#### **ORIGINAL ARTICLE**



# A layer-level multi-scale architecture for lung cancer classification with fluorescence lifetime imaging endomicroscopy

Qiang Wang<sup>1</sup> · James R. Hopgood<sup>2</sup> · Susan Fernandes<sup>1</sup> · Neil Finlayson<sup>2</sup> · Gareth O. S. Williams<sup>1</sup> · Ahsan R. Akram<sup>1</sup> · Kevin Dhaliwal<sup>1</sup> · Marta Vallejo<sup>3</sup>

Received: 7 October 2021 / Accepted: 26 May 2022 / Published online: 25 June 2022  $\circledcirc$  The Author(s) 2022

#### Abstract

In this paper, we introduce our unique dataset of fluorescence lifetime imaging endo/microscopy (FLIM), containing over 100,000 different FLIM images collected from 18 pairs of cancer/non-cancer human lung tissues of 18 patients by our custom fibre-based FLIM system. The aim of providing this dataset is that more researchers from relevant fields can push forward this particular area of research. Afterwards, we describe the best practice of image post-processing suitable per the dataset. In addition, we propose a novel hierarchically aggregated multi-scale architecture to improve the binary classification performance of classic CNNs. The proposed model integrates the advantages of multi-scale feature extraction at different levels, where layer-wise global information is aggregated with branch-wise local information. We integrate the proposal, namely ResNetZ, into ResNet, and appraise it on the FLIM dataset. Since ResNetZ can be configured with a shortcut connection and the aggregations by *Addition* or *Concatenation*, we first evaluate the impact of different configurations on the performance. We thoroughly examine various ResNetZ variants to demonstrate the superiority. We also compare our model with a feature-level multi-scale model to illustrate the advantages and disadvantages of multi-scale architectures at different levels.

**Keywords** Convolutional neural networks  $\cdot$  Fluorescence lifetime imaging endomicroscopy  $\cdot$  Lung cancer classification  $\cdot$  Multi-scale feature extraction  $\cdot$  Hierarchically aggregated architectures  $\cdot$  ResNetZ

#### 1 Introduction

Fluorescence lifetime is characterized by a decay from the excited state to the ground state, which is independent of fluorescence concentration but sensitive to the biological environment [1]. Fluorescence lifetime imaging endo/microscopy (FLIM) utilizes lifetime contrast between heal-thy/unhealthy biological tissue to distinguish them effectively. Due to the independence, typically, lifetime images are more homogeneous than intensity images which show higher contrast. This introduces challenges for visual recognition. For example, when imaging the same physical point on tissue with different hardware configurations, lifetime images are usually visually indistinguishable compared to intensity images. Jo et al. [2] reported that oral cancer has a shorter lifetime, whereas McGinty et al. revealed that tumours have a longer lifetime [3]. In

addition, other hardware factors, such as excitation bandwidth (wavelength) and exposure time, also affect lifetime derivation [4]. We similarly observed that, as wavelength increases, the contrast or difference in lifetime values between a pair of normal/cancerous tissue becomes so small that, although there is a classification boundary, it cannot be *a priori* deduced which tissue state has the lower lifetime.

Machine Learning (ML), particularly Deep Learning (DL), has revolutionized biomedical image processing in many aspects, such as in classification and segmentation [5]. However, little attention has been paid to the application of ML/DL to FLIM images in, for example, the automatic discrimination of cancer. Existing approaches in this area usually employ conventional ML algorithms with handcrafted features, which inevitably requires professional knowledge on feature engineering. For instance, Chen et al. [6] apply support vector machines (SVM) for skin lesion detection with artificial features retrieved from

Extended author information available on the last page of the article

lifetime decay fitting parameters. For DL, the situation is even worse: there is very limited research concerning DL in FLIM-based cancer differentiation, apart from a few preliminary studies [7]. Unlike ML, which can perform well on small-scale data, DL usually requires large-scale datasets for effective learning without over-fitting. Unfortunately, there is no such dataset publicly available in this domain, which severely affects the development and application of FLIM.

Contemporary architectures, such as ResNet [8] and DenseNet [9], have advanced the state-of-the-art of classification performance significantly. A common practice in those models is the modularization of convolution blocks, particularly the usage of bottleneck blocks. Those relatively fixed patterns enable them to be easily expanded with more sophisticated blocks, and multi-scale architectures are prevalent among the expansions. The primary disadvantage of ResNet is that it produces many redundant features but struggles to create new features [9, 10]. Different strategies could be applied to avoid this effect. For example, DenseNet [9] employs very narrow networks to reduce the redundancy and dense aggregation for new feature creation, and Res2Net [11] splits redundant features and applies a hierarchical multi-scale module to create new features per the separated features. Due to the split, however, Res2Net is unable to retain the correlations among input features as global information since the grouped features are handled separately.

