Computerized Medical Imaging and Graphics 93 (2021) 101992

Contents lists available at ScienceDirect

Computerized Medical Imaging and Graphics

journal homepage: www.elsevier.com/locate/compmedimag

Classification of squamous cell carcinoma from FF-OCT images: Data selection and progressive model construction

Chi-Jui Ho [a,b], Manuel Calderon-Delgado [c], Ming-Yi Lin [d], Jeng-Wei Tjiu [d], Sheng-Lung Huang [c], Homer H. Chen [a,b,e,*]

[a] Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan
[b] Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan
[c] Graduate Institute of Photonics and Optoelectronics, National Taiwan University, Taipei 10617, Taiwan
[d] Department of Dermatology, College of Medicine, National Taiwan University Hospital, Taipei 10002, Taiwan
[e] Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei 10617, Taiwan

ARTICLE INFO

Keywords:
Optical coherence tomography
Deep learning
Regularization
Training strategy
Convolutional neural network
Squamous cell carcinoma

ABSTRACT

We investigate the speed and performance of squamous cell carcinoma (SCC) classification from full-field optical coherence tomography (FF-OCT) images based on the convolutional neural network (CNN). Due to the unique characteristics of SCC features, the high variety of CNN, and the high volume of our 3D FF-OCT dataset, progressive model construction is a time-consuming process. To address the issue, we develop a training strategy for data selection that makes model training 16 times faster by exploiting the dependency between images and the knowledge of SCC feature distribution. The speedup makes progressive model construction computationally feasible. Our approach further refines the regularization, channel attention, and optimization mechanism of SCC classifier and improves the accuracy of SCC classification to 87.12% at the image level and 90.10% at the tomogram level. The results are obtained by testing the proposed approach on an FF-OCT dataset with over one million mouse skin images.

1. Introduction

Non-invasive medical techniques, which extract biomedical information without contacting the internal body, have become increasingly popular in diagnostic imaging, clinical staging, and therapy for effectivity and convenience reasons (Gollakota et al., 2011; Huang et al., 1991; Dubois and Boccara, 2008; Xiong et al., 2018; Olsen et al., 2018; Wang et al., 2013). Full-field optical coherence tomography (FF-OCT) is a typical non-invasive imaging technique that uses low-coherency light to rebuild sample structures at sub-micron resolution (Dubois and Boccara, 2008; Dalimier and Salomon, 2012; Scholler et al., 2020). Fast data acquisition and high resolution make FF-OCT a desirable 3D imaging technique for retinal image alignment (Wang et al., 2020b; Zhang et al., 2019), blood vessel detection (Lee et al., 2018; Lee et al., 2016) and red blood cell segmentation (Mekonnen et al., 2019), among others.

In this paper, we focus on the classification of squamous cell carcinoma (SCC), the second most common form of skin cancer (Muzic et al., 2017). Algorithms based on convolutional neural network (CNN) can extract features relevant to SCC (Ho et al., 2021). However, since the variety of CNN-based approaches is high and the training workload for each approach is heavy, the exhaustive model construction process is time-consuming. The speed may be improved by using the weights of pre-trained models. However, such weights obtained from other tasks are not applicable to SCC classification because the features for SCC classification are unique, which makes transfer learning powerless (Tan et al., 2018). A novel approach to speed up model construction is required.

The workload required for training each CNN approach has to do with the size of the FF-OCT dataset. In our work, the image size is 576 × 256 pixels, each tomogram contains 439 images, and there is a total of 3373 tomograms in the dataset. In other words, our dataset contains nearly 1.5 million mouse skin images. In practice, it takes at least two days to train a CNN approach on an NVIDIA Tesla V100 (NVIDIA, NVIDIA Tesla V100.Technical Report.⟨https://www.nvidia.com/en-gb/data-center/tesla-v100/⟩.). The actual amount of time is 10 times longer if a 10-fold validation is adopted.

We find that the image data available are not equally relevant to, or effective for, SCC classification. Since the FF-OCT data are densely

* Corresponding author at: Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan.
E-mail address: homer@ntu.edu.tw (H.H. Chen).

https://doi.org/10.1016/j.compmedimag.2021.101992
Received 16 February 2021; Received in revised form 19 May 2021; Accepted 6 September 2021
Available online 23 September 2021

sampled, neighboring images contain similar information and hence present a redundancy. Likewise, most features relevant to SCC classification are located in the upper half of the tomograms; the bottom half of the tomograms may not contain as much information for SCC classification. To reduce the workload and speed up the progressive model construction, we design a strategy to selectively use the training data. Hopefully, by excluding redundant or ineffective data from training, the progressive model construction process can be speeded up with little or no performance drop.

Progressive model construction is often applied to improve classification accuracy. In our work, we fit an ordinary CNN model to the SCC classification task. To achieve the best result, the model construction must take the characteristic of SCC classes into consideration. Since there is no clear cut between different SCC classes, a hard-labeling approach would make the classifier prone to overfitting. Consequently, a soft-labeling approach and a regularization method using mixed labeling are preferred. Channel attention is another important component to consider for model construction. It emphasizes key features while downplaying the significance of other features. This enhances the discrimination power of an SCC classifier. The optimizer of a classifier entails a self-governing mechanism to make an iterative process converge effectively. It is the third component considered in this work for progressive model construction.

The contributions of the work presented in this paper are as follows. We develop an effective training strategy that speeds up the progressive model construction process by a factor of 16. We incorporate channel attention (Hu et al., 2018; Wang et al., 2020a) and cutmix (Yun et al., 2019) into the residual neural network and develop a CNN-based model for SCC classification of mouse skin. Integrated with an FF-OCT device, the overall system provides fast, non-invasive, and accurate SCC classification. It achieves 87.12% SCC classification accuracy at the image-level and 90.10% at the tomogram level.

The remaining parts of this paper are organized as follows. In Section II, we provide a review of FF-OCT, SCC classification, and image classification techniques. In Section III, we describe the characteristics of the FF-OCT dataset and the proposed training strategy. In Section IV, we describe our approach to progressive model construction through regularization (DeVries and Taylor, 2017; Zhang et al., 2017; Yun et al., 2019), channel attention (Hu et al., 2018; Wang et al., 2020a) and optimization (Kingma and Ba, 2014) of CNN, followed by a discussion of the experimental results in Section V. Further discussions of the experiments are provided in Section VI. Finally, the concluding remarks are made in Section VII.

2. Related work

In this section, we describe the FF-OCT imaging techniques used in our system. Then, we discuss the characteristic of SCC and the previous CNN approaches to image classification.

2.1. Full-field optical coherence tomography

Optical coherence tomography (OCT) has been developed for non-invasive diagnosis and high-resolution imaging (Huang et al., 1991; Izatt and Choma, 2008; Drexler et al., 2014). OCT can work in either the time domain (Huang et al., 1991) or the frequency domain (Kalkman, 2017). Our FF-OCT is a variant of time-domain OCT, aiming for significant imaging speed improvement. It performs interferometry using a broad and bright light source and collects the back-scattered light to reconstruct the tissue anatomy at the cellular-level scale. A schematic diagram of our system is shown in Fig. 1. The sample, beam splitter, and reference mirror are placed in front of an objective that focuses the light into the sample to minimize the impact of environmental vibration and disturbance (Tsai et al., 2014).

