local features, a global geometric consistency verification strategy, i.e., RANSAC [49], [50], is very popular for this task. It randomly samples several pairs of local matches many times to estimate the affine transformations between images, and then verifies the geometric consistency of the local matches to filter out inconsistent ones. However, RANSAC can only be applied to a small number of top-ranked candidate images due to its high computational complexity. Zhou *et al.* [51] proposed a novel geometric coding scheme, which describes the geometric relationships among SIFT features in three geo-maps for the verification of SIFT matches. The proposed model can efficiently filter the false matches that are geometrically inconsistent under rotation and scaling transformations, partial-occlusion, and background clutter. However, many false matches between similar images may satisfy geometric consistency, thus cannot be removed effectively with these strategies.

Therefore, in this paper, we propose a novel Siamese network to address the challenge of detecting short-term lesion change. In order to extract high-level features, our proposed network utilizes a ResNet CNN structure for learning local features and a newly proposed Tensorial Regression Process (TRL) for learning global features [52]. The Spatial Transformer Network (STN) is utilized as a sub-layer in the framework to deal with the rotation between a pair of lesion images. In order to impose more emphasis on representative features of lesion regions in the decision-make process of melanoma screening, such as lesion shape and texture of regions of interest, we further devise a segmentation loss function, SegLoss, as a regularization term.

III. Proposed Method

As shown in Fig. 2, the overall architecture of our proposed network mainly consists of two feature learning streams: the Improved ResNet Convolutional Process for local feature extraction and the Tensorial Regression Process for global feature extraction. Weight sharing has been applied to different image channels for extracting image features. The input to the network is a pair of lesion images taken at different times. At the end two fully connected layers taking the concatenated global and local features as input and a softmax regression layer are utilized to produce classification results, changed or unchanged. In this section, we will first introduce our proposed Tensorial Regression Layer (TRL), including the novel Tensorial Global Average Pooling (TGAP) strategy, then discuss the Spatial Transformer Network (STN) which improves ResNet, and finally explain our optimization function.

A. Tensorial Regression Layer (TRL)

In general, tensors could be regarded as structural objects that encode useful linear correlation among different modes. That is, a D-way or D^{th}-order tensor is an element of the tensor product of D vector spaces, each of which has its own coordinate system [53]. In such a representation, image data and the hidden layer outputs of a deep neural network can be viewed as tensors.

The l-th layer of conventional CNNs performs a non-linear projection on its input X^l:

$$X^{l+1} = \sigma^{(l)}(W^l X^l + b^l), \tag{1}$$

where $\sigma^{(l)}$ is a non-linear activation function and b is the offset of the current layer. However, such a non-linear mapping only assumes a vectorial input, which means X^l and X^{l+1} are vectors and W is a linear mapping matrix in appropriate size. This could be problematic as the spatial information of multi-dimensional data (e.g., images containing different color channels) will not be well exploited during vectorization. As a result, tensorial representation has been proposed as a versatile method to preserve the data correlation along different dimensions so that global features could be better extracted.

In our previous work, we proposed MatNN [54] which assumes matrix (2nd-order) data as inputs. As a tensor is essentially an organized multi-dimensional array of numerical values and a D^{th}-order tensor could be regarded as a D-way array, in this paper, we consider Tucker Decomposition [53] and further propose a Tensorial Regression Layer (TRL) for general multi-dimensional data. A simple illustration of our TRL is shown in Fig. 3.

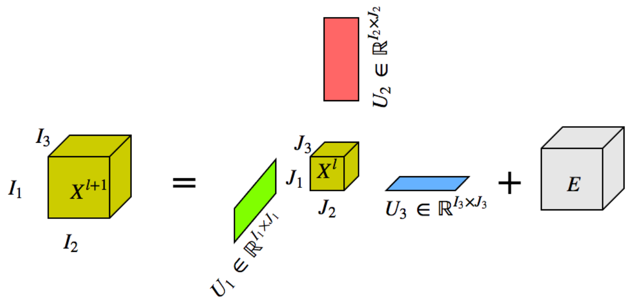

$U_2 \in \mathbb{R}^{I_2 \times J_2}$
I_3
I_1
X^{l+1}
I_2
=
J_3
J_1
X^l
J_2
$U_1 \in \mathbb{R}^{I_1 \times J_1}$
$U_3 \in \mathbb{R}^{I_3 \times J_3}$
+
E

Fig. 3: Illustration of our proposed Tensorial Regression Process, where the relationship between X^{l+1} and X^l can be viewed as a Tucker Decomposition with residual E. Similar to convolutional neural networks, each Tensorial Block built on such tensorial process contains one batch normalization (BN) layer, one Tensorial Regression Layer (TRL), and one Rectified Linear Unit (RELU) activation layer.

A D-order tensor X is an element of the tensor product of D vector spaces with D coordinates indexed by $(i_1, i_2, ..., i_D)$:

$$X = (x_{i_1,i_2,...,i_D})_{1 \le i_1 \le I_1, 1 \le i_2 \le I_2, ..., 1 \le i_D \le I_D}, \tag{2}$$

where D is called the order/dimensionality and $I_1, I_2, ..., I_D$ are the corresponding dimensions along each of the tensor's D modes. Hence, all the $I_1 \times I_2 \times \cdots \times I_D$ elements of the tensor X are arranged in a high-dimensional rectangular structure.

The d-mode product of a D-order tensor X with a matrix $U^{(d)} = (u_{j_d,i_d}) \in \mathbb{R}^{J_d \times I_d}$ is denoted by $X \times_d U^{(d)}$. The result is a D-order tensor of dimension $I_1 \times \cdots \times I_{d-1} \times J_d \times I_{d+1} \times \cdots \times I_D$. Elementwise, the d-mode product can be expressed as

$$\begin{aligned}&(X \times_d U^{(d)})_{i_1,\cdots,i_{d-1},j_d,i_{d+1},\cdots,i_D} \\ &= \textstyle\sum_{i_d=1}^{I_d} x_{i_1,\cdots,i_{d-1},i_d,i_{d+1},\cdots,i_D} u_{j_d,i_d}.\end{aligned} \tag{3}$$

The basic model of a TRL illustrated in Fig. 3 is described as the following mapping:

$$\begin{aligned}X^{l+1} &= X^l \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)} ... \times_D U^{(D)} + B^l \\ &= \left[\left[X^l; U^{(1)}, U^{(2)}, U^{(3)}, ..., U^{(D)}, B^l\right]\right],\end{aligned} \tag{4}$$

where $U^{(i)} \in \mathbb{R}^{J_i \times I_i}$ represents the i-th way projection and B^l is the intercept term. It is not difficult to write the mapping formula at the elementwise level according to the definition of tensor product with a mode matrix.