Here, we introduce our unique dataset of over 100, 000 FLIM images from 18 pairs of normal/cancerous tissues of 18 patients. The images were collected by a custom FLIM system [12, 33] aiming for online in-vivo in-situ lung disease diagnostics, with various user-specified configurations. The dataset consists of multi-dimensional images rich in spatial and spectral information, which can reflect the diversity of fluorescence lifetime to a large extent. Afterwards, we describe the image post-processing procedure, which applies intensity image as a soft weight to the corresponding lifetime images. With this, lifetime independence from its intensity can be addressed, increasing the classification performance of classic CNNs. To further improve the discrimination and address the broad spread of correlating pixels with similar lifetime values in lifetime images, we propose a hierarchically aggregated multi-scale architecture at a layer-level, namely ResNetZ. We integrate the model into ResNet, and evaluated the performance on three aspects, including the impact of a shortcut connection and different aggregations; the overall performance with state-of-the-art CNNs and ResNet variations; and the comparison between Res2Net and ResNetZ. Accuracy, precision, recall, the area under the receiver operating characteristic (ROC) curve (AUC), and Cohen's Kappa [13] were used as metrics.

The rest of the paper is organized as follows. Section 2 reviews the related work in FLIM for cancer classification and multi-scale architectures. Section 3 introduces the technical details of our method. Experimental results are presented in Sect. 4 and discussed in Sect. 5, followed by the conclusion and future work in Sect. 6.

#### 2 Related work

#### 2.1 FLIM in cancer classification

As shown in Fig. 1, a common practice is to derive the averaged lifetime by histogramming and discriminate cancer based on lifetime difference, with the assistance of histological images. Here, cancer tissue has an average lifetime of 1.48 ns, while a non-cancer sample has an average lifetime of 1.9 ns. Little effort has been made on automatic classification of cancer using ML algorithms on FLIM images. Gu et al. [14] utilized a feed-forward neural network-based extreme learning for the diagnostic of early cervical cancer using FLIM on H &E stained samples, with expert-engineered features. Cuenca et al. [15] and Jo et al. [2] applied a quadratic discriminant analysis binary classifier for distinguishing oral cancer and dysplasia, with six handcrafted features extracted from FLIM images. In [6], Chen et al. deployed a SVM model to distinguish nonmelanoma skin lesions, where features were engineered from lifetime reconstruction. Marsden et al. [16] applied ML technologies for intraoperative cancer margin assessment with FLIM, where a dual-path architecture retrieved information at different scales for predicting the point-wise probability of cancer. Nonetheless, all those works investigated conventional ML methods with engineered features, and none of them concerns lung cancer classification using DL.

Considerable effort has been made by the authors to investigate the automatic classification of ex-vivo lung cancer from FLIM images. In [17], we applied four popular ML methods to FLIM images for ex-vivo lung cancer classification, namely K-nearest neighbour, SVM, neural network, and random forest. A significant difference between our approach and the existing ones is that we applied pixel values of lifetime images as features, instead of artificial ones. Later, we investigated the feasibility of traditional CNNs for the same classification problem [7]. With five classic CNNs, i.e. ResNet, ResNeXt, DenseNet, Inception, and Xception, the results were dramatically better than the ML-based solutions. We also showed that integrating lifetime with intensity information can achieve better results than using lifetime images only. We further expanded the study by embedding dilated convolutions to multi-scale technologies [18, 19]. Comparing with previous



Fig. 1 Lifetime contrast of non-cancerous (row 1) and cancerous (row 2) lung tissue using histograms (column 3) of lifetime images (column 2) obtained from intensity images (column 1), along with histological images (column 4) as the ground truth [7]

studies, this one does not use dilated convolutions as, empirically, they do not contribute much to performance improvement on the dataset. In addition, this study thoroughly compares the performance of different configurations to understand the impacts of the configurations on the results. Meanwhile, we also introduce the optimal image post-processing to maximise the performance improvement.

#### 2.2 Multi-scale architectures

Multi-scale architectures have become very popular in contemporary CNNs, which are usually epitomized by employing a number of single/composite operations in parallel at different levels. Typical examples include a multi-path CNN for brain image segmentation [20], Inception, using several parallel convolution branches at a layer-level [21], and Res2Net with a hierarchical featurelevel multi-scale model. One reason for their success is their capability to simultaneously extract features at different scales and, later, integrate the multi-scale features together, so that more information passes through their backbone networks.

Architecture-level multi-scale strategies are usually developed to deal with multiple inputs or for special purposes. Setio et al. [22] proposed a multi-view model to decline false-positive cases in pulmonary nodule detection. Moeskops et al. applied a multi-path architecture for magnetic resonance brain image segmentation [20]. Despite their success, the major problem of architecturelevel multi-scale models is that the underlying ideas are usually problem-specific and, hence, it is challenging to migrate them to other architectures.

Layer-level models concern the features extracted after each layer as a whole and utilize more sophisticated operations to process the information. The operations could be simple operators, such as multiple parallel convolutions in Inception [21, 23], or a set of complicated ones, e.g. multiple dilated convolutions in densely connected blocks [24]. In addition, they can substitute a few or the entire original operations. For example, DeepLab [25] used an atrous spatial pyramid pooling with several parallel dilated convolutions for better semantic segmentation. Due to the modularization, layer-level styles are usually easy to be integrated into other backbone networks with similar architectures, such as the inception-like convolution blocks in DRINET [26]. With the paralleling, more features different in space, scale, and context can be retrieved. However, an apparent disadvantage is the increase of complexity due to parallel operations.