Conventional time-domain OCT systems perform single-point scanning using a photo-diode with single-pixel, therefore, the data collection

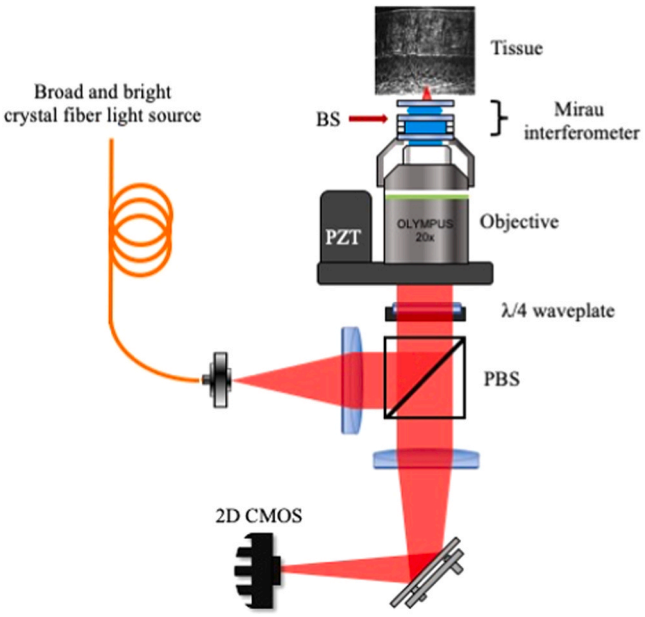

Fig. 1. The schematic diagram of our FF-OCT system. BS: beam splitter; PBS: polarizing beam splitter; 2D-CMOS: two-dimensional complementary metal-oxide-semiconductor.

process is time-consuming. To speed it up, the FF-OCT uses a high-speed 2D camera to perform all the A-scans simultaneously, avoiding the need for lateral scans. Consequently, a 3D tomogram can be obtained in 2 min in average. To have a reasonable signal-to-noise ratio, the full well capacity of each image pixel is 17,000 electrons. As a result, more than 40 dB of dynamic range can be achieved. FF-OCT is often used for clinical analysis that needs non-invasive imaging and timely diagnosis (Dalimier and Salomon, 2012; Scholler et al., 2020). For pathological analysis of skin cancer, the hematoxylin and eosin staining is a popular technique (Chan, 2014; Wells et al., 2007). But it is invasive, and the data collection process takes a few days. Therefore, FF-OCT is a good alternative when timeliness is desired.

2.2. Squamous cell carcinoma

Though skin cancer is rarely life-threatening, it accounts for 40% of cancer cases (World Cancer Research Fund, American Institute for Cancer, 2019.Skin cancer report. Technical Report.⟨https://www.wcrf.org/dietandcancer/skin-cancer⟩.). The incidence rate of skin cancer has been rapidly increasing in recent years (Muzic et al., 2017). SCC is the second most common type of skin cancer, and is usually found on skins that are exposed to intensive ultraviolet radiation. It affects two layers, epidermis and dermis, of the epithelial tissue. When skin is affected by SCC, the size and number of keratinocytes cells grow in the epidermis, and the stratum corneum in epidermis becomes thicker, resulting in an increase of the depth of dermal-epidermal junction (DEJ).

In the context of this work, we consider three categories of skin samples: normal, dysplasia, and SCC. Healthy skin samples belong to the normal category. Abnormal skin samples, which are usually observed in inflamed and reddish tissue, belong to dysplasia or SCC category. The dysplasia category is similar to the actinic keratosis in human, which is difficult to discriminate clinically. Abnormal skin samples that have developed into cancer belong to SCC category, which are usually observed in slowly growing warts. Fig. 2 shows an example of FF-OCT image of each category of skin. As we can see, the DEJ of normal skin is closer to the surface than that of skin affected by dysplasia. On the other hand, the DEJ of an SCC skin sample is too deep to be captured by the FF-OCT system. Another feature of SCC is the size of cells. The epidermic cells of the dysplasia and SCC skin samples are larger than

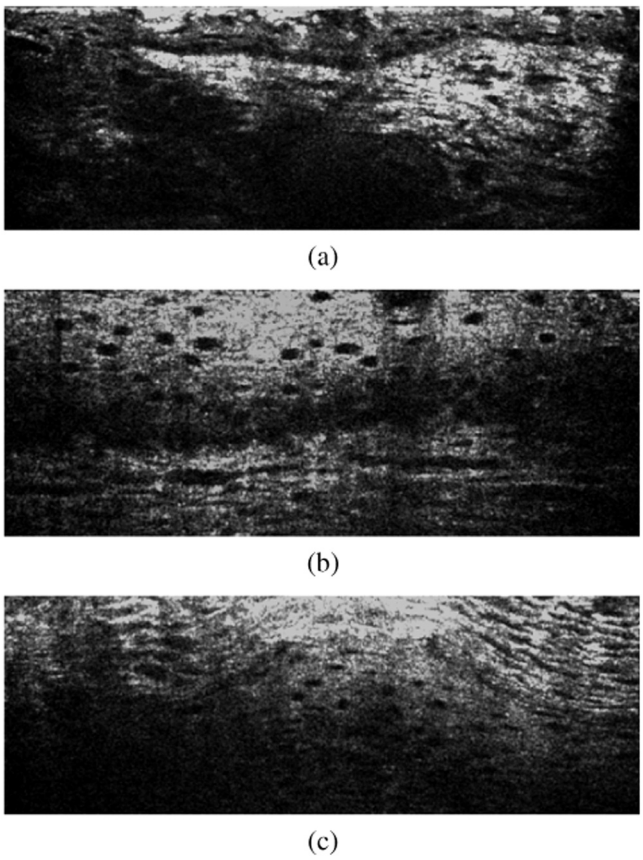

Fig. 2. Example FF-OCT images of (a) normal, (b) dysplasia, and (c) SCC skin.

those of normal skin samples. Moreover, the stratum corneum of an SCC skin sample is much thicker than that of a normal or dysplasia skin sample. Note also that the boundary between these categories may be blurry; there is no clear cut. This characteristic should be considered for SCC classification.

2.3. CNN-based approaches to image classification

Image classification is a core task in computer vision and image processing (Tan and Le, 2019; Russakovsky et al., 2015). CNN is a deep learning approach that has achieved great success in image classification. Krizhevsky et al. pioneered the deep CNN model for image classification (Krizhevsky et al., 2017). He et al. proposed the residual neural network (ResNet) that improves the convergence rate by using shortcuts to connect different layers in each residual block (He et al., 2016). He et al. further applied multi-branch residual blocks to enhance the classification accuracy without increasing the depth or width of the network (Xie et al., 2017).

A key component of CNN-based approaches is the non-linear activation function. The first non-linear function introduced to a deep CNN model is the rectification linear unit (ReLU) (Nair and Hinton, 2010). However, the fragility issue may arise when a large gradient is propagated through the network (Xu et al., 2015)]). To address this issue, many variants with a non-zero slope have been proposed (Xu et al., 2015; He et al., 2015; Clevert et al., 2015). A popular one is the parametric rectified linear unit (PReLU), which enables CNN model to surpass human performance in an image classification task (He et al., 2015).

The representational power of CNN can be improved by adopting the channel attention mechanism. The squeeze-and-excitation (SE) block (Hu et al., 2018) is a pioneering channel attention module that rescales the feature map according to the channel-wise dependency. Wang et al. enhanced the learnability and saved the parameters by handling cross-channel interaction without dimensional reduction (Wang et al., 2020a).

