ZooME: Efficient Melanoma Detection Using Zoom-in Attention and Metadata Embedding Deep Neural Network

Xiaoyan Xing1,2, Pingping Song1, Kai Zhang1, Fang Yang3,4,† and Yuhan Dong1,2,†, Member, IEEE

Abstract—Melanoma detection is a crucial yet hard task for both dermatologists and computer-aided diagnosis (CAD). Many traditional machine learning algorithms including deep learning-based methods are employed for melanoma classification. However, more and more complex network architectures do not harvest a leap in model performance. In this paper, we aim to enhance the credibility of CAD approach for melanoma by paying more attention to clinically important information. We propose a Zoom-in Attention and Metadata Embedding (ZooME) melanoma detection network by: 1) introducing a Zoom-in Attention model to better extract and utilize unique pathological information of dermoscopy images; 2) embedding patients’ demographic information including age, gender, and anatomic body site, to provide well-rounded information for better prediction. We apply a ten-fold cross-validation on the latest ISIC-2020 dataset with 33,126 dermoscopy images. The proposed ZooME achieved state-of-the-art results with 92.23% in AUC score, 84.59% in accuracy, 85.95% in sensitivity, and 84.63% in specialty, respectively.

Clinical relevance—This work establishes an efficient melanoma detection method with pathological and demographic information in consideration.

I. INTRODUCTION

Melanoma is a serious type of cancer accounting for a majority of skin cancer deaths. According to the American Cancer Society, new diagnoses and new deaths of melanoma are expected to reach 106,110 and 7,180 respectively in the United States for 2021, and the numbers are still increasing [1]. Developing detection procedures to make early and accurate diagnosis of melanoma is of great importance to increase the survival rate.

Traditionally, dermatologists used to diagnose melanoma based on dermoscopy images with characteristics including shape, color and texture in consideration. However, early-stage melanomas have similar morphological features as skin moles, even experienced experts would be easily confused. Over the past two decades, computer-based methods for automatic melanoma diagnosis emerged [2] and freed dermatologists from the predicament of relying on manual judgement alone. Various algorithms have been proposed for the segmentation, feature extraction and melanoma classification [3]. The task has been further simplified with the advent of deep learning methods [4]. However, the lack of large collections of labeled data and the poor interpretability lead to the gap between the existing deep convolutional networks designed for melanoma diagnosis and clinical significance. To bridge the gap, we aim to give more information to the model from both pathological and epidemiological perspectives. From pathological aspect, one of the most common used procedural assessment for dermatologists is referred to as ABCD rule [5], which includes asymmetry, border, color, and diameter. Aside from common traits existed in natural images, medical images carry more pathological information which are hard to be extracted by traditional convolutional neural networks (CNN). Moreover, patients’ demographic characteristics including age, gender etc. are recorded and can be further utilized. Prior studies verified the relationship between melanoma incidence and demographic factors [6] [7]. To the best of our knowledge, few deep learning based methods have taken this relationship into consideration.

In this paper we propose a Zoom-in Attention and Metadata Embedding (ZooME) network, to bridge the gap between computer-aided diagnosis of melanoma and clinical significance. Our work has three main contributions:

• We provide an efficient melanoma detection framework ZooME and yield state-of-the-art results on the newest ISIC-2020 datasets.

• We present a flexible Zoom-in Attention module based on squeeze-excitation (SE) mechanism [8]. This module

1Xiaoyan Xing, Pingping Song, Kai Zhang, and Yuhan Dong are with Shenzhen Internation Graduate School, Tsinghua University, Shenzhen, China. dongyuhan@sz.tsinghua.edu.cn
2China International Exchange and Promotive Association for Medical and Health Care, Beijing, China.
3Fang Yang is with the Department of Dermatology, Shenzhen People’s Hospital (The Second Clinical Medical College, Jinan University; The First Affiliated Hospital, Southern University of Science and Technology), Shenzhen, Guangdong, China. yangfang3013@gmail.com
4Candidate Branch of National Clinical Research Center for Skin Diseases.
† Co-corresponding author
can be embedded in any existing CNN models to fully utilize more pathological information.

- We introduce Metadata Embedding (ME) mechanism, which strengthens the framework by taking demographic factors into consideration.

## II. Related Work

### A. Learning Based Melanoma Detection

Rubegni et al. [3] introduced artificial neural network (ANN) to automatic melanoma diagnosis, which only applied a few feature annotations and obtained a better accuracy than traditional pattern recognition methods. Recently, CNN has been studied to complete the classification task for dermoscopy images [4]. Maron et al. [9] proposed a CNN based melanoma classification by employing ResNet50 as the feature extraction backbone, which outperformed on average a team of senior dermatologists in terms of sensitivity. Reisinho et al. [10] combined three deep CNN architectures as the backbone for melanoma detection but obtained a limited increments.

However, simply increasing network complexity does not result in significant improvements in terms of classification performance. Clinically, dermatologists make diagnosis followed by ABCD rules, which have not been adequately learned in common CNN architectures.