Feature-level multi-scale styles are normally characterized by splitting input features into groups, processing grouped features individually, and fusing processed features. The operations can be committed by group convolutions [27], depthwise separable convolution and its variations [28, 29], or pointwise group convolutions [30]. More sophisticated operations can be integrated into feature-level architectures. For example, Res2Net proposed a hierarchical aggregation into the processing, and ResNeSt [31] introduced split-attention into group convolutions. Besides the advantages of multi-scale models, feature-level architectures significantly decrease the complexity, compared with the conventional convolutions. However, due to the separation of the input features, the correlations among the features are partially ignored. Our inspiration comes from the hierarchical style reported in [11] and [32], except that we incorporate our model at a layer-level instead of at a feature-level, to retain the correlations among features, which can be further reused and aggregated. In addition, our ResNetZ architecture also introduces other possible configurations, such as *Concatenation* aggregation rather than *Addition* applied in [11] and [32]. This is explored further in Sect. 4.1.

## 3 Methodology

The FLIM dataset was gathered by a continuous collection of *ex-vivo* human lung tissue using our custom-built FLIM system. Raw FLIM images contain a certain level of noise for visual recognition, and thus post-processing is required before being fed into the CNNs for classification. It is worth noting that our intention is to introduce the best practice we have learned so far on the FLIM dataset for reproducible research. The overall procedure is depicted in Fig. 2, and technical details will be addressed in depth in the following sections.

#### 3.1 Data collection

A custom fibre-based FLIM system was deployed to acquire data with various user-specified configurations, including different exposure time and two spectral bands [33]. For online imaging and diagnostic purposes, our custom FLIM imaging system recorded sequences of lifetime images with a resolution of  $128 \times 128$  pixels, at a frame rate of 9 frames per second, which were aggregated across a line sensor of single-photon detectors [12]. Each frame contains four images, yielding an intensity and the corresponding lifetime image for each of the two customizable spectral bands. Lifetime values can be reconstructed by different algorithms, such as the Rapid Lifetime Determination method (RLD) [34]. Data Collection in Fig. 2 depicts an example of the experiment workflow, where a lung tissue was fixed on a corkboard and the 128  $\times$ 128 images of autofluorescence intensity and lifetime were reconstructed with an exposure time 6 µs, a spectral band of 498-570 nm, and the RLD decay fitting approach. These settings were chosen to represent intended characteristic conditions for future clinical trials.

For each ex-vivo experiment, a pair of cancerous/noncancer tissue from each patient was scanned using direct contact between the fibre and tissue, and multiple measurements were extracted at different physical points on each tissue to enrich the variety of the images on the same tissue. Over 100,000 raw FLIM images were collected from 18 pairs of lung normal/cancerous tissues. For this study, we removed some images which may introduce extra variance. For example, we excluded images whose lifetime was not reconstructed by RLD, since their lifetime is significantly different. After cleaning there were, in total, 61,816 FLIM images remaining, including 25,372 from cancerous tissue and 36,444 from normal tissue. The detailed information regarding the remaining images is listed in Table 1. Note that each frame contains an intensity and its corresponding lifetime image.

#### 3.2 Image post-processing

The raw images collected were very noisy (see the grayscale images at the right of *Data Collection* in Fig. 2). In order for the images to be suitable for human and machine perception, post-processing is therefore needed. The overall post-processing is depicted in Fig. 2, *image post-processing*. One of the criteria for achieving reasonable postprocessing results is to ensure that the histograms derived from the averaged lifetime remain unchanged, while the images become visually plausible, as shown in Fig. 1.

Given a relatively short exposure time, e.g. 20 µs, the total number of recorded photons per pixel is usually in the order of 100 – 1000, or even lower, where the pixel record is easily affected by photon quantum noise [35]. In order for the signals to be recorded and processed efficiently, optimal signal–noise ratio (SNR) of fluorescence intensity measurement is required. In this study, we utilize a threshold value of  $\sqrt{\hat{N}}$  to approximate SNR [35], where  $\hat{N}$  is the mean of the measured fluorescence concentration. It is, therefore, assumed that pixel intensity larger than  $\sqrt{\hat{N}}$  is essential for a lifetime derivation with acceptable accuracy.

Let  $I^{I} = \{i_{x,y}^{I} | i_{x,y}^{I} \ge 0 \text{ and } x, y \in [0, M]\}$  denote an intensity image  $I^{I}$  with size of  $M \times M$ , and  $I^{L} = \{i_{x,y}^{L} | i_{x,y}^{L} \ge 0 \text{ and } x, y \in [0, M]\}$  denote the corresponding lifetime image  $I^{L}$  with size of  $M \times M$ . The denoising approach can be defined as [7]:

$$i_{x,y}^{I} = \begin{cases} 0 & i_{x,y}^{I} \le \sqrt{\hat{N}} \\ i_{x,y}^{I} & \text{otherwise} \end{cases}$$
(1a)

Fig. 2 Schematic diagram of the proposed method adapted from [7]. Raw FLIM images were collected on an ex-vivo lung tissue fixed on a corkboard (Step 1). Raw images were then post-processed to obtain FLIM images suitable for the classification (Step 2). Finally, all processed images were input into CNN models for binary classification purposes (Step 3)



$$\hat{i}_{x,y}^{L} = \begin{cases} 0 & i_{x,y}^{L} \le \sqrt{\hat{N}} \\ i_{x,y}^{L} & \text{otherwise} \end{cases}$$
(1b)