In this paper, we treat one lesion image as a 3-way tensor. That is, each input data contains both spatial (mode-1 and mode-2) and color channel (mode-3). Therefore the corresponding mapping function within one hidden layer can be written as:

$$\begin{aligned}X^{l+1} &= X^l \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)} + B^l \\ &= \left[\left[X^l; U^{(1)}, U^{(2)}, U^{(3)}, B^l\right]\right],\end{aligned} \tag{5}$$

or, element-wise, as:

$$x^{l+1}_{i_1,i_2,i_3} = \sum_{j_1=1}^{J_1} \sum_{j_2=1}^{J_2} \sum_{j_3=1}^{J_3} (x^l_{j_1,j_2,j_3} u^{(1)}_{i_1 j_1} u^{(2)}_{i_2 j_2} u^{(3)}_{i_3 j_3} + b^l_{i_1,i_2,i_3}), \tag{6}$$

where J_1, J_2, and J_3 are the dimensions for mode-1, mode-2 and mode-3, respectively, and $b^l \in \mathbb{R}^{i_1 \times i_2 \times i_3}$ is the elementwise intercept term. Therefore, for each Tensorial Regression Layer, we treat the input and output as a tensorial structure without discarding the spatial context by eliminating the vectorization step in conventional CNNs. Therefore the learned representation through the TRL process could be viewed as a global feature.

B. Tensorial Global Average Pooling (TGAP)

In conventional neural networks, data are first vectorized before being fed into the connected layers, which will destroy its representation continuity. In addition, a well-trained fully connected layer always requires numerous parameters, and thus may lead to overfitting. Therefore, in order to have a continuous representation as well as to reduce the feature size, we propose a novel Tensorial Global Average Pooling (TGAP) step.

Each frontal slice of a 3-way tensor can be regarded as one single view of the tensorial data. Instead of using a single vectorization on top of the tensorial feature maps, we utilize global average pooling for each slice of different modes (i.e., horizontal mode, vertical mode, and frontal mode), as shown in Fig. 4. The averaged confidence output of each slice can be easily interpreted as the tensor's representation.

Mathematically, given a tensorial data as an input, the horizontal confidence H_y, vertical confidence V_x and frontal confidence F_z could be calculated as follows:

$$H_y = \frac{1}{h \times j} \sum_{x=1}^{h} \sum_{z=1}^{j} X_{x,y,z}, \tag{7}$$

$$V_x = \frac{1}{i \times j} \sum_{y=1}^{i} \sum_{z=1}^{j} X_{x,y,z}, \tag{8}$$

$$F_z = \frac{1}{h \times i} \sum_{x=1}^{h} \sum_{y=1}^{i} X_{x,y,z}, \tag{9}$$

where $X \in \mathbb{R}^{h \times i \times j}$.

The advantage of TGAP is three-fold. First, the averaged confidence of each slice has physical meaning and therefore the output of TGAP can be easily interpreted as the tensor's global representation. In our case, the averaged confidence of a horizontal or vertical slice can be regraded as the spatial changes of a lesion's spatial representation, e.g. location and shape, whereas the averaged confidence of a frontal slice can be viewed as the different activation result along the different channel. Second, there are no parameters to be optimised in our TGAP layer, which helps reduce the computational complexity. Third, TPAG significantly reduces the feature dimension from $h \times i \times j$ to only $h + i + j$.

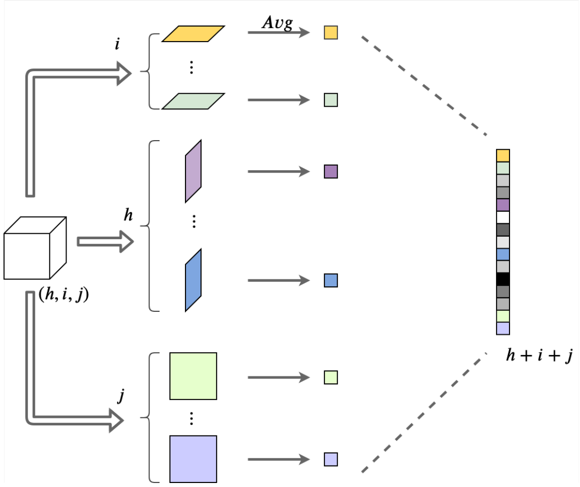


Fig. 4: Illustration of Tensorial Global Average Pooling where each horizontal, vertical and frontal slice is regarded as one standard feature map, because each of them represents specific spatial information of a given tensorial output.

C. Spatial Transformer Network

In order to overcome the deformation issue caused by the acquisition of input image pairs (e.g., rotation), we utilize the Spatial Transformer Network [55] module as a sub-set of our CNN structure for a robust local representation. This module could be viewed as a generalisation of differentiable attention to any spatial transformation. Though the original implementation introduces many geometrical transformation, such as scaling, translation, and rotation, in this paper, we utilize the affine transform for this module.

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}, \tag{10}$$

where (x_i^s, y_i^s) and (x_i^t, y_i^t) represent the source and target coordinates of the regular grid in the output feature map respectively, and A_θ is the affine transform matrix. Note that the affine transform matrix A_θ is trainable via backward propagation. By applying such an affine transform, the local region of interest can be extracted. The detailed structure of our STN is available in Table V. We did not introduce STN into our TRL process as such an affine transform is used to locate the salient sub-regions which may further destroy the representation continuity.

D. Loss Function

We utilize the Softmax regression for classification. Assume that we are given a training dataset $D = \left\{(X_i^1, X_i^2, t_i)\right\}_{i=1}^{N}$ where X_i^1 and X_i^2 represents the samples of each pair for Channel 1 (Tensorial Regression) and Channel 2 (ResNet), t_i is the target. The cross entropy loss function is designed as:

$$L = -\frac{1}{N} \sum_{i=1}^{N} t_i \log(o_i) + (1 - t_i) \log(1 - o_i) + Re, \tag{11}$$

$$o_i = \frac{\exp(f_k)}{\exp(f_0) + \exp(f_1)}, \tag{12}$$

where k equals to either 0 or 1 and o_i is softmax output and f is the output of the last fully connected layer. In (11), Re represents all the regularization factors and will be discussed in the following sub-sections.

1) Regularization: Norm-2 regularization has been applied upon the loss function so that the proposed model could clamp the size of the weights on the connections. In addition, in order to obtain a sparse model, a penalty term R is utilized to penalize the weights. The SegLoss S which is designed for extracting the region specific patterns is also introduced as a factor, which will be described in details in Section III-D.3:

$$Re = \lambda_1 \sum_l {W^{(l)}}^2 + \lambda_2 \sum_l {U_{1,2,3}^{(l)}}^2 + \beta \sum_l R^{(l)} + \gamma \sum_l S^{(l)}, \tag{13}$$

where W includes all the parameters in the CNN network structure, U is the weight matrix in our TRL.