### B. Demographic Characteristics for Melanoma Detection

Apart from image information, patients’ physiological data can play an important role in clinical diagnosis. Andrew et al. [6] concluded that patients’ age is highly relevant with melanoma morbidity through a cohort study of 8772 patients. Yuan et al. [7] found that patients’ demographic characteristics including gender, age and race are risk factors of melanoma.

Inspired by the relationship between melanoma incidence rate and patients’ physiological traits verified by previous works, we aim to utilize this relationship to assist clinical diagnosis.

## III. Method

As introduced in Section II, the common architecture of CNN based classification models are mostly designed for natural images. A fine-tune CNN by possibly utilizing more unique pathological information carried in dermoscopy images is of great clinical importance. Meanwhile, demographic characteristics including age, gender, etc. are not fully considered in learning based melanoma detection methods nowadays.

To address these two issues mentioned above, several things need to be done: 1) a comprehensive feature extraction backbone which can pay more attention to pathological information of the dermoscopy images; 2) a mechanism which can utilize patients’ demographic information.

Thus we propose ZooME, which can: 1) give more emphasis to pathological traits a special Zoom-in SE module; 2) introduce patients’ age, gender and anatomic sites to the last step of feature extraction by Metadata Embedding mechanism. Details will be demonstrated below.

### A. Zoom-SE Module

SE module [8] is a comprehensive attention mechanism in natural image classification. Inspired by this, we propose Zoom-in SE (Zoom-SE) module to give extra attention to the clinically interested information of dermoscopy images. Zoom-SE consists of three parts: 1) zoom-in and squeeze; 2) channel re-weighting; 3) spatial re-weighting. The details of each component are introduced below.

1) **Zoom-in and Squeeze:** Zoom-in is employed to find the most unique area in the given feature map $u$. The input feature map $u \in \mathbb{R}^{C \times H \times W}$ is transformed by the $4 \times 4$ adaptive max pooling (AMP) into a statistic $z \in \mathbb{R}^{C \times 4 \times 4}$. We then squeeze $z$ to a channel-wise statistic $s \in \mathbb{R}^{C \times 1 \times 1}$. Unless otherwise stated, the subscript $c$ of all variables in the following formulas represents the $c$-th component of the corresponding variables, i.e. $s_c$ represents as the $c$-th element of $s$. $s_c$ can be transformed by the squeeze operation $F_{sq}$:

$$s_c = F_{sq}(z_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} z_c(i, j), \quad (1)$$

where $H$ is height and $W$ is weight of $z_c$, $z_c(i, j)$ is the $(i, j)$-th pixel of $z_c$.

**Remark:** In this task, $H$ and $W$ are set to 4 as the same scale as the adaptive pooling. Mathematically, the adaptive pooling size can be set to any scale, $4 \times 4$ is the best selection after multiple scale experiments.

2) **Channel Re-weighting:** After Zoom-in and Squeeze operation, $s$ is squeezed into a $C$ channel statistic, therefore, we can apply a channel re-weighting mechanism by using fully connected (FC) layer $W$. Same as the SE-Module, the channel re-weighting $F_{cr}$ can be formulated in:

$$q = F_{cr}(s, W) = \sigma(W_2\sigma(W_1s)), \quad (2)$$

where $\sigma$ represents the ReLU fuction, $W_1$ and $W_2$ are the FC layers. $W_1 \in \mathbb{R}^{C \times C}$ and $W_2 \in \mathbb{R}^{C \times C}$.

3) **Spatial Re-weighting:** After channel re-weighting, $q$ is a re-weighted channel-wise statistic. A spatial reweighting transformation $F_{sr}$ is employed to generate the final output $x$ with Sigmoid activation:

$$x_c = F_{sr}(u_c, s_c) = s_c \otimes u_c, \quad (3)$$

where $x = \{x_1, x_2, \ldots, x_c\}$ and $\otimes$ refers to channel-wise multiplication.

### B. Metadata Embedding

This module consists of two FC layers and an embedding operation. Firstly, metadata $m$ is encoded to a $N$ channels tensor $t \in \mathbb{R}^{N \times 1 \times 1}$ ($N$ equals to the number of channels of $x$) through the FC layers; then $t$ is embedded in the feature map $x$ from the dermoscopy images through channel-wise multiplication $\otimes$:

$$o = t \otimes x, \quad (4)$$

where $o$ is the feature map after the embedding operation, $o \in \mathbb{R}^{N \times 1 \times 1}$. 


C. Example of Using ZooME

Zoom-SE and Metadata Embedding can be easily implanted to the commonly used CNN architectures and enhance the classification ability of CNN on melanoma classification. In this work, we choose ResNeXt50 [11] as our baseline backbone, and implant ZooME in its architecture. Fig. 2 illustrates the architecture of the proposed network\(^1\).

The network consists of two parts:

- Feature extractor: an activate convolution layer and four convolution groups with residual blocks to extract the feature map from input dermoscopy images.
- Classifier: a fully connected layer with Sigmoid function to predict the probability of melanoma.