Next, the intensity images are normalized with dark background D and lightfield images L, adapted from [36]:

$$\vec{I}^{I} = \frac{G * (\vec{I}^{I} - D)}{G * (L - D)}$$
(2)

where \* denotes the 2D convolution operator,  $\hat{I}^{\vec{l}}$  is the intensity image, and *G* is a convolutional Gaussian smoothing filter with a 3×3 kernel defined in [37] as:

$$[G]_{ij} = \frac{1}{2\pi\mu^2} e^{-\frac{i^2+j^2}{2\mu^2}}$$
(3)

where *i* and *j* are the distance from the origin in the horizontal and vertical axis, respectively, and  $\mu$  is the standard deviation of the distribution. Notice that since the corresponding dark background *D* and lightfield *L* images are not always available, we simply apply the 2D Gaussian smoothing filter *G* to the intensity images. The post-processed intensity  $\overline{I}^{1}$  image is therefore derived.

Afterwards, the normalized intensity image is binarized to yield a binary mask applied to the denoised lifetime image. Finally, a histogram-based contrast-enhancing algorithm from [37] is utilized to further improve the visual effect of the lifetime image, and the post-processed lifetime image is obtained.

In [7], we showed that combining both intensity and lifetime information together achieves better performance than using false-colour lifetime images alone for CNN-based cancer classification. In this study, we use intensity-weighted lifetime images as the output of the post-processing to be the input to the proposed model. With the evolution of the technologies, we observed that by feeding intensity-weighted lifetime images, the CNNs were able to obtain even better scores than the stacked images. The comparison of these two different formats on the classic CNNs can be found in "Appendix".

#### 3.3 Layer-level multi-scale architecture

As discussed in Sect. 2, multi-scale architectures can be implemented at layer and feature levels. The primary concern of the feature-level is that the correlations among

 Table 1 The effective number of images collected from 18 patients, including measured points on cancerous / non-cancerous tissue, and collected frames from each patient

Patient	Measured points		Frames	
	Cancer	Non-cancer	Cancer	Non-cancer
1	8	9	904	1712
2	10	14	880	1300
3	11	15	1080	1320
4	0	30	0	2400
5	4	8	792	1186
6	2	2	396	396
7	8	13	1196	1994
8	8	9	1196	1202
9	5	20	598	2790
10	9	8	1200	1200
11	13	2	1698	200
12	6	6	992	894
13	7	9	140	180
14	14	6	280	120
15	12	9	240	180
16	7	6	126	120
17	7	7	128	140
18	10	30	840	888
Total	141	203	12686	18222

the input features are partially ignored. Since both layerand feature-level multi-scale models can be easily and, in most cases, seamlessly integrated into the networks with residual or similar blocks, we apply the replacement at layer-level so that the correlations among them are retained. To benefit from the advantages of layer-level multi-scale architecture and hierarchical aggregation, we propose a new layer-level multi-scale convolution architecture, called *ResNetZ*.

**Fig. 3 a** original residual block in ResNet, **b** proposed ResNetZ module, where *A* is an aggregation operator, and **c** Res2Net module as a featurelevel multi-scale example. Both ResNetZ and Res2Net blocks contain a shortcut connection (the leftmost blue dash line)



ResNetZ (Fig. 3b) and Res2Net (Fig. 3c) are visually similar since our ResNetZ is inspired by the Res2Net hierarchical aggregation. However, they are conceptually distinct: one major difference is that our ResNetZ performs multi-scale feature extraction on the input features as a whole to retain the correlations among the input features (Fig. 3b), whereas Res2Net splits input features into groups and performs multi-scale feature extraction per grouped features (Fig. 3c).

#### 3.3.1 Block-wise shortcut connections

Comparing [11] and [32], an apparent difference, besides the utilization of dilated convolutions, is a shortcut connection used in [11]. Unlike the identity mapping in ResNet, which is used for a better flow of information through residual blocks, the shortcut connections in Res2Net (Fig. 3c) are located within the computational block. The advantage of this is that it helps the information and gradient flow. However, it also introduces extra complexity. For ResNetZ, the increased complexity is due to more feature maps being concatenated as input to the output  $1 \times 1$  convolution. Moreover, Res2Net also needs more feature maps extracted from the input  $1 \times 1$  convolution because of the splitting performed before the hierarchical aggregation.

#### 3.3.2 Hierarchical aggregation

Another configurable hyperparameter is the aggregation of the global and local features before the  $3\times3$  convolution. Both [11] and [32] employed the ResNet-like *Addition* operation, which is able to spatially integrate features without sacrificing complexity. In DenseNet, a major difference from ResNet is that it replaces the *Addition* by *Concatenation*, which increases variation in the input of successive layers and improves efficiency. While *Addition* requires input features that have identical dimensions, *Concatenation* is flexible in dimensional terms. The main disadvantage of *Concatenation*, however, is the extra parameters introduced. Inspired by DenseNet, *Concatenation* can also be used as a viable alternative to *Addition*. As a result, there are potentially four different configurations by combining the shortcut connection with the aggregation.