Another popular refinement is the regularization method, which alleviates overfitting for model training. Traditionally, regularization is performed by adding an L2-norm to the loss function. Most regularization methods for CNN perform regional dropout to generate soft labeling while alleviating overfitting. Devries et al. proposed the cutout algorithm that randomly masks an image region and enforces the model to learn every local feature (DeVries and Taylor, 2017). To make the model learn the dependency between different classes, Zhang et al. proposed the mixup algorithm that mixes the training images from two different classes by linear interpolation and labels the mixed images according to the ratio between the two classes (Zhang et al., 2017). Exploiting the cutout and mixup algorithms, Yun et al. developed the cutmix algorithm that considers the completeness of image structure and the dependency between classes when generating mixed images (Yun et al., 2019).

3. Proposed training strategy

As described in Section I, the redundancy between FF-OCT images and the knowledge of SCC feature distribution can be exploited to reduce the computational complexity of model training for SCC classification. Therefore, it is essential to discuss the characteristics of the FF-OCT data. Before that, we describe how the data were collected from the FF-OCT system described in Section II-A. Then, we describe the proposed training strategy.

3.1. FF-OCT data characteristics

The FF-OCT data used in this work were collected from around 40 Friend Virus B NIH Jackson (FVB/N) female mice aging 6–8 weeks (Calderon-Delgado et al., 2021). We induced tumor growth in their back skin by combining an immunosuppressant solution with a tumor promoter (Hennings et al., 1993). We took normal samples from abdominal tissue, which was left untreated, and abnormal samples from excised back skin. Abnormal samples were further categorized into dysplasia and SCC. Inflamed, reddish back skin samples were categorized as dysplasia, whereas tumor samples that grow over 5 mm were categorized as SCC.

Three pre-processing steps were performed after data collection. Firstly, a 1-μm mean filter was applied to reduce the noise of tomograms. Secondly, tomograms were resized to obtain an isotropic voxel size of 0.5 μm. The last step padded the tomograms to a homogeneous size of $576 \times 256 \times 439$ (width $\times$ height $\times$ length) pixels. Each pre-processed tomogram corresponds to a tissue volume of physical size $288 \times 128 \times 219.5\ \mu m^3$. The pixel sampling period satisfies the Nyquist criterion. Each tomogram contains 439 cross-sectional images for SCC classification. The interval between neighboring cross-sectional images is 0.5 μm, but the diameter of cancer nuclei usually ranges from 20 μm to 30 μm. Consequently, a cancerous nucleus may appear in more than 40 cross-sectional images, and consecutive cross-sectional images of a tomogram may bear significant similarity. In other words, there exists significant redundant information between consecutive cross-sectional images.

Furthermore, the features relevant to SCC classification are not uniformly distributed. For example, features of the dermis layer, which is located deep into a cross-sectional image, are mostly irrelevant to SCC classification. Therefore, the lower part of a tomogram is not as contributive as the upper part to SCC classification. In contrast, the upper part of a tomogram contains rich relevant features, such as the thickness of stratum corneum and the size of cancer cells.

Therefore, it is feasible to improve the computational efficiency of model training by leveraging the feature distribution and data redundancy and by removing image regions irrelevant to SCC classification.

3.2. Training strategy

Our goal is to speed up model training. A training strategy here refers to a plan of action that makes the training of neural network models computationally feasible. For SCC classification, we need to optimize each candidate model on a large dataset and select among all candidate models the best one. The operations involved in this process must be performed in an efficient way.

An overview of our training strategy is depicted in Fig. 3. We crop the cross-sectional images to remove irrelevant features. An important consideration of the training strategy is that the region to be cropped should be carefully selected so that the epidermis and hence SCC features can be well preserved. In our design, we reduce the height and width of cross-sectional images by half if the performance drop due to the reduction is within 1%. The process continues for the remaining half until the performance drop exceeds 1%. This operation makes the CNN focus on regions that contain rich features for SCC classification and, in the meantime, reduces the memory usage for model training. It allows us to train more candidate classification models per graphical processing unit and quickly select the optimal model.

On the same design principle, we sub-sample images in each tomogram to reduce the amount of data to be processed. Noting that the reduction of training data may sacrifice data diversity and induce overfitting, we strike a balance between data diversity and training efficiency. Specifically, we monitor the performance drop while doubling the sampling interval. The process continues until the performance drop is greater than 1%.

These two data reduction operations save the computational time and make progressive model construction computationally manageable. In addition, since the features for SCC classification are largely preserved, the classification accuracy is maintained as much as possible.

4. Progressive model construction

We improve the regularization, channel attention, and optimization mechanisms of model construction. In particular, to cope with the characteristic of SCC categorization, we apply soft-labeling to refine the regularization and loss function. In addition, we adjust the architecture of SCC classifier to make it focus on key features of FF-OCT images. Finally, we fine tune the optimizer and the activation function to control the convergence process. The details of these operations are discussed in this section.

4.1. Soft Labeling

As described in Section II, there is not a clear cut between different SCC classes. Therefore, a soft labeling method is applied by using the cutmix as the regularization method to alleviate overfitting and by applying the smooth cross-entropy as the loss function to optimize the result.

As illustrated in Fig. 4, the concept of cutmix entails the generation of a mixed image from two images of distinct classes by cropping a region from one image and pasting it to the other image at the corresponding position. The label of the mixed image is determined by the ratio between the two regions of the mixed image, as shown in Fig. 4. This operation can be applied to generate images in between two image classes and to let a model learn the soft boundary between classes.

We also try to discourage the SCC classification model from being over-confident of its prediction by using smooth cross entropy as the loss function (Szegedy et al., 2016). Denote the three SCC class labels by 1 for normal, 2 for dysplasia, and 3 for SCC. Also denote the ground truth of an image by y, which is a triplet of three items. Initially, each item of y is either 0 or 1, and only one item is 1. Let y_i be an item of y, i ∈ {1, 2, 3}. During model training, the ground truth label is converted to a soft label. The value of y_i is updated by

$$y_i = \begin{Bmatrix} 1-\epsilon, & \text{if} \quad i=j \\ \frac{\epsilon}{2}, & \text{otherwise.} \end{Bmatrix}, \tag{1}$$

where ϵ is a small constant much less than 1, and j is the ground truth class label of y. This way, a hard label is converted to a soft label that is slightly deviated from the ground truth. The small perturbation introduced through this operation reduces the impact of images near the border of an SCC class on model training.

The loss function $L(y, y')$ is defined by the cross entropy of the ground truth y and the prediction y′. That is

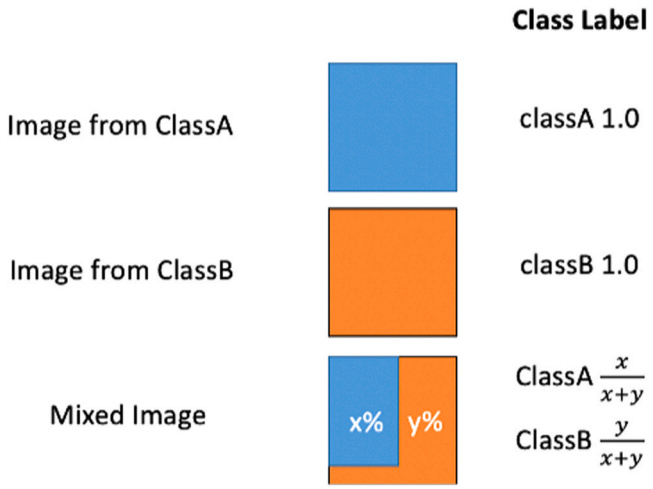

Fig. 4. Illustration of the cutmix algorithm. Regions belong to image classes A and B are shown in blue and orange, respectively.