2) Sparsity: We also refer to the idea in [54] to design the sparsity term R. This factor is obtained through Kullback-Leibler divergence by penalizing the average activations $\overline{\rho}^{(l)}$ at the l-th layer which deviates significantly from $\rho^{(l)}$:

$$R^{(l)} = \text{sum}(\rho \log \frac{\rho}{\overline{\rho}^{(l)}} + (1 - \rho) \log \frac{1 - \rho}{1 - \overline{\rho}^{(l)}}). \tag{14}$$

Note that in (14), ρ is called the sparsity parameter, typically a small value close to zero, and $\overline{\rho}^{(l)}$ is the average activation defined as:

$$\overline{\rho}^{(l)} = \frac{1}{M \times N} \sum_x^M \sum_y^N f^{(l)}, \tag{15}$$

where f is the feature maps with size $M \times N$ in our TRL only.

3) SegLoss: In order to mimic the decision-making process of dermatologists who often focus more on regions with specific patterns, a segmentation mechanism is also incorporated into our proposed network as a regularization term. We treat each slice of the hidden layer as a standard feature map. We utilize the Otsu segmentation algorithm [56] which maximizes the inter-class difference to achieve the segmentation within an image, then maximize the variance of the average gray level of foreground (i.e., Region of Interest) and background so that each feature map within the model could be segmented.

As defined in (16), the SegLoss factor S is equivalent to the inverse of the inter-class variance:

$$S^{(l)} = \frac{1}{\delta^2}, \tag{16}$$

$$\delta^2 = p_1(I_1 - I_g)^2 + p_2(I_2 - I_g)^2 + ... + p_n(I_n - I_g)^2, \tag{17}$$

$$I_g = p_1 I_1 + p_2 I_2 + ... + p_n I_n \ , \tag{18}$$

where p_i in (17) is the pixel proportion of the corresponding "mask" class, I_n and I_g are the average gray levels of the foreground and the background, respectively, and n is the number of main objects that the foreground could be segmented into.

When calculating $I_n^{(l)}$ and $I_g^{(l)}$ at the l-th layer, we firstly normalize the feature maps into $[-1, 1]$, and then decompose the normalized feature maps into $n+1$ channels (including the background). This is to accelerate the calculation of S and simplify the gradient calculation. The down-sampled I does not influence the calculation of lateral hidden layers. Note that this SegLoss factor is applied to TRL only.

We rewrite (16) with $n = 2$ here:

$$S^{(l)} = \left(p_1^{(l)}(I_1^{(l)} - I_g^{(l)}))^2 + p_2^{(l)}(I_2^{(l)} - I_g^{(l)}))^2\right)^{-2}. \tag{19}$$

Given $p_2^{(l)} = 1 - p_1^{(l)}$ for this case, we have:

$$S^{(l)} = \left(p_1^{(l)}(I_1^{(l)} - I_g^{(l)})^2 + (1 - p_1^{(l)})(I_2^{(l)} - I_g^{(l)})^2\right)^{-2}, \tag{20}$$

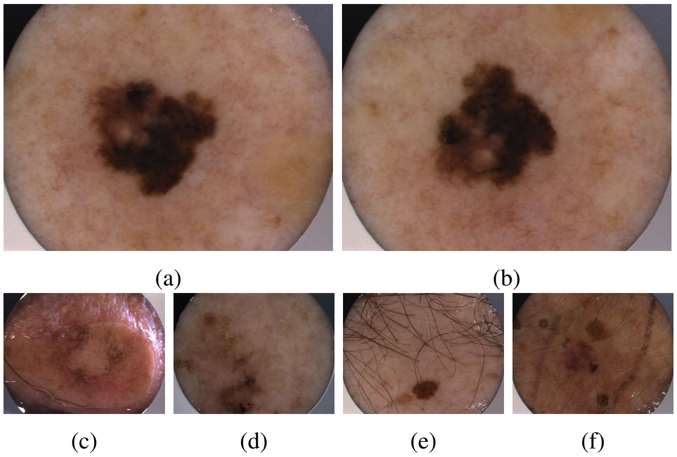
$$\begin{aligned} S^{(l)} = \Big(& p_1^{(l)}({I_1^{(l)}}^2 - 2I_1^{(l)} I_g^{(l)} + {I_g^{(l)}}^2) \\ &+ (1 - p_1^{(l)})({I_2^{(l)}}^2 - 2I_2^{(l)} I_g^{(l)} + {I_g^{(l)}}^2)\Big)^{-2}, \end{aligned} \tag{21}$$

$$\begin{aligned} S^{(l)} =& (p_1^{(l)} {I_1^{(l)}}^2 - 2p_1^{(l)} I_1^{(l)} I_g^{(l)} + (1 - p_1^{(l)}) {I_2^{(l)}}^2 \\ &- 2I_g^{(l)}(1 - p_1^{(l)}) I_2^{(l)} + {I_g^{(l)}}^2 + 2p_1^{(l)} I_2^{(l)} I_g^{(l)})^{-2}. \end{aligned} \tag{22}$$

Therefore, we have:

$$\frac{\partial S^{(l)}}{\partial p_1} = \delta^{-3}({I_1^{(l)}}^2 - 2I_1^{(l)} I_g^{(l)} - {I_2^{(l)}}^2 + 2I_2^{(l)} I_g^{(l)} + 2I_2^{(l)} I_g^{(l)}), \tag{23}$$

subject to $I_1^{(l)} = V_{x,label=1}$ and $I_2^{(l)} = V_{x,label=2}$.


(a) (b)

(c) (d) (e) (f)

Fig. 5: Sample lesion images in the dataset. (a) and (b) are a pair of lesion images used in our experiments. Lesion images were excluded from our experiments due to noises caused by various factors, such as (c) air bubble, (d) incomplete capturing of the lesion, (e) dense hair coverage, and (f) unexpected ink.

IV. Experimental Results and Discussions

A. Dataset

To conduct experiments with the proposed method, we established a melanocytic lesion dataset with Human Ethical Approval HREC/16/RPAH/496 at Sydney Local Health District, NSW, Australia. This retrospective dataset collected from the Sydney Melanoma Diagnostic Centre (SMDC), Royal Prince Alfred Hospital (RPAH) consists of approximately 100,000 dermoscopy skin lesion images captured with a dedicated video instrument at a resolution of 760 × 570 (SolarScan, Polartechnics Ltd, Sydney, Australia).

The only measurement made was the binary decision: *changed* versus *unchanged*. The ground truth was assessed from visual observation by Boyan Zhang after consultation with Scott W. Menzies (the first describer of the technique of short-term digital dermoscopy monitoring [57]) and using published tutorials [58]. A random selection of paired images were assessed with an inter-observer agreement of ground truth between Boyan Zhang and Scott W. Menzies greater than 92% in terms of the percentage of identical assessments.