#### 3.3.3 ResNetZ block definition

Let  $x_g$  denote the global features from the first  $1 \times 1$  convolution as the input of the ResNetZ block,  $y_i$  denote the output features extracted from the branch  $3 \times 3$  convolution for  $i \in \{1, 2, ..., n\}$ , A be the aggregation, and  $\Gamma$  be a composite operation consisting of a  $3 \times 3$  convolution, batch normalization [23], and a rectified linear unit [38]. Let  $y_g$  be the output of the ResNetZ block. Accordingly,  $y_i$  and  $y_g$  without the shortcut are governed by:

$$y_i = \begin{cases} \boldsymbol{\Gamma}(x_g) & i = 1\\ \boldsymbol{\Gamma}(\boldsymbol{A}(x_g, y_{i-1})) & 1 < i \leq n \end{cases}$$
(4)

$$y_g = C([y_1, y_2, ..., y_n])$$
 (5)

where *C* is a concatenation operator. Let  $y'_i$  and  $y'_g$  be the model with the shortcut, which can be defined as:

$$y'_{i} = \begin{cases} x_{g} & i = 1\\ \Gamma(x_{g}) & i = 2\\ \Gamma(A(x_{g}, y'_{i-1})) & 2 < i \le n \end{cases}$$
(6)

$$y'_{g} = \boldsymbol{C}([y'_{1}, y'_{2}, ..., y'_{n}])$$
(7)

Since Res2Net splits the features  $x_g$  (Fig. 3c), the shortcut only passes partial information to the output. In contrast, with the shortcut, the ResNetZ block is able to pass the whole set of features  $x_g$  to the output and, thus, enhances the information flowing in forward and backward propagation within the block. Aggregation A can be further split into  $A^a$  and  $A^c$  for Addition and Concatenation operators, respectively. Since Addition is pixel-wise,  $A^{a}$  implicitly conveys local  $(y_i \text{ or } y'_i)$  information to the subsequent branches, whereas with  $A^c$ , local  $(y_i \text{ or } y'_i)$  information is explicitly carried to the remaining branches. In addition, by concatenating  $x_g$  and  $y_{i-1}/y'_{i-1}$  from different receptive fields, more information is expected to be integrated and retrieved as the output of the ResNetZ block. In consequence, the sequentially integrated output  $y_g$  and  $y'_g$  contains features rich in spatial and contextual information.

#### 3.3.4 ResNetZ complexity

When the number of parallel  $3 \times 3$  convolutions is fixed, given the same backbone network, the aggregation style (A symbol in Fig. 3b) and the shortcut connection (blue dash

line in Fig. 3b) will also affect the complexity and performance of the model. The shortcut connection introduces more features to be concatenated as the output of the ResNetZ block. When paralleling several  $3 \times 3$  convolutions with aggregation, the receptive field of each branch will also increase exponentially, due to the features from the previous branch. In addition, compared with Addition, *Concatenation* doubles the features to be fed into the  $3 \times 3$ convolution. Taking into account the shortcut and aggregation styles, when the width and the scale are small, the complexity of the model is  $C_{\rm add} < C_{\rm concat} <$  $C_{\text{add+shortcut}} < C_{\text{concat+shortcut}}$ . When they become larger, the complexity of the model changes to  $C_{add} < C_{add+shortcut} <$  $C_{\text{concat}} < C_{\text{concat+shortcut}}$ .

#### 3.4 Implementation details

All models were implemented in PyTorch.<sup>1</sup> For the existing CNNs, we used their official implementation included in PyTorch or published by their authors. For ResNetZ, we integrated the layer-level multi-scale architecture into the PyTorch-implemented ResNet. As per Fig. 3b, we substituted the original residual block (Fig. 3a) with the ResNetZ model (Fig. 3b). Unlike ResNeXt and Res2Net, which keep the width of the ResNet backbone, we used a narrower version of ResNet, so the width of the input  $1 \times 1$  convolution is retained as for the branch  $3 \times 3$  convolution, thereby reducing the overall complexity of our ResNetZ model.

To ensure a fair comparison, we adapted the authors' official implementation of Res2Net to a similar version of our proposal, adjusting the width of the backbone ResNet, so that the scale and width of ResNetZ and Res2Net are equal. In addition, ResNetZ followed the same configuration of Res2Net by using *Addition* as the aggregation with a shortcut connection.

All models were examined on 61, 816 FLIM images from 18 patients, as described in Sect. 3.1. Images from 17 patients (patients 1 - 17 in Table 1) were used as the training set, where 10% training images were split as the validation set. The images from the remaining patient (patient 18) served as the independent testing set, which contains 840 cancerous images and 888 normal images. For all evaluated CNNs, we applied a stochastic gradient descent for optimization with momentum 0.9. The learning rate was initially set to 0.1, and divided by 10 at epochs 50, 100, 150, and 175 for in total 200 epochs, with batch size 128. We used binary cross-entropy as the loss function. Weights were initialized using He's method [39]. In addition, we also employed weight decay 1e - 4. For data augmentation, we utilized a simple strategy of vertical and

<sup>&</sup>lt;sup>1</sup> http://pytorch.org.

horizontal flipping, as well as random crop with zero-value padding of 16 pixels. All training and testing were performed using NVidia V100 GPU provided by JADE.<sup>2</sup>

## **4** Results

In order to fully quantify the performance of the proposed layer-level multi-scale architecture, we first evaluate the influence of the aggregation style and the shortcut connections. We then fix these two parameters and evaluate the impact of integrating the model in different ResNets. Finally, we compare our layer-level multi-scale model with a feature-level model (Res2Net) to understand how levels impact the results.