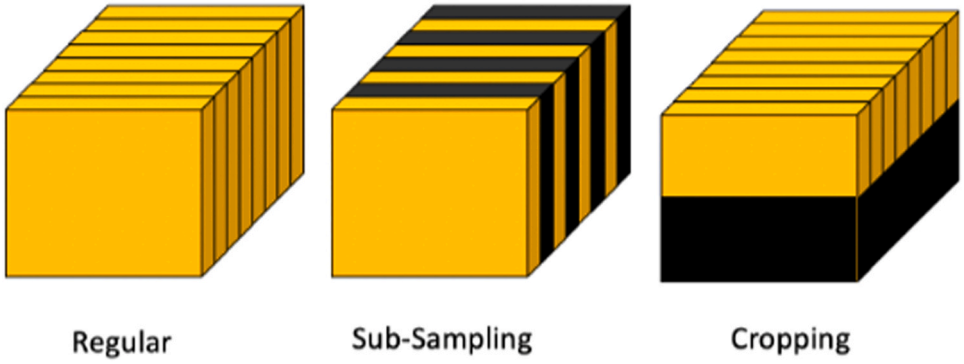

Fig. 3. Illustration of the proposed training strategy. Yellow parts represent the data used for training, while the black part represents the data discarded.

$$L(y, y') = -\sum_{i=1}^{3} y_i log(y'). \quad (2)$$

4.2. Model architecture adjustment

The architecture of SCC classification model determines its learnability and hence accuracy. Because of its low complexity, ResNet-18 is adopted as the backbone architecture of our model (He et al., 2016, 2019). It consists of four down-sampling blocks, each of which is divided into two paths, A and B, as shown in Fig. 5. Path A contains two consecutive 3 × 3 convolutional layers. One of them is responsible for reducing the width and height of the input tensor by one half, and the other performs a non-linear transformation without dimension reduction. Path B contains a 1 × 1 convolutional layer with a stride of 2 to halve the height and width of the input. The output of each down-sampling block is the sum of the tensors generated by its two paths.

The down-sampling process performed in path B overlooks three-quarters of the information from the input tensor. To fix the problem, we set the stride of the convolutional layer to 1 and insert an average pooling layer in front of it. In this way, all the information of the input tensor is conveyed in path B. The resulting model is called ResNet-18A in this paper.

In view of the impact of significant features on model training, we apply a channel attention mechanism to the SCC classification model. Specifically, we add an SE or an efficient channel attention (ECA) block described in Section II-C to every block of ResNet-18A. The architectures of SE and ECA are shown in Fig. 6. Consider an input tensor $T \in R^{H\times W\times C}$. A global average pooling (GAP) is performed in these two additional blocks to obtain a transformed tensor $GAP(T) \in R^{1\times 1\times C}$. Then, a scaling vector $S(T) \in R^{1\times 1\times C}$ is generated by two fully-connected networks in SE or by a 1D convolutional layer with a kernel size of 5 in ECA. The $S(T)$ rescales the input tensor T to generate an output tensor $O(T) \in R^{H\times W\times C}$. This operation recalibrates the features of T.

4.3. Convergence mechanism

The convergence mechanism of model training controls how an optimizer reaches a solution. A desirable convergence mechanism adapts to the dynamics of the loss function so that the solution can be effectively and efficiently found.

We achieve the adaptivity by increasing the decay momentum of the optimization process. That is, the beta parameter, which is between 0 and 1, of the Adam optimization algorithm is set to a small value (Kingma and Ba, 2014).

If ReLU is used as an activation function and if negative gradients are propagated through some neurons of the SCC classification model, these neurons may stop working and hence affect the convergence of model training. Therefore, we use PReLU instead, because it is robust to negative gradients [30].

5. Experiments

We conducted experiments to evaluate the performance of the proposed training strategy and progressive model construction described in Secs. III and IV. In this section, we first discuss the implementation details, including data partitioning, experimental setup, and metrics for performance evaluation. Then, we show the experimental results of the proposed training strategy and progressive model construction and investigate the relationship between model complexity and classification accuracy. Finally, we verify our selection of the sampling interval and check how our SCC classifier performs in comparison with a human expert.

5.1. Data partition

We partitioned the 3373 tomograms of the dataset described in Sec. I into three sets: training, validation, and testing. The testing set consisting of 677 tomograms was first selected from the dataset. Then, we divided the remaining tomograms into 10 subsets and performed a 10-fold cross validation procedure (Arlot and Celisse, 2010)]). In each

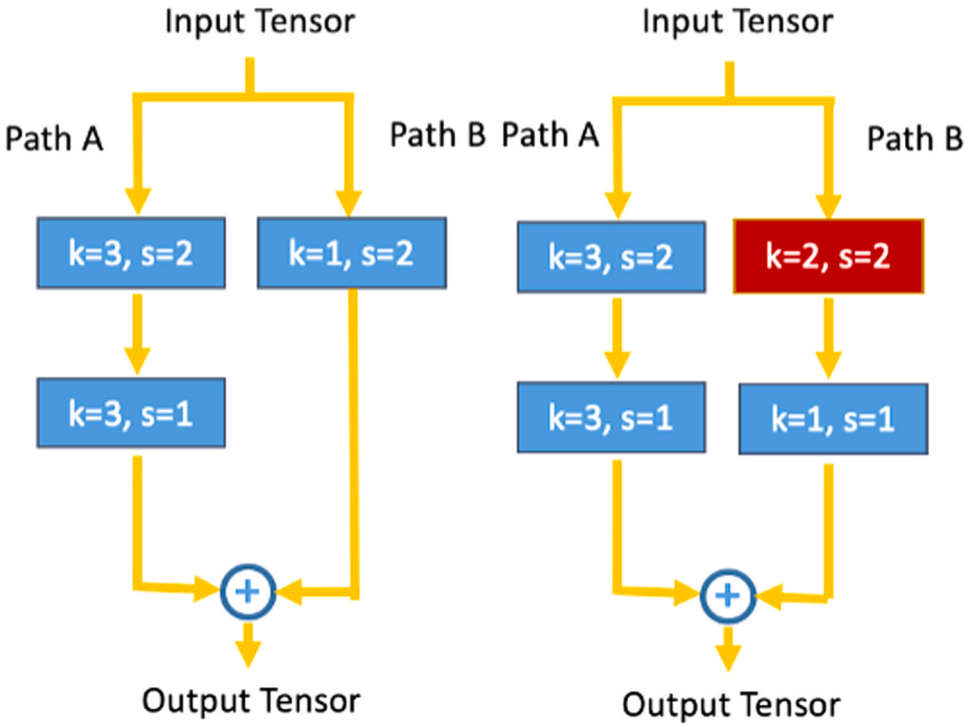

Fig. 5. Comparison between the down-sampling blocks of (left) a conventional ResNet18 and (right) an adjusted version of ResNet18. Blue and red boxes denote convolutional layers and average pooling layers, respectively. The kernel size (k) and strides (s) of each layer are shown in the diagrams.