As some dermoscopy images were contaminated with noises caused by various factors such as illumination change and occlusion with bubbles or hair shown in Fig. 5, we manually chose 1,000 useful pairs with lesions clearly presented and a time interval between 3 months and 6 months (short-term monitoring). As the first study in the field of detecting change in dermoscopy images, uncommonly detected artefacts were considered best removed from our set (as also seen in the ISIC 2018 study). However, these artefacts need to be included in future studies to ensure that such confounders do not effect accuracy.

In order to increase the number of training samples, we augmented the dataset with flipping and cropping transformations. Finally, we obtained 9,500 pairs of augmented short-term dermscopy images in total (3,435 pairs changed and 6,065 pairs unchanged). All the dermoscopy images were resized

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2020.3037761, IEEE Transactions on Medical Imaging

into a resolution of 224×224.

B. Experimental Settings

We utilized the ResNet-50 pre-trained with ImageNet as the base model for the convolutional process. The weights in the convolutional process and the TRL processing were shared for both input channels. All models were trained with the Adam optimizer and a total of 8 images (4 pairs) per mini-batch over 2 NVIDIA GTX 1080 GPUs. The learning rate α was pre-set as 0.0001. We trained our entire dataset over 400 epochs. The sparsity parameter ρ in our regularization factor R was set to 0.05 in our experiments. The ratio between training and test sets was set to 8:2, the experiments were repeated with 5 replicates, and the hyperparameters were determined on one replicate shown in Table IV.

C. Evaluation Metrics

In our study, four popular evaluation metrics: Accuracy, Sensitivity, Specificity, and AUC (Area Under Curve) are used. Accuracy measures the overall correctness of the decisions produced by our algorithm, Sensitivity measures the correctness of our algorithm over changed lesions, and Specificity measures the correctness over unchanged lesions. As the consequence of missing a changed lesion is fatal, for a change detection algorithm, in general the higher sensitivity the better. AUC aims to provide an overall evaluation of a classifier across various sensitivity and specificity values. The greater an AUC value is, the better a classifier is.

D. Performance Analysis

We first compared our proposed model with two baseline models: SIFT method [59] and conventional ResNet-50 model [60]. Due to the significant advantages such as robustness against rotation and scale, Scale-invariant feature transform (SIFT) has been widely used for various image matching tasks such as change detection [18], [59] and near duplicate image detection [61]. When the keypoint matching based similarity between a lesion pair exceeds a given threshold, it is concluded that the keypoints of the lesion pair match well and the lesion pair is unchanged; otherwise, the pair will be regarded as changed. The similarity measurement of a given lesion pair is defined as:

$$Sim = \frac{|k_1 \bigcap k_2|}{|k_1 \bigcup k_2|}, \tag{24}$$

where k_i is the set of keypoints in the i-th image of the pair.

We used the ResNet-50 model pre-trained on ImageNet and fine-tuned the parameters on our melanoma dataset. Similar with our proposed Siamese network structure, 2 lesion images were fed into the network to extract local features. Then we concatenated the features together as one vector and used 2 fully connected layers and one softmax regression layer to generate the decision (i.e., changed or unchanged).

As shown in Table I, when STNs are utilized in the convolutional blocks 1, 2 and 3, the proposed model achieves 72.0% accuracy and 74.1% AUC with a sensitivity of 84.1% and a specificity of 65.2%, whereas the accuracy and AUC values of the conventional SIFT model and ResNet-50 model are both around 60%, which is just slightly better than random guessing. Similarly, when STNs are utilized in the convolutional blocks 1, 2, 3, 4 and 5, the proposed model achieves 74.1% accuracy and 74.8% AUC with a sensitivity of 87.1% and a specificity of 66.8%, which also significantly outperformed the baseline models. Note that it is challenging to choose an optimal similarity threshold for SIFT matching, we empirically chose a value to produced the best sensitivity on the dataset showed in the table.

As shown in the Table I, the proposed TRL is very helpful and can lead to a huge performance boost, about 7% accuracy over original ResNet-50 model. Furthermore, STN is helpful. This is because that STN utilizes the affine transformation within the hidden layers so that the model could overcome the problems caused by different viewing angles during the processes acquiring dermoscopy images. Finally, introducing different regularization factors further improves the performance, such as 2.9% on accuracy and 2.7% on AUC improvement when STNs are utilized in convolutional blocks 1, 2, and 3. Similar improvements are also observed when STNs are utilized in convolutional blocks 1, 2, 3, 4, and 5. Further evaluations on the settings of different modules in our proposed framework will be discussed in Sec IV-F, Sec IV-G and Sec IV-H.

In addition, we compared our proposed model with two state-of-art models in both image change detection and near duplicate detection fields: Deep Siamese Convolutional Network (DSCN) [36] and Overlapping Region-based Global Context Descriptor (OR-GCD) [45]. Since DSCN is designed for dealing with labeled images, which is different from our task, we used DSCN pre-trained model for our image pairs (we do not have labeled lesion images which are required for training their network), and then thresholded the generated maps to generate the final label (changed vs not changed). For OR-GCD, we applied their model with the reported best hyper-parameter $M = 4$ and $N = 8$ to our experiments. As shown in Table I, our proposed model clearly outperforms these two models in terms of both Accuracy and Sensitivity. DSCN and OR-GCN achieved 70.4% and 69.0% for sensitivity, which is trailing almost 17% with our best proposed model. This may be caused by the lack of accurate image descriptor extraction in their proposed model. For example, it is very challenging, if not impossible, to train DSCN on our dataset, and OR-GCD only utilized low-level visual descriptors.

We further evaluated the result of our proposed method with independent melanographers. 204 original lesion pairs (i.e. non-augmented pairs) of one testing split were scored independently (unchanged, changed, uncertain), without knowledge of the proposed method, by two experienced trained melanographers (Observer 1 and Observer 2) at one institution (The University of Queensland Diamantina Institute) who have previously assessed more than 5000 cases of digital dermoscopy monitored lesions. The agreements between ground truth, our proposed method (Algorithm) and the two independent observers were assessed using Cohen's Kappa with uncertain cases removed from the analysis. The sensitivity and specificity of Algorithm to detect change was calculated

TABLE I: Experimental results on different models. The experiments are conducted on the same set for 10 rounds.