#### 4.1 ResNetZ configurations

According to Fig. 3, multi-scale architectures potentially have different configurations. We tested the proposed model on ResNet50 as the backbone with width 8, in parallel with scales 2, 4, and 6. Following the naming convention in [11], we also use *S* for scale and *W* for width. For ResNetZ50-W8-S2, the complexity is consistent with the former situation, whereas the other two (ResNetZ50-W8-S4 and ResNetZ50-W8-S6) are with the latter. To simplify the presentation, we append *A*, *AS*, *C*, and *CS* to the model names to represent the model with *Addition*, *Addition with Shortcut*, respectively.

The scores, depicted in Fig. 4, are grouped by accuracy, precision, recall, AUC, and Kappa. All variations except recall show a very similar tendency, with different configurations and scales. For ResNetZ50-W8-S2, the shortcut is very helpful for Addition aggregation, with 4% gain, but has little effect on Concatenation. The Concatenation achieves higher scores than Addition, except on recall, regardless of having the shortcut or not. When the scale increases from 2 to 4, i.e. ResNetZ50-W8-S4, the shortcut still leads to performance improvement, especially for Addition. In contrast, Concatenation is not always better than Addition. In ResNetZ50-W8-S4-AS and ResNetZ50-W8-S4-C, the Addition with shortcut achieves very similar results than Concatenation with or without shortcut on accuracy, AUC, and Kappa. Further increasing scale to 6, the shortcut still improves the performance of Addition, but considerably deteriorates on Concatenation, which is even lower than Addition alone. ResNetZ50-W8-S6-AS and ResNetZ50-W8-S6-C produce comparable results on accuracy, AUC, and Kappa, where the discrepancy is less than 1%.

In general, the shortcut connection always introduces performance gain in terms of accuracy, precision, AUC, and Kappa. The gain is usually more remarkable on Addition than on Concatenation, mainly when the scale is small. An exception occurs when the model is relatively complex since an extra shortcut connection does not improve the performance for Concatenation. That is, ResNetZ50-W8-S6-CS is inferior to ResNetZ50-W8-S6-C for all metrics except recall. For the aggregation styles, Concatenation is usually superior to Addition, with or without the shortcut, except for ResNetZ50-W8-S6 with the shortcut. This is not unexpected since the features introduced by Concatenation are twice than in Addition for the convolution branch. In summary, Concatenation without shortcut is overall superior almost for all the metrics than the other three configurations with three different scales.

#### 4.2 Overall performance

Based on the results of Sect. 4.1, we evaluate the performance of the proposed model with *Concatenation* as the aggregation without the shortcut connection. We first evaluate six state-of-the-art CNNs, namely ResNet50, DenseNet121, Inception, Xception, SENet, and Res2Net. It is worth noting that the classification of FLIM images may not benefit from very complex CNNs as prior experience. Therefore, we appraise three shallow ResNet variations, including ResNet38 and ResNet50, with two different widths. Further, we use these three variations as the backbone networks and integrate the ResNetZ block into the backbone, but with a smaller width. The results are listed in Table 2 and ResNetZ ROC curves, along with two backbone ResNet in Fig. 5.

Amongst the contemporary CNNs, Res2Net50 achieves the overall best scores. Meanwhile, DenseNet121 is inferior to ResNet and its variations. Its performance may improve with deeper configurations. In the backbone ResNet, with depth growing from 38 to 50 and width from 32 to 64, the outcomes increase consistently, except for ResNet50-w64. Considering these are relatively shallow networks and the relatively small FLIM dataset, these outcomes are anticipated. However, note that as ResNet50w32 yields better scores than ResNet50-w64, other stateof-the-art CNNs with less width may perform better.

In general, ResNetZ surpasses the backbone ResNet with three different depths, but with significantly fewer parameters since ResNetZ employs parallel 3 X 3 convolutions with narrow width. Specifically, ResNetZ38-W16-S2 achieves the overall best scores, but with fewer than 3.5M parameters. For ResNetZ38, all ResNetZ models are superior to ResNet38 but require considerably fewer parameters. A further deepening of ResNet to 50 layers,

<sup>&</sup>lt;sup>2</sup> http://www.jade.ac.uk.