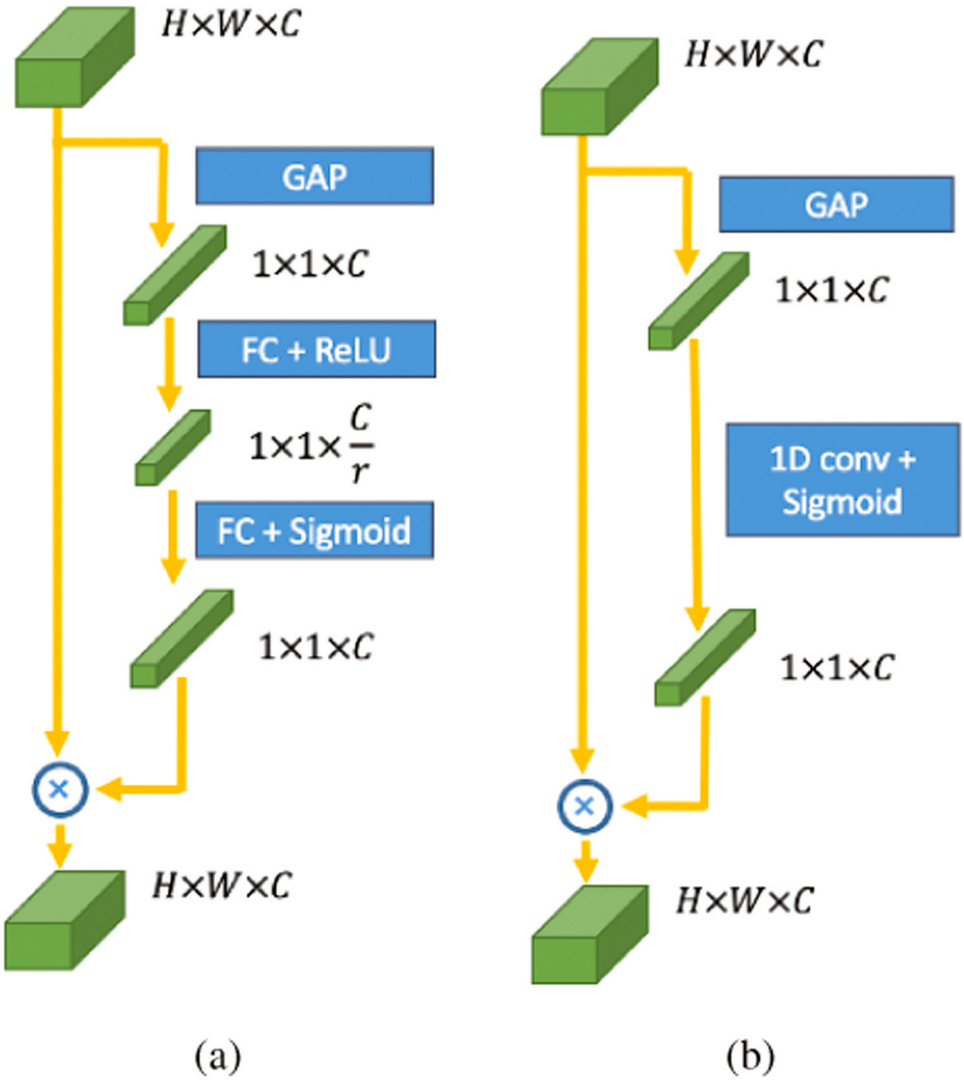

Fig. 6. Illustration of (a) SE and (b) ECA. The green volume denotes the tensor whose size (height, width, and channels) is shown. The multiplier performs element-wise multiplication. GAP: global average pooling; FC: fully-connected layer; 1D conv: 1D convolutional layer with kernel size 5.

step of the procedure, one of the subsets was chosen as the validation set and all the other subsets as the training set. This step continued 10 times for all subsets.

Recall that one to ten tomograms were collected from each tissue sample. To make the tomograms in different sets as independent of each other as possible, those collected from the same tissue sample were grouped in the same set.

5.2. Experimental setup

Our experiments were implemented on an NVIDIA Tesla V100 (NVIDIA, NVIDIA Tesla V100.Technical Report.⟨https://www.nvidia.com/en-gb/data-center/tesla-v100/⟩.) using PyTorch (Paszke et al., 2017), which is a deep learning framework for fast implementation. The batch size was set to 32 and the learning rate of the Adam optimizer was set to 0.001. We monitored the loss of SCC classification model evaluated on the validation set in each epoch and terminated the training process if the minimum loss did not decrease for 10 epochs.

5.3. Metric

In this work, we evaluate the performance of SCC classification model at both image and tomogram levels. For the image-level evaluation, we compute the average accuracy of the predictions generated by the SCC classification model for all images. Denote an input image by x_i and the function performed by the SCC classification model by $F(\cdot)$. Then, the prediction for x_i generated by the SCC classification model is $F(x_i) \in R^3$, and the formula for average accuracy is

$$A_I = \frac{1}{N}\sum_{i=1}^{N} 1(\text{argmax}_{j=1,2,3} F(X_i) = y_i), \tag{3}$$

where N denotes the number of images, y_i as defined in Eq. (1) denotes the class label of x_i, and $1(\cdot)$ denotes the 0–1 indicator function (Gerber et al., 2003).

For the tomogram-level evaluation, we take the average of the predictions for images in a tomogram and decide an aggregated prediction for the whole tomogram. Consider a tomogram X_k with 439 cross-sectional images $\{x_1, \ldots, x_{439}\}$. The class label prediction $F(X_k)$ of the tomogram is computed from the predictions $F(x_1)$,., and $F(x_{439})$ by

$$F(X_k) = \frac{1}{439}\sum_{i=1}^{439} F(x_i). \tag{4}$$

Then, the tomogram-level accuracy is obtained by

$$A_T = \frac{1}{M}\sum_{k=1}^{M} 1(\text{argmax}_{j=1,2,3}\, F(X_k) = Y_k), \quad (5)$$

where M denotes the number of tomograms in the dataset and Y_k denotes the ground truth label of the tomogram X_k.

The 10-fold validation generates 10 SCC classification models denoted by F_1,., and F_{10}. We evaluate their performance in two ways. One takes the average accuracy at both image and tomogram levels; the result is called overall accuracy. The other considers an ensemble SCC classification function $F_E(\cdot)$ obtained by

$$F_E(x_i) = \frac{1}{10}\sum_{k=1}^{10} F_k(x_i) \quad (6)$$

Then, the accuracy of the ensemble SCC classifier is computed by substituting $F_E(\cdot)$ for $F(\cdot)$ in Eqs. (3) and (5); the result is called ensemble accuracy. Totally, four different accuracy measurements (overall A_I, overall A_T, ensemble A_I, and ensemble A_T) of each SCC classification model are obtained.

5.4. Results of training strategy

Recall that the goal of the training strategy is to make model training computationally efficient. Table 1 shows the effect of image height reduction on model training. As we can see, the best result is obtained when the height is 128. When the height is 256, the bottom half of image may contain the features of dermis, which have little to do with SCC. Their appearance in the image affects the performance of SCC classification. On the other hand, when the image height is 64, the image does not contain sufficient epidermis features for SCC classification.

Table 2 shows the effect of image width reduction on model training. We can see that reducing the image width results in a performance drop. A possible reason for the performance drop is that DEJ may not present in the whole cross-section. As discussed in Sec. II-B, the depth of DEJ is an attribute of SCC. The DEJ is a curved segment along the horizontal axis. When we crop the image horizontally, the DEJ may become absent in the image, making the SCC classification model misjudge the actual depth of DEJ and resulting in a misclassification. Table 2 also shows that reducing the width in either direction yields the same effect, suggesting that the left and right halves of a tomogram are equally important to SCC classification.