Model	Acc	Sens	Spec	AUC
SIFT	53.1%	76.4%	39.8%	56.1%
Conv	58.3%	69.9%	51.7%	61.9%
DSCN	67.1%	70.4%	65.2%	69.3%
OR-GCD (M=4, N=8)	67.1%	69.0%	66.1%	68.5%
Conv+TRL	67.3%	76.5%	62.1%	68.0%
STN in Conv_Block 1,2,3				
Conv+STN	65.0%	71.4%	61.3%	67.7%
Conv+TRL+STN	69.1%	78.8%	63.7%	71.4%
Conv+TRL+STN+SR	69.9%	80.5%	63.9%	72.7%
Conv+TRL+STN+SR+SL	**72.0%**	**84.1%**	**65.2%**	**74.1%**
STN in Conv_Block 1,2,3,4,5				
Conv+STN	65.3%	71.3%	61.9%	68.2%
Conv+TRL+STN	69.2%	79.0%	63.6%	71.6%
Conv+TRL+STN+SR	70.3%	81.2%	64.2%	73.2%
Conv+TRL+STN+SR+SL	**74.1%**	**87.1%**	**66.8%**	**74.8%**

Conv denotes the ResNet-50 model, **SR** and **SL** for **S**parisity **R**egularization R and **S**eg**L**oss S, respectively. **DSCN** denotes Deep Siamese Convolutional Network [36], and **OR-GCD** denotes Overlapping Region-based Global Context Descriptor [45].

TABLE II: Agreements (Cohen's Kappa with 95% confidence interval (CI)) between Algorithm, Ground Truth (GT) and two independent observers.

	Algorithm	GT	Observer 1	Observer 2
Algorithm	1	0.38 (0.21-0.54)	0.32 (0.13-0.50)	0.27 (0.08-0.46)
GT	0.38 (0.21-0.54)	1	0.79 (0.67-0.90)	0.71 (0.58-0.84)
Observer 1	0.32 (0.13-0.50)	0.79 (0.67-0.91)	1	0.92 (0.84-1.00)
Observer 2	0.27 (0.08-0.46)	0.71 (0.58-0.84)	0.92 (0.84-1.00)	1

Cohen's Kappa can be interpreted as follows: 0-0.20 no to slight agreement; 0.21-0.40 fair agreement; 0.41-0.60 as moderate agreement; 0.61-0.80 as substantial agreement; 0.81-1.00 almost perfect agreement.

using a modified consensus ground truth of the three human scorers (original ground truth and the two independent observers), where at least two of three scores were identical. If the consensus was 'uncertain' these were also removed. All analyses were carried out in the programming language R. The agreements between two independent observers, ground truth and Algorithm are shown in Table II. An almost perfect agreement occurred between the two observers (Kappa 0.92) and a substantial agreement between the observers and ground truth used to develop the Algorithm (Kappa 0.71-0.79). However, only a fair agreement occurred between the Algorithm and two observers (Kappa 0.27-0.32). The sensitivity and specificity for the Algorithm to detect change using a consensus ground truth of the three human observers was 73% (95%CI (Confidence Interval) 0.58-0.84) and 65% (95%CI 0.54-0.76), respectively. While a substantial agreement occurred between the ground truth and the two experienced melanographers trained in dermoscopy monitoring, there was only a fair agreement with the Algorithm and those melanographers. The proposed method is required to be further improved before its use in the clinic.

E. Effectiveness of TRL

Table III shows the detailed network settings of our proposed TRL process. In order to obtain a generalized spatial representation, we down-sample the spatial resolution. Therefore, we introduce smaller sizes of U_1 and U_2 in the later TRL blocks. In addition, mode-3 in the tensor represents the correlation of different feature maps. As a large size of U_3 results in more feature maps in the current TRL block, we introduce a larger size of U_3 in the later TRL blocks.

Fig. 6 shows some examples of the extracted feature maps from different TRL blocks in our proposed model. The original test images are displayed in the first row, whereas the feature maps from different TRL blocks are displayed in the remaining rows. It can be observed that the proposed TRL mechanism is able to extract the global representation of a melanoma image at different levels. Along the increase of the block depth, the extracted features focus more on the salient regions of a lesion.

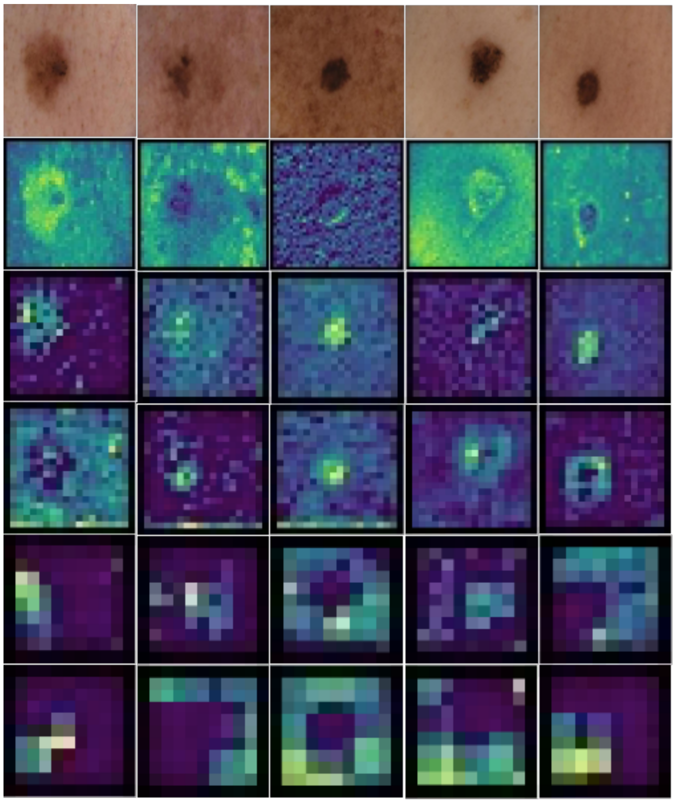


Fig. 6: Sample feature maps learned by our proposed TRL blocks. Note that we only displayed 5 feature maps of the feature cube from each TRL block for the illustration purpose.

TABLE III: Detailed network settings of proposed TRL. In each hidden layer, the size of the offset B is identical with its output.

Layer	U_1	U_2	U_3	B & **Output Size**
TRL 1	112 × 224	112 × 224	10 × 3	112 × 112 × 10
TRL 2	56 × 122	56 × 122	20 × 10	56 × 56 × 20
TRL 3	28 × 56	28 × 56	40 × 20	28 × 28 × 40
TRL 4	14 × 28	14 × 28	100 × 40	14 × 14 × 100
TRL 5	7 × 14	7 × 14	100 × 100	7 × 7 × 100
TGAP	-	-	-	7+7+100

F. Penalty Term Selection

In order to select the best fitting penalty term for the regularization factor, we investigated different combinations of $\lambda_1, \lambda_2, \beta$ and γ. To avoid exploring exhaustive combinations which is computationally expensive, we selected 4 combinations. Detailed settings and corresponding performance are shown in Table IV. Note that we select the best fitting penalty term, and add the following constraint into the model in order to prevent the gradient vanishing problem:

$$\lambda_1 + \lambda_2 + \beta + \gamma = 0.1, \tag{25}$$

subject to $[\lambda_1, \lambda_2, \beta, \gamma] \in (0, 0.1)$.