Feature extractor takes dermoscopy image as input. Zoom-SE module takes each residual block’s output feature map \(u\) as input, and outputs the re-weighted feature map \(x\). In the next residual block, \(x\) is regarded as input. Metadata Embedding can be regarded as a bypass branch with metadata label as input. The encoded metadata \(m\) is embedded in the output of last Zoom-SE module. Through multiple iterations of the above operations, we hope the pathological and the demographic information will be given extra attention through channel and spatial, respectively.

Classifier takes the final feature map \(x\) as input and outputs the final predicted probability \(p_t\) of melanoma through an adaptive pooling layer and a FC layer.

D. End-to-end optimization

Considering the imbalance of the dataset as well as the realistic (the melanoma is a rare tumor of skin), we choose focal loss [12] as the loss function. The loss function is given by:

\[
L_{focal}(p_t) = -\alpha_t(1 - p_t)\gamma \log(p_t),
\]

where \(p_t\) is the predicted probability of melanoma. \(\alpha_t\) and \(\gamma\) are the hyper-parameters and set to \(\alpha_t = 0.25\), and \(\gamma = 2\) in this work. By adopting such loss, the whole network could be trained end-to-end.

IV. EXPERIMENT

In this section, we firstly present quantitative evaluation compared with several state-of-the-art methods [13] [9] on international skin image classification (ISIC) 2020 dataset\(^2\); secondly, we apply ablation study on the proposed network and common backbones.

A. Dataset

We use the ISIC dataset, which is an open-access repository dataset. The ISIC dataset contains 33,126 images with 32,532 labeled as ‘benign’ and 594 labeled as ‘malignant’. Following the distribution of the whole dataset, we split the dataset into ten folds each contains 3,313 images (58 to 60 images are malignant) for cross-validation.

B. Settings

Data augmentation: To overcome over-fitting and to simulate the real situation, we apply a data augmentation. All images are randomly cropped from the central with zoom from 0.8 to 1. Meanwhile, the images are rotated by random angle between \(-30^\circ\) and \(+30^\circ\). Left-right flipping with a probability of 0.5 is also applied. After augmentation, all images are resized to 224\(\times\)224 \(\times\) 3.

Implementation and Training: We implement the proposed network based on Pytorch using RTX 2080 Ti GPU with the learning rate of \(1\times10^{-4}\) and batch size of 64. The Adam optimizer [14] is employed to train the network.

\(^1\)Codes are available from the corresponding author on reasonable request.

\(^2\)https://www.kaggle.com/c/siim-isic-melanoma-classification/data
C. Metrics
The accuracy (ACC) serving as the only metric may be misleading since the number of samples is imbalanced severely. Alternatively, we adopt area under the curve (AUC) as the premier metric and sensitivity (SE) and specialty (SP) as supplementation.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specialty</th>
<th>ROC_AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maron et al. [9]</td>
<td>77.77±5.04</td>
<td>81.63±6.65</td>
<td>76.97±4.58</td>
<td>85.57±2.13</td>
</tr>
<tr>
<td>Kaur et al. [13]</td>
<td>79.88±4.38</td>
<td>80.77±7.01</td>
<td>79.17±4.23</td>
<td>85.85±2.01</td>
</tr>
<tr>
<td>AlexNet [15]</td>
<td>78.50±4.61</td>
<td>78.28±6.98</td>
<td>76.81±4.04</td>
<td>85.02±1.54</td>
</tr>
<tr>
<td>ZooME</td>
<td>84.59±2.94</td>
<td>85.95±2.81</td>
<td>84.63±3.03</td>
<td>92.23±1.69</td>
</tr>
<tr>
<td>w/o Zoom-SE</td>
<td>80.34±4.16</td>
<td>82.30±4.77</td>
<td>80.32±3.22</td>
<td>86.78±2.15</td>
</tr>
<tr>
<td>w/o Metadata</td>
<td>84.16±2.36</td>
<td>85.62±3.70</td>
<td>84.44±2.32</td>
<td>91.66±1.27</td>
</tr>
</tbody>
</table>

D. Experiment on ISIC dataset
For a more fair comparison, we reproduce some state-of-the-art results. It should be noticed that all the hyperparameters are carefully adjusted to achieve the best performance. The quantitative results are reported in Table I. Our proposed ZooME network outperforms other latest methods, in terms of all the four metrics. Particularly, the proposed ZooME achieved state-of-the-art result with 92.23% in AUC score, 84.59% in accuracy, 85.95% in sensitivity, and 84.63% in specialty.

Moreover, We also apply ablation studies on the proposed network to evaluate the contribution of each module, which are listed in Table II. The two proposed modules enhance the performance of ResNet50 [16] and AlexNet [15] in melanoma detection.

V. CONCLUSIONS
In this paper, we proposed an efficient melanoma detection framework, ZooME, to help CNN architectures to better identify melanoma by not only taking pathological information of dermoscopy images, but also combing demographic characteristics. Experimental results suggested that ZooME network achieved state-of-the-art performance and outperformed traditional approaches in terms of AUC score, accuracy, sensitivity and specialty. As for future work, We will work closely with dermatologists to apply this method in real clinical diagnosis.

ACKNOWLEDGMENT
This research is supported in part by the National Key R&D Program of China (2018YFC0116800). The corresponding authors are Yuhan Dong and Fang Yang.

REFERENCES