Table 3 shows the results of model training using larger sampling intervals. We can see that the performance drop is contained within a 0.8% range when the sampling interval is less than or equal to 16. However, an abrupt performance drop occurs when the sampling interval increases from 16 to 32, at which the physical distance between neighboring images becomes 16 μm. This means that the cancer cell whose diameter is 20–30 μm can only appear in one or two consecutive images at most. Therefore, the sub-sampled tomogram may lack sufficient information of the cancer cell for SCC classification, resulting in poor classification accuracy.

5.5. Results of progressive model construction

All experiments on progressive model construction were performed by setting the sampling interval to 16 and the image height to 128. Table 4 shows the results of an ablation test on the cutmix. As we can see, without the cutmix, the accuracy of SCC classification drops 2%. This shows that the cutmix is an important component for training an SCC classifier. Without it, the model training would suffer from overfitting and hence result in poor classification accuracy.

Table 5 shows the results of model training using ResNet-18A and ResNet-18. We can see that the former has better accuracy because it attempts to preserve the features of FF-OCT images while downsampling the images. This is a distinct feature of our design.

Table 6 shows the contributions of channel attention to the SCC classification performance. We can see that the adoption of ECA and SE (Sec. IV-B) improves the classification accuracy of ResNet-18A, suggesting that reweighting FF-OCT image features is helpful. Table 6 also shows that ECA can better control the cross-channel interactions of features extracted by each residual block than SE.

The results of using different beta values in Adam for model training are shown in Table 7. We can see that similar SCC classification accuracy is obtained for beta = 0.1, 0.3, and 0.5. It should be noted that the default value 0.9 of beta yields the worst accuracy. Note that the classification accuracy varies more than 2% in this experiment, larger than the variations of all other experiments. Therefore, the beta value should be carefully selected.

Table 8 shows the results of model training using ReLU and PReLU. We find that PReLU performs consistently better than ReLU for ResNet-18A, although the difference is moderate. We can also see that the integration of all the refinement methods enables ResNet-18A to achieve 90.1% SCC classification accuracy at the tomogram level. This is the best result so far.

Table 1
Results of Model Training with Different Image Heights.

Image height	Overall		Ensemble	
	A_I	A_T	A_I	A_T
256	0.8584	0.8700	0.8654	0.8759
128	**0.8636**	**0.8874**	**0.8769**	**0.8936**
64	0.8100	0.8457	0.8245	0.8463

Table 2
Results of Model Training with Different Image Widths.

Image width	Overall		Ensemble	
	A_I	A_T	A_I	A_T
576	**0.8636**	**0.8874**	**0.8769**	**0.8936**
288 (L)	0.8339	0.8753	0.8511	0.8803
288 (R)	0.8377	0.8735	0.8504	0.8744

Table 3
Results of Model Training Using Different Sampling Interval.

Sampling interval	Overall		Ensemble	
	A_I	A_T	A_I	A_T
1	0.8636	0.8874	0.8769	**0.8936**
2	**0.8656**	**0.8924**	**0.8773**	**0.8936**
4	0.8649	0.8849	0.8749	0.8862
8	0.8588	0.8901	0.8714	0.8921
16	0.8610	0.8821	0.8710	0.8862
32	0.8468	0.8700	0.8579	0.8729

Table 4
Accuracy of SCC Classification Before and After the Cutmix is Removed.

Cutmix	Overall		Ensemble	
	A_I	A_T	A_I	A_T
with	**0.8610**	**0.8821**	**0.8710**	**0.8862**
without	0.8428	0.8809	0.8585	0.8833

Table 5
Accuracy of Adjusted and Original ResNet-18.

Cutmix	Overall		Ensemble	
	A_I	A_T	A_I	A_T
ResNet-18A	**0.8659**	**0.8889**	**0.8710**	**0.8862**
ResNet-18	0.8428	0.8809	0.8585	0.8833

Table 6
Results of Model Training Using Different Channel Attention Mechanisms.

Mechanism	Overall		Ensemble	
	A_I	A_T	A_I	A_T
without	0.8659	0.8889	0.8710	0.8862
SE	0.8676	**0.8961**	0.8774	0.8995
ECA	**0.8689**	0.8954	**0.8781**	**0.9010**

Table 7
Accuracy of SCC Classifier Using Different Beta Values of the Optimizer for Model Training.

Beta	Overall		Ensemble	
	A_I	A_T	A_I	A_T
0.9	0.8494	0.8895	0.8623	0.8892
0.7	0.8669	0.8927	0.8748	0.8951
0.5	0.8689	**0.8954**	0.8781	**0.9010**
0.3	**0.8704**	0.8930	**0.8799**	0.8980
0.1	0.8698	0.8939	0.8790	0.8956

Table 8
Results of Model Training using Different Activation Functions.

Activation function	Overall		Ensemble	
	A_I	A_T	A_I	A_T
ReLU	0.8689	0.8954	0.8781	**0.9010**
PReLU	**0.8712**	**0.8974**	**0.8792**	**0.9010**

5.6. Model complexity

An experiment was performed to investigate how the complexity of the SCC classification model relates to the classification accuracy. In this experiment, we applied the proposed approach to train and refine ResNet-50 and ResNet-101, both are deeper than ResNet-18A. All the refinements are implemented on ResNet-50 and ResNet-101, so the only difference is the depth of the network. Deeper models have a higher complexity. We can see from the results shown in Table 9 that increasing the model complexity does not necessarily enhance the classification accuracy. In fact, the performance of the ResNet-18A is as good as ResNet-50. In view of the trade-off between complexity and performance, ResNet-18A is a good choice.

5.7. Retraining with small sampling intervals

The sampling interval 16 selected in Sec. V-D is the result of a tradeoff between classification accuracy and computational efficiency of model training. To verify the appropriateness of this selection, we retrain the refined SCC classification model with different sampling intervals smaller than the selected value. A smaller sampling interval means a higher sampling rate.

Table 10 shows the SCC classification accuracy of retraining the refined model using different sampling intervals. We can see that, in general, using smaller sampling intervals slightly improves the classification accuracy; however, the improvement is only 0.56% at best. Therefore, we are confident that the sampling interval 16 is an appropriate choice.

Table 9
Results of Model Training Using Different Depths.

Model Depth	Overall		Ensemble	
	A_I	A_T	A_I	A_T
18	**0.8712**	**0.8974**	0.8792	**0.9010**
50	0.8701	0.8909	**0.8821**	0.8980
101	0.8674	0.8870	0.8798	0.8966

Table 10
Results of the Refined SCC Classifier Using Different Sampling Interval.