TABLE IV: Performance comparison of different penalty terms.

Case	λ_1	λ_2	β	γ	Acc	Sens
Comb_1	0.025	0.025	0.025	0.025	68.4%	75.8%
Comb_2	0.03	0.03	0.02	0.02	69.7%	81.1%
Comb_3	0.04	0.04	0.01	0.01	**72.0%**	**84.1%**
Comb_4	0.045	0.045	0.005	0.005	71.8%	83.7%

As shown in Table IV, the best accuracy 72.0% and the best sensitivity are achieved with **Comb_3**. Therefore, we used the **Comb_3** as the penalty term setting for the subsequent experiments.

From Table IV, we can also conclude that emphasizing more on Norm-2 regularization is more helpful for improving the Sensitivity of our proposed algorithm. For instance, when all the penalty terms are equal to 0.025 (ratio $= (0.025 + 0.025)/(0.025 + 0.025) = 1$), the performance is relatively low: Acc $= 68.4\%$ and Sens $= 75.8\%$. When λ_1 and λ_2 take a larger proportion (e.g., λ_1, $\lambda_2 = 0.4$ and ratio $= (0.04+0.04)/(0.01+0.01) = 4$), the performance is relative better: Acc $= 72.0\%$ and Sens $= 84.1\%$. That is, the L2-norm penalty is more sensitive to the short-term melanoma image change detection task than the Sparsity term or other high-level regularization factor, such as segmentation loss.

G. Effectiveness of STN

We add STNs as a sub network before the conv_1, conv_2, conv_3, conv_4 and conv_5 blocks of the ResNet-50. The detailed settings of the STN structure are shown in Table V.

TABLE V: The model setting for STN.

type	filter size	stride	number of filters
conv1	7x7	2	64
conv2	5x5	2	128
FC	6 neurons	-	-

As shown in Table I, introducing the STN layer could boost the performance. However, we were also curious about whether STN is helpful at each layer or some special layers only. Therefore, we removed STN from some layers and compared the performance. When we include STN in every Conv block in ResNet, the best performance was achieved with 74.1% accuracy and 87.0% sensitivity shown in Table VIII. However, introducing too many STNs in our model leads to a significant increase of the training time. We also found that too many hidden layers in the convolutional process may result in the overfitting problem and the vanishing gradient problem.

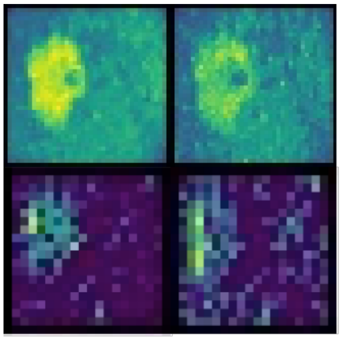


Fig. 7: Example of the extracted feature maps with and without SegLoss in the TRL process (Block#1 in first row and Block#2 in the second row). Features in the first column are extracted from the model with SegLoss and features in the second column are extracted from the model without SegLoss.

H. Effectiveness of SegLoss

We first investigated the hyperparameter n in (17) and (18) which is used to define how many "colors" are in the current feature map. We set n within a range of $[2, 5, 10, 20]$ to investigate its impact on change detection performance. In addition, we investigated the impact of SegLoss on different TRL layers.

As shown in Table VI for the performance comparison of SegLoss in different TRL blocks with STNs in convolutional block 1, 2, and 3, the best accuracy 72.0% and best sensitivity 84.1% can be achieved when $n = 5$ and SegLoss is applied in TRL blocks #1,2,3,4,5. Similarly, it is also observed that when $n = 5$ and 10, the accuracy is promising in general, ranging from 68.1% to 72.0%. In contrast, when n is too large or too small, the accuracy decreases, ranging from 67.6% to 69.9% for $n = 20$ and 69.4% to 70.1% for $n = 2$, which is very close to the benchmark (69.9%) without using any SegLoss in our model. We also observed similar results in terms of sensitivity. When $n = 5$ or 10, the model outperforms the case when $n = 2$ or 20. The top ranked sensitivity score (84.1%) is achieved when $n = 5$ and SegLoss is introduced in all the TRL blocks. When $n = 20$, we can see a significant decrease with SegLoss introduced in more TRL blocks. For example, if SegLoss is introduced in all TRL blocks, the sensitivity is 80.8%, which is very close to that of the benchmark model (80.5%) without any SegLoss. In addition, a greater n will result in higher computational demands. Therefore, we finally set $n = 5$ and added SegLoss in all the TRL blocks.

We also investigated the performance of different STNs setting in all the five convolutional layers. As shown in Table VII, the best accuracy 74.1% and best sensitivity 87.1% can be achieved when $n = 5$ and SegLoss is applied in TRL blocks #1,2,3,4,5.

In order to further investigate the differences caused by different n settings, we compared the extracted feature maps from the TRL process with and without the SegLoss. As shown

TABLE VI: Performance (%) of different settings for our proposed SegLoss with STN in Conv_# Block 1,2,3

	No SegLoss	SegLoss in TRL Block #1	TRL Block #1, 2	TRL Block #1,2,3	TRL Block #1,2,3,4	TRL Block #1,2,3,4,5
Acc n = 2	69.9	69.8	69.4	70.1	70.1	69.9
Acc n = 5	-	69.3	70.2	69.5	71.3	**72.0**
Acc n = 10	-	68.1	68.4	69.5	69.8	71.3
Acc n = 20	-	67.6	68.3	69.2	69.0	69.9
Sens n = 2	80.5	76.8	78.8	80.9	81.6	80.3
Sens n = 5	-	80.8	81.7	82.8	82.6	**84.1**
Sens n = 10	-	77.4	77.5	80.2	80.0	82.1
Sens n = 20	-	76.0	76.5	79.6	79.7	80.8

TABLE VII: Performance (%) of different settings for our proposed SegLoss with STN in Conv_# Block 1,2,3,4,5

	SegLoss in TRL Block #1	TRL Block #1, 2	TRL Block #1,2,3	TRL Block #1,2,3,4	TRL Block #1,2,3,4,5
Acc n = 2	69.8	69.3	70.4	70.6	71.7
Acc n = 5	69.8	70.6	70.5	72.4	**74.1**
Acc n = 10	68.2	68.7	68.7	69.8	67.5
Acc n = 20	67.5	68.7	69.1	68.5	68.2
Sens n = 2	76.5	79.0	81.3	82.5	83.4
Sens n = 5	81.0	82.6	83.7	83.8	**87.1**
Sens n = 10	78.2	77.2	78.7	79.3	75.5
Sens n = 20	75.5	77.3	79.2	80.2	76.2

TABLE VIII: Analysis of STNs in different convolutional layers.