Fig. 4 Impact of different ResNetZ configurations on the results

Table 2Performancecomparison of the backboneResNet, ResNetZ, and threevariations of ResNet family

	Acc (%)	Prec (%)	Recall (%)	AUC (%)	Kappa (%)	Params (10 <sup>6</sup> )
ResNet50	83.56	89.43	77.14	83.75	67.23	23.51
DenseNet121	82.99	89.08	76.24	83.18	66.09	6.95
Inception	82.45	88.36	74.24	85.03	64.39	6.95
Xception	81.08	78.19	87.61	80.89	61.99	20.81
SENet50	83.96	90.59	81.08	84.06	68.26	26.0
Res2Net50	84.61	88.68	80.29	84.73	69.27	23.65
ResNet38-w32	84.20	89.47	78.49	84.36	68.49	7.4
ResNet38-w64	85.36	85.16	86.60	85.32	70.68	16.8
† ResNetZ38-W12-S2	86.28	88.34	84.46	86.34	72.58	2.3
† ResNetZ38-W16-S2	88.14	92.42	83.78	88.26	76.32	3.5
† ResNetZ38-W12-S4	86.46	88.93	84.12	86.53	72.94	3.3
† ResNetZ38-W16-S4	87.44	89.52	85.59	87.50	74.90	5.1
ResNet50-w32	86.98	90.38	83.56	87.08	73.99	8.3
ResNet50-w64	83.56	89.43	77.14	83.75	67.23	23.5
† ResNetZ50-W12-S2	82.81	87.45	77.70	82.96	65.71	2.9
† ResNetZ50-W16-S2	85.76	87.24	84.68	85.79	71.53	4.5
† ResNetZ50-W12-S4	86.05	91.63	80.18	86.22	72.18	8.5
† ResNetZ50-W16-S4	83.28	86.39	80.07	83.37	66.59	12.6

The best scores of each group are in bold, and the overall best scores are highlighted in italic. Our proposed architecture is indicated by a <sup>†</sup>

ResNet50-W32 yields the best scores in accuracy, AUC, and Kappa. However, ResNetZ50-W12-S4 also produces very comparable results.

Considering Fig. 4 and Sect. 4.1, with a relatively simple ResNet, the performance of ResNetZ improves by increasing depth, width, and scale. Due to the parallel  $3 \times 3$  convolutions, which concatenate more features, the model

produces better scores than the backbone with considerably fewer parameters. Although there are exceptions in ResNet50, the decline is understandable since the model already achieved the best outcomes with ResNet38 and, hence, the scores of more complex ResNetZ variations may also drop.





#### 4.3 Multi-scale at layer and feature levels

ResNetZ extracts and fuses features at layer-level, whereas Res2Net performs similar operations at feature-level. We conduct further experiments to compare both models. To make the comparison fair, we follow the architecture of Res2Net, i.e. using addition for the aggregation with the shortcut, and applying the same width and scale to both models. The results are shown in Fig. 6. In general, ResNetZ yields promising scores, slightly better than Res2Net but with significantly fewer parameters. As shown in Fig. 3, given the same width and scale, Res2Net requires a much wider 1 X 1 convolution as input to maintain the width of the branch, in contrast to ResNetZ. When the width and scale are relatively large, the difference in complexity becomes significant. Additionally, given the same configurations and backbone networks, the highest score achieved by Res2Net is larger than for ResNetZ. Regarding accuracy, both architectures improve almost consistently when growing the number of parameters, followed by a decline after reaching the peak (Fig. 6, plot 1). For four of the seven variations, ResNetZ produces higher scores than Res2Net and the highest, 88.83% compared to 88.19%. The same tendency is found on AUC and Kappa. For precision (Fig. 6, plot 2), Res2Net is marginally better than ResNetZ, but their best scores are very close, 90.83% and 91% for Res2Net and ResNetZ, respectively. As for recall (Fig. 6, plot 3), both models obtain comparable results. In this case, all scores of our model are over 80%, whereas Res2Net produces both the best 87.05% and the worst 77.25% scores.



Fig. 6 ResNetZ vs Res2Net over accuracy (first plot), precision (second plot), recall (third plot), AUC (fourth plot), and Kappa (fifth plot)

18891



Fig. 7 Visualization of class activation map (CAM) using Grad-CAM on normal and cancerous images with ResNet, Res2Net, and ResNetZ

# 5 Discussion

#### 5.1 FLIM images

Unlike other biomedical images, FLIM provides an extra dimension and introduces several visual recognition challenges. With the capability of user-specified configurations, the custom fibre-based FLIM system is able to deliver multi-dimensional images rich in spatial and spectral information. Although the dataset is relatively small in terms of the number of patients, it will gradually increase over time as the sample collection is still ongoing. Since the FLIM system was designed for in-vivo in-situ diagnostics with endoscopically delivered fibres, reliable classification is, therefore, of great clinical importance in realtime human lung cancer diagnostic pathways.

To attract more engineers, researchers, and enthusiasts to overcome the challenges and together push forward this particular area, the FLIM dataset is available on https:// github.com/qiangwang57/flim\_cancer\_ml.

#### 5.2 Model configurations

Multi-scale strategies are flexible in configurations. An identity shortcut has proved to be helpful for information and gradient flowing, achieving better outcomes [8, 9, 11]. In this study, when the model is simple in scale, the performance gain by the shortcut is significant, especially for Addition aggregation. However, for Concatenation, the shortcut improvement is not significant and, in an extreme case, harms the performance. Consequently, the shortcut should be used with Addition. As far as the aggregation is concerned, Concatenation is overall better than Addition in most cases, which is expected since it doubles the input features. As a result, Concatenation can be used, in general, if the priority is model performance. When complexity is a major concern, particularly when ResNetZ has relatively more branch convolutions, Addition with shortcut can substitute Concatenation for comparable performance with relatively less complexity.