Sampling interval	Overall		Ensemble	
	A_I	A_T	A_I	A_T
1	0.8682	0.8867	**0.8848**	0.8906
2	0.8706	0.8902	0.8795	0.8936
4	**0.8726**	0.8952	0.8789	**0.9010**
8	0.8683	0.8834	0.8789	0.8906
16	0.8712	**0.8974**	0.8792	**0.9010**
32	0.8502	0.8836	0.8665	0.8921

5.8. Comparison with human performance

We also asked a medical expert with over 10 year experience on OCT imaging to classify the same testing tomograms. The training and testing data were partitioned as described in Sec. V-A. Before the test, the expert was given a chance to get familiar with the training data and perform labeling tests on the training data. In labeling tests, each tomogram was presented as a collection of ten cross-sectional images uniformly distributed along the two lateral directions. This human training process continued until the expert surpassed 90% classification accuracy. Then, the expert was asked to label each of the 677 testing tomograms. Among these, the human expert made a correct prediction for 569 tomograms and a wrong prediction for 103 tomograms, leaving the other 5 tomograms unpredicted. Thus, the classification accuracy of the expert is 84.04%. Comparing it with the classification accuracy 90.1% of our approach, we can see that our SCC classifier is superior to the human expert.

6. Discussion

In this section, we discuss the difference between the image-level accuracy and the tomogram-level accuracy. We also discuss why transfer learning and knowledge distillation are not applicable to SCC classification.

6.1. Accuracy

As described in Sec. V-C, we measure the classification accuracy at both image and tomogram levels. We note that the tomogram-level accuracy is always higher than the image-level accuracy. The existence of this difference can be explained by an example tomogram shown in Fig. 7. The ground truth class label of this tomogram is dysplasia. The first 100 images of the tomogram are noisy and contain sparse SCC features, as shown in the top image of Fig. 7. Therefore, the predictions for those images are inaccurate. For the other images of the tomogram, the structure of DEJ is clearly presented, as shown in the bottom image of Fig. 7. The SCC classifier makes an accurate prediction for these images. For this tomogram, the model achieves 75.85% image-level accuracy and 100% tomogram-level accuracy. This example shows that, as long as the tomogram receives a majority of the correct predictions, the incorrect predictions do not affect the final tomogram-level accuracy. In other words, aggregating predictions from different images effectively alleviates the impact of noise on classification accuracy. This is why the tomogram-level accuracy is always higher than the image-level accuracy.

6.2. Transfer learning

As described in Section I, transfer learning is a popular training strategy. It facilitates model training by using the weights of another model pre-trained in other tasks to initialize the model to be trained. In this way, model training does not have to start from scratch. However,

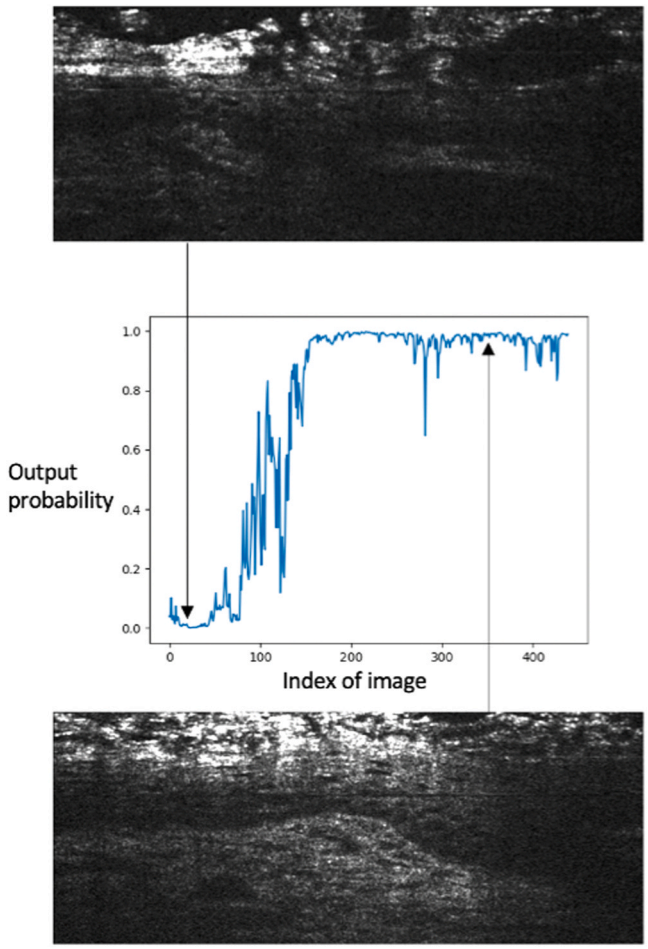

Fig. 7. Output probability of dysplasia (ground truth class label) obtained from a tomogram that contains noisy (top) and clear (bottom) images. The index of image was given according to the scanning position of the horizontal axis.

transfer learning is not suitable for the problem considered in this work for two reasons. The first reason is that the features of SCC are unique; therefore, the feature extraction learned from other tasks is not helpful for the training of an SCC classifier. The second reason is that most transfer learning methods are designed for RGB images, but FF-OCT images are grayscale, making transfer learning inappropriate for SCC classification.

6.3. *Knowledge distillation*

Study has shown that knowledge distillation (Hinton et al., 2015) is a common technique for model construction. However, it is not applicable to SCC classification. Using a deep network to enhance the learnability of the shallow network does not work for the model training considered in this work because the assumption that classification accuracy increases with model complexity does not hold for SCC classification, as suggested by the results shown in Table 9.

7. Conclusion

Squamous cell carcinoma classification from FF-OCT images is an elaborative process even for a well-trained physician. In this paper, we have described an efficient training strategy that takes the characteristics of FF-OCT data into consideration and achieves a significant reduction of computation time required for model training. It is an essential step that makes the progressive model construction of CNN-based SCC classification computationally feasible. The performance of our proposed model construction is attributed to the consideration of both morphological and cellular characteristics of squamous cell carcinoma in the design.

We have also provided a thorough analysis of the effectiveness of the proposed training strategy and progressive model construction through various experiments. We believe the lessons learned from this work and the techniques developed herewith are useful for other medical applications of machine learning as well.

CRediT authorship contribution statement

Chi-Jui Ho: Algorithm design and test; paper drafting. **Manuel Calderon-Delgado:** Software implementation and data analysis. Paper revision. **Ming-Yi Lin:** Data collection. **Jeng-Wei Tjiu:** Data collection. **Sheng-Lung Huang:** Project development and supervision; paper revision. **Homer H. Chen:** Project development and supervision. Paper revision and finalization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Ministry of Science and Technology of Taiwan under Contracts 107-2634-F-002-017 and 103-2325-B-002-044 and by National Taiwan University under Contract 109L891707. The authors would like to thank the technical support of the National Center for High-Performance Computing, Taiwan.

References

Arlot, S., Celisse, A., et al., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79.

Calderon-Delgado, M., Lin, M.Y., Tjiu, J.W., Huang, S.L., 2021. OCT-MoS, a dataset of mouse skin squamous cell carcinoma stages by full-field optical coherence tomography. Image Data Resource, idr0098. https://doi.org/10.17867/10000155. https://idr.openmicroscopy.org/webclient/?show=project-1605.

Chan, J.K., 2014. The wonderful colors of the hematoxylin-eosin stain in diagnostic surgical pathology. Int. J. Surg. Pathol. 22, 12–32.

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2015.Fast and accurate deep network learning by exponential linear units (elus). arXiv: ⟨http://arXiv.org/abs/arXiv:1511.07289⟩.

Dalimier, E., Salomon, D., 2012. Full-field optical coherence tomography: a new technology for 3d high-resolution skin imaging. Dermatology 224, 84–92.

DeVries, T., Taylor, G.W., 2017.Improved regularization of convolutional neural networks with cutout. arXiv:.⟨http://arXiv.org/abs/arXiv:1708.04552⟩.