STN in Conv_# Block	Acc	Sens
0	65.9%	74.5%
1	66.1%	75.7%
1,2	70.7%	79.6%
1,2,3	**72.0%**	**84.1%**
1,2,3,4,5	**74.1%**	**87.0%**

in Fig. 7, the feature maps in the first row are extracted in Block#1 and those in the second row are extracted in Block#2. It is clearly observed that utilizing the SegLoss module contributes to a better localized and cleaner feature representation for the TRL process.

V. Conclusion

In this paper, we present the first study of deep learning based lesion change detection for short-term melanoma screening. We formulate change detection as a task measuring the similarity between two dermoscopy images taken on the same lesion over a short time-frame and propose a novel Siamese deep learning network. In order to complement local features obtained through convolutional neural networks such as ResNet, we propose a tensorial neural network to extract global features of a dermoscopy image. In order to overcome the image distortion and alignment problem between two dermoscopy images taken at different times, we improve the ResNet model with a spatial transformer network. As STN could not always locate the most significant segments, we introduced the SegLoss regularization factor, which helps improve the detection performance. While the experimental results and comprehensive analysis demonstrate a great promise of deep learning based short-term lesion change detection, there was only a fair agreement between the algorithm and those independent melanographers, which requires more algorithmic improvements before the use in the clinic.

The aim of our study was to explore image analysis methods to detect change using the conventional clinical definitions as currently described [6]–[8], [57]. We believe that this is a paradigm shift of what is currently found. Furthermore, given a large enough data set, with changed excised lesions labelled with the histological diagnosis of melanoma versus benign, and incorporating both static and sequential image analysis with deep learning techniques, the end-point of benign versus melanoma could be achieved with a significantly greater accuracy than is found today.

References

[1] "What is skin cancer?" https://www.cancer.org/cancer/skin-cancer/prevention-and-early-detection/what-is-skin-cancer.html.

[2] "Cancer data in australia," https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/contents/summary.

[3] "Melanoma: Stages, types, causes, and pictures - medical news today," https://www.medicalnewstoday.com/articles/154322.php.

[4] J. Grogan, C. L. Cooper, T. J. Dodds, P. Guitera, S. W. Menzies, and R. A. Scolyer, "Punch 'scoring': a technique that facilitates melanoma diagnosis of clinically suspicious pigmented lesions," *Histopathology*, vol. 72, no. 2, pp. 294–304, 2018.

[5] H. Kittler, A. A. Marghoob, G. Argenziano, C. Carrera, C. Curiel-Lewandrowski, R. Hofmann-Wellenhof, and *et al.*, "Standardization of terminology in dermoscopy/dermatoscopy: Results of the third consensus conference of the international society of dermoscopy," *Journal of the American Academy of Dermatology*, vol. 74, no. 6, pp. 1093 – 1106, 2016.

[6] D. Altamura, M. Avramidis, and S. W. Menzies, "Assessment of the optimal interval for and sensitivity of short-term sequential digital dermoscopy monitoring for the diagnosis of melanoma," *Archives of Dermatology*, vol. 144, no. 4, pp. 502–506, April 2008.

[7] H. Kittler, P. Guitera, E. Riedl, M. Avramidis, L. Teban, M. Fiebiger, and *et al.*, "Identification of Clinically Featureless Incipient Melanoma Using Sequential Dermoscopy Imaging," *Archives of Dermatology*, vol. 142, no. 9, pp. 1113–1119, Sep 2006.

[8] S. W. Menzies, A. Chamberlain, P. Guitera, H. P. Soyer, and Cancer Council Australia Melanoma Guidelines Working Party. (2018) What is the role of sequential digital dermoscopy imaging in melanoma diagnosis? https://wiki.cancer.org.au/australiawiki/index.php?oldid=186222.

[9] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2092–2103, Sep. 2019.

[10] K. Matsunaga, A. Hamada, A. Minagawa, and H. Koga, "Image classification of melanoma, nevus and seborrheic keratosis by deep neural network ensemble," *CoRR*, vol. abs/1703.03108, 2017.

[11] I. González-Díaz, "Dermaknet: Incorporating the knowledge of dermatologists to convolutional neural networks for skin lesion diagnosis," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 547–559, March 2019.

[12] P. Tschandl, N. Codella, B. N. Akay, G. Argenziano, R. P. Braun, H. Cabo, and *et al.*, "Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study," *The Lancet Oncology*, vol. 20, no. 7, pp. 938 – 947, 2019.

[13] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and *et al.*, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[14] N. A. Koohbanani, M. Jahanifar, N. Z. Tajeddin, A. Gooya, and N. M. Rajpoot, "Leveraging transfer learning for segmenting lesions and their attributes in dermoscopy images," *CoRR*, vol. abs/1809.10243, 2018.

[15] F. Navarro, M. Escudero-Viñolo, and J. Bescós, "Accurate segmentation and registration of skin lesion images to evaluate lesion change," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 501–508, March 2019.

[16] D. Lu, P. Mausel, E. S. Brondízio, and E. Moran, "Change detection techniques," *International Journal of Remote Sensing*, vol. 25, no. 12, pp. 2365–2401, 2004.

[17] P. Coppin, I. Jonckheere, K. Nackaerts, B. Muys, and E. Lambin, "Review articledigital change detection methods in ecosystem monitoring: a review," *International Journal of Remote Sensing*, vol. 25, no. 9, pp. 1565–1596, 2004.

[18] V. Nika, P. Babyn, and H. Zhu, "Change detection of medical images using dictionary learning techniques and principal component analysis," *Journal of Medical Imaging*, vol. 1, no. 2, pp. 1 – 21, 2014.

[19] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning cnn image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1655–1668, July 2019.

[20] F. Nachbar, W. Stolz, T. Merkle, A. B. Cognetta, T. Vogt, M. Landthaler, and *et al.*, "The ABCD rule of dermatoscopy: High prospective value in the diagnosis of doubtful melanocytic skin lesions," *Journal of the American Academy of Dermatology*, vol. 30, no. 4, pp. 551 – 559, 1994.

[21] G. Argenziano, G. Fabbrocini, P. Carli, V. De Giorgi, E. Sammarco, and M. Delfino, "Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions: Comparison of the ABCD Rule of Dermatoscopy and a New 7-Point Checklist Based on Pattern Analysis," *Journal of American Medical Association of Dermatology*, vol. 134, no. 12, pp. 1563–1570, Dec 1998.