Drexler, W., Liu, M., Kumar, A., Kamali, T., Unterhuber, A., Leitgeb, R.A., 2014. Optical coherence tomography today: speed, contrast, and multimodality. J. Biomed. Opt. 19, 071412.

Dubois, A., Boccara, A.C., 2008. Full-field optical coherence tomography. In: Optical Coherence Tomography. Springer, pp. 565–591.

Gerber, H.U., Leung, B.P., Shiu, E.S., 2003. Indicator function and hattendorff theorem. N. Am. Actuarial J. 7, 38–47.

Gollakota, S., Hassanieh, H., Ransford, B., Katabi, D., Fu, K., 2011. They can hear your heartbeats: non-invasive security for implantable medical devices. Proc. ACM SIGCOMM 2011 Conf. 2–13.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proc. IEEE Int. Conf. Comput. Vis. 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 770–778.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., Li, M., 2019. Bag of tricks for image classification with convolutional neural networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 558–567.

Hennings, H., Glick, A.B., Lowry, D.T., Krsmanovic, L.S., Sly, L.M., Yuspa, S.H., 1993. Fvb/n mice: an inbred strain sensitive to the chemical induction of squamous cell carcinomas in the skin. Carcinogenesis 14, 2353–2358.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. Statics 1050, 9.

Ho, C.J., Calderon-Delgado, M., Chan, C.C., Lin, M.Y., Tjiu, J.W., Huang, S.L., Chen, H. H., 2021. Detecting mouse squamous cell carcinoma from submicron full-field optical coherence tomography images by deep learning. J. Biophoton. 14, e202000271.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 7132–7141.
Huang, D., Swanson, E.A., Lin, C.P., Schuman, J.S., Stinson, W.G., Chang, W., Hee, M.R., Flotte, T., Gregory, K., Puliafito, C.A., et al., 1991. Optical coherence tomography. science 254, 1178–1181.
Izatt, J.A., Choma, M.A., 2008. Theory of optical coherence tomography. In: Optical Coherence Tomography. Springer, pp. 47–72.
Kalkman, J., 2017. Fourier-domain optical coherence tomography signal analysis and numerical modeling. Int. J. Opt.
Kingma, D.P., Ba, J., 2014.Adam: A method for stochastic optimization. arXiv: ⟨http://arXiv.org/abs/arXiv:1412.6980⟩.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. Imagenet classification with deep convolutional neural networks. Commun. ACM 60, 84–90.
Lee, P., Chan, C., Huang, S., Chen, A., Chen, H.H., 2016. Blood vessel extraction from oct data by short-time rpca. 2016 IEEE Int. Conf. Image Process. 394–398. https://doi.org/10.1109/ICIP.2016.7532386.
Lee, P.H., Chan, C.C., Huang, S.L., Chen, A., Chen, H.H., 2018. Extracting blood vessels from full-field oct data of human skin by short-time rpca. IEEE Trans. Med. Imaging 37, 1899–1909.
Mekonnen, B.K., Tsai, D.F., Hsieh, T.H., Yang, F.L., Liaw, S.K., Huang, S.L., 2019. Deep learning approach for red blood cell segmentation from full-field oct data of human skin. 2019 IEEE Int. Conf. 1–2. https://doi.org/10.1109/BioPhotonics.2019.8896748.
Muzic, J.G., Schmitt, A.R., Wright, A.C., Alniemi, D.T., Zubair, A.S., Lourido, J.M.O., Seda, I.M.S., Weaver, A.L., Baum, C.L., 2017. Incidence and trends of basal cell carcinoma and cutaneous squamous cell carcinoma: a population-based study in olmsted county, minnesota, 2000 to 2010. Mayo Clin. Proc. 890–898 (Elsevier).
Nair, V., Hinton, G.E., 2010.Rectified linear units improve restricted boltzmann machines, in: ICML.
NVIDIA, NVIDIA Tesla V100.Technical Report. ⟨https://www.nvidia.com/en-gb/data-center/tesla-v100/⟩.
Olsen, J., Holmes, J., Jemec, G.B., 2018. Advances in optical coherence tomography in dermatology-a review. J. Biomed. Opt. 23, 040901.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Autom. Differ. Pytorch.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al., 2015. Imagenet large scale visual recognition challenge. Int. J. Comput. Vis. 115, 211–252.
Scholler, J., Groux, K., Goureau, O., Sahel, J.A., Fink, M., Reichman, S., Boccara, C., Grieve, K., 2020. Dynamic full-field optical coherence tomography: 3d live-imaging of retinal organoids. Light 9, 1–9.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 2818–2826.
Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C., 2018. A survey on deep transfer learning. In: International Conference on Artificial Neural Networks. Springer, pp. 270–279.
Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. Int. Conf. Mach. Learn. 6105–6114.
Tsai, C.C., Chang, C.K., Hsu, K.Y., Ho, T.S., Lin, M.Y., Tjiu, J.W., Huang, S.L., 2014. Full-depth epidermis tomography using a mirau-based full-field optical coherence tomography. Biomed. Opti. Express 5, 3001–3010.
Wang, K.X., Meekings, A., Fluhr, J.W., McKenzie, G., Lee, D.A., Fisher, J., Markowitz, O., Siegel, D.M., 2013. Optical coherence tomography-based optimization of mohs micrographic surgery of basal cell carcinoma: a pilot study. Dermatol. Surg. 39, 627–633.
Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q., 2020a. Eca-net: Efficient channel attention for deep convolutional neural networks. Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogni. 11534–11542.
Wang, Y., Zhang, J., An, C., Cavichini, M., Jhingan, M., Amador-Patarroyo, M.J., Long, C. P., Bartsch, D.U.G., Freeman, W.R., Nguyen, T., 2020b. A segmentation based robust deep learning framework for multimodal retinal image registration. IEEE Int. Conf. Acoust. Speech Signal Process. IEEE 1369–1373.
Wells, W.A., Barker, P.E., MacAulay, C.E., Novelli, M., Levenson, R.M., Crawford, J.M., 2007. Validation of novel optical imaging technologies: the pathologists' view. J. Biomed. Opt. 12, 051801.
World Cancer Research Fund, American Institute for Cancer, 2019.Skin cancer report. Technical Report. ⟨https://www.wcrf.org/dietandcancer/skin-cancer⟩.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. Proc. IEEE Conf. Comput. Visi. Pattern Recognit. 1492–1500.
Xiong, Y.Q., Mo, Y., Wen, Y.Q., Cheng, M.J., Huo, S.T., Chen, X.J., Chen, Q., 2018. Optical coherence tomography for the diagnosis of malignant skin tumors: a meta-analysis. J. Biomed. Opt. 23, 020902.
Xu, B., Wang, N., Chen, T., Li, M., 2015.Empirical evaluation of rectified activations in convolutional network. arXiv: ⟨http://arXiv.org/abs/arXiv:1505.00853⟩.
Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y., 2019. Cutmix: regularization strategy to train strong classifiers with localizable features. Proc. IEEE Int. Conf. Comput. Vis. 6023–6032.
Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2017.mixup: Beyond empirical risk minimization. arXiv: ⟨http://arXiv.org/abs/arXiv:1710.09412⟩.
Zhang, J., An, C., Dai, J., Amador, M., Bartsch, D.U., Borooah, S., Freeman, W.R., Nguyen, T., 2019. Joint vessel segmentation and deformable registration on multi-modal retinal images based on style transfer. IEEE Int. Conf. Image Process. IEEE 839–843.