[22] S. Jain, V. jagtap, and N. Pise, "Computer aided melanoma skin cancer detection using image processing," *Procedia Computer Science*, vol. 48, pp. 735 – 740, 2015.

[23] R. Amelard, J. Glaister, A. Wong, and D. A. Clausi, "High-level intuitive features (hlifs) for intuitive skin lesion description," *IEEE Transactions on Biomedical Engineering*, vol. 62, no. 3, pp. 820–831, March 2015.

[24] G. Pellacani, A. M. Cesinaro, C. Longo, C. Grana, and S. Seidenari, "Microscopic In Vivo Description of Cellular Architecture of Dermoscopic Pigment Network in Nevi and Melanomas," *Journal of the American Medical Association of Dermatology*, vol. 141, no. 2, pp. 147–154, Feb 2005.

[25] M. E. Celebi, H. Iyatomi, W. V. Stoecker, R. H. Moss, H. S. Rabinovitz, G. Argenziano, and *et al.*, "Automatic detection of blue-white veil and related structures in dermoscopy images," *Computerized Medical Imaging and Graphics*, vol. 32, no. 8, pp. 670 – 677, 2008.

[26] M. Sadeghi, T. K. Lee, D. McLean, H. Lui, and M. S. Atkins, "Detection and analysis of irregular streaks in dermoscopic images of skin lesions," *IEEE Transactions on Medical Imaging*, vol. 32, no. 5, pp. 849–861, May 2013.

[27] T. Yao, Z. Wang, Z. Xie, J. Gao, and D. D. Feng, "A multiview joint sparse representation with discriminative dictionary for melanoma detection," in *International Conference on Digital Image Computing: Techniques and Applications*, Nov 2016, pp. 1–6.

[28] F. Xie, H. Fan, Y. Li, Z. Jiang, and R. M. A. Bovik, "Melanoma classification on dermoscopy images using a neural network ensemble model," *IEEE Transactions on Medical Imaging*, vol. 36, no. 3, pp. 849–858, March 2017.

[29] N. Codella, J. Cai, M. Abedini, R. Garnavi, A. Halpern, and J. R. Smith, "Deep learning, Sparse coding, and SVM for melanoma recognition in dermoscopy images," in *Machine Learning in Medical Imaging*, L. Zhou, L. Wang, Q. Wang, and Y. Shi, Eds. Cham: Springer International Publishing, 2015, pp. 118–126.

[30] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation," *CoRR*, vol. abs/1802.06955, 2018.

[31] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, and *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Journal of the American Medical Association*, vol. 316, no. 22, pp. 2402–2410, 2016.

[32] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 80, pp. 91 – 106, 2013.

[33] L. Jia, M. Li, P. Zhang, and Y. Wu, "Sar image change detection based on correlation kernel and multistage extreme learning machine," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 10, pp. 5993–6006, Oct 2016.

[34] A. F. Zuur, E. N. Ieno, and G. M. Smith, "Principal component analysis and redundancy analysis," *Analysing Ecological Data*, pp. 193–224, 2007.

[35] T. Kasetkasem and P. K. Varshney, "An image change detection algorithm based on markov random field models," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 40, no. 8, pp. 1815–1823, Aug 2002.

[36] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 10, pp. 1845–1849, Oct 2017.

[37] J. Zhao, M. Gong, J. Liu, and L. Jiao, "Deep learning to classify difference image for image change detection," in *International Joint Conference on Neural Networks*, July 2014, pp. 411–417.

[38] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred deep learning-based change detection in remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–14, 2019.

[39] W. Zhao, S. Du, and W. J. Emery, "Object-based convolutional neural network for high-resolution imagery classification," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 7, pp. 3386–3396, July 2017.

[40] F. Zhang, B. Du, and L. Zhang, "Saliency-guided unsupervised feature learning for scene classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 2175–2184, April 2015.

[41] W. Zhao and S. Du, "Spectral-spatial feature extraction for hyperspectral image classification: A dimension reduction and deep learning approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 8, pp. 4544–4554, Aug 2016.

[42] R. Simões and C. Slump, "Change detection and classification in brain mr images using change vector analysis," in *The Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug 2011, pp. 7803–7807.

[43] M. D. Li, K. Chang, B. Bearce, C. Y. Chang, A. J. Huang, J. P. Campbell, and *et al.*, "Siamese neural networks for continuous disease severity evaluation and change detection in medical imaging," *NPJ Digital Medicine*, vol. 3, no. 48, 2020.

[44] L. Liu, Y. Lu, and C. Y. Suen, "Variable-length signature for near-duplicate image matching," *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1282–1296, April 2015.

[45] Z. Zhou, Y. Wang, Q. M. J. Wu, C. Yang, and X. Sun, "Effective and efficient global context verification for image copy detection," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 48–63, Jan 2017.

[46] R. Connor and F. A. Cardillo, "Quantifying the specificity of near-duplicate image classification functions," in *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 01 2016, pp. 647–654.

[47] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European Conference on Computer Vision*. Springer, 2008, pp. 304–317.

[48] W. Zhao, X. Wu, and C. Ngo, "On the annotation of web videos by efficient near-duplicate search," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 448–461, Aug 2010.

[49] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[50] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," 06 2007.

[51] W. Zhou, H. Li, Y. Lu, and Q. Tian, "SIFT match verification by geometric coding for large-scale partial-duplicate web image search," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 9, no. 1, pp. 1–18, 2013.

[52] M. Bai, B. Zhang, and J. Gao, "Tensorial neural networks and its application in longitudinal network data analysis," in *International Conference on Neural Information Processing*, Oct 2017, pp. 1–5.

[53] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *Society for Indystrial and Applied Mathematics*, vol. 51, no. 3, pp. 455–500, 2008.

[54] J. Gao, Y. Guo, and Z. Wang, "Matrix neural networks," in *International Symposium on Neural Networks*, May 2017, pp. 313–320.

[55] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2017–2025.

[56] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, Jan 1979.

[57] S. W. Menzies, A. Gutenev, M. Avramidis, A. Batrac, and W. H. McCarthy, "Short-term digital surface microscopic monitoring of atypical or changing melanocytic lesions," *Archives of Dermatology*, vol. 137, no. 12, pp. 1583–1589, Dec 2001.

[58] S. W. Menzies, K. Crotty, C. Ingvar, and W. McCarthy, *Dermoscopy: An Atlas, 3rd Edition*. McGraw-Hill Education Australia, 2009.

[59] D. G. Lowe, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, vol. 2, Sep. 1999, pp. 1150–1157.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 770–778.

[61] S. Vitaladevuni, F. Choi, R. Prasad, and P. Natarajan, "Detecting near-duplicate document images using interest point matching," in *International Conference on Pattern Recognition*, Nov 2012, pp. 347–350.