#### 5.3 Layer-level multi-scale architecture

A remarkable advantage of ResNetZ is the complexity compared with the backbone ResNet. The primary reason is the parallel  $3 \times 3$  convolutions, along with the aggregation. This enables the fusion and extraction of features at different scales, which are further concatenated as the output. A direct consequence is that each  $3 \times 3$  convolution is much narrower than the original ResNet. With concatenation as the aggregation, our ResNetZ has similarities with DenseNet. Within the block, every branch convolution is supervised directly by the input, enforcing the convolution to learn different features, except that the supervision is performed at block scope. Although the results in Sect. 4 show that the performance gain is not always consistent with the number of parameters, we believe this is not because of the model itself but due to the relatively small patient diversity in the dataset.

Figure 7 illustrates the CAM areas generated by ResNet, Res2Net, and ResNetZ on images of normal and cancer tissues, based on Grad-CAM [40]. For normal and cancerous images, the CAM areas produced by ResNet generally cover almost entire image, but with partial concentration on a particular portion. Res2Net tends to have larger focusing areas than ResNet, with more concentration on the part with moderate brightness. The CAM results by ResNetZ are further enhanced, comparing with Res2Net. In particular, ResNetZ tends to have multiple concentrating areas in the images with mild brightness. This indicates the superiority of ResNetZ over ResNet and Res2Net.

Both layer- and feature-level multi-scale architectures perform comparably on the FLIM dataset given the same width, depth, scale, and configuration. Since feature-level multi-scale needs to split the input features, unlike layerlevel, it contains more parameters due to wider input features. The difference in the number of parameters becomes significant when the backbone network is complex. Overall, they are comparable in terms of metrics.

# 6 Conclusion

This paper formally introduced a unique FLIM image dataset, described the best practices to improve raw input images, and proposed a novel multi-scale CNN, called ResNetZ, to further improve lung cancer classification. Through 61, 816 FLIM images on 18 pairs of normal/cancerous tissue collected from 18 patients, we show the superiority of the proposed method over the backbone ResNet with significantly fewer parameters. In particular, ResNetZ38-W16-S2 presented the overall best performance but with only 3.5 M parameters. We also compared our layer-level multi-scale model with feature-level one (Res2Net) to demonstrate the advantages and disadvantages of the ResNetZ model. Given the same model configurations, ResNetZ is superior to Res2Net in most cases. It is notable that with the same configurations, Res2Net is up to 20% more complex than ResNetZ. Since the FLIM system is designed for online *in-vivo in-situ* imaging, fewer parameters are more convenient for real-time clinical diagnostics provided by clinicians at bedside. As ResNetZ is designed to be independent of the backbone network, it could be easily migrated to other networks with similar convolutional blocks, such as segmentation-oriented networks. Future research will be conducted on the migration of our approach to other backbone networks and different research scenarios.

# Appendix: Image formats input to CNNs

Integrating lifetime with intensity information (third image in Fig. 8) has proved to be more effective for CNN-based lung cancer classification using FLIM images [7]. To further improve the performance, we evaluated intensityweighted lifetime images (fourth image in Fig. 8), adapted from [3]. To make the comparison unbiased, we strictly follow the methodology shown in [7]. The results are listed in Table 3. Generally, the outcomes using the weighted



Fig. 8 Three-channel stacked image (third) and intensity-weighted lifetime image (fourth) by intensity (first) and lifetime (second) image

Table 3 Channel-stacked images versus intensity-weighted images

	Accuracy	Precision	Recall	AUC					
CNNs on intensity+lifetime 3-channel input [7]									
ResNet50	0.816	0.804	0.943	0.767					
ResNeXt50	0.839	0.839	0.902	0.824					
DenseNet121	0.865	0.88	0.895	0.858					
Inception	0.852	0.876	0.874	0.846					
Xception	0.817	0.895	0.785	0.825					
CNNs on intensity-weighted lifetime									
ResNet50	0.854	0.876	0.88	0.848					
ResNeXt50	0.859	0.908	0.851	0.861					
DenseNet121	0.837	0.888	0.832	0.838					
Inception	0.79	0.839	0.803	0.787					
Xception	0.879	0.961	0.832	0.891					

lifetime images are better than those using the stacked images, except for recall. Particularly, ResNet50 and ResNeXt50 yield significantly better scores on the intensity-weighted lifetime images than on the stacked images. As a result, the combination of intensity and lifetime by weighting is chosen for this and for future studies.

Acknowledgements We thank Dr Catharine Ann Dhaliwal for annotating the histological images. This project made use of time on Tier 2 HPC facility JADE, funded by EPSRC (EP/P020275/1).

**Funding** This work is supported by the Engineering and Physical Sciences Research Council (EPSRC, United Kingdom) Interdisciplinary Research Collaboration (Grant Number EP/K03197X/1 and EP/R005257/1). Dr A. Akram is supported by a Cancer Research UK Clinician Scientist Fellowship (A24867).

#### Declarations

**Conflict of interest** The authors declare that they have no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

#### References

- Suhling K, Hirvonen LM, Levitt JA, Chung PH, Tregidgo C, Le Marois A, Rusakov DA, Zheng K, Ameer-Beg S, Poland S, Coelho S, Henderson R, Krstajic N (2015) Fluorescence lifetime imaging: Basic concepts and some recent developments. Med Photonics 27:3–40. https://doi.org/10.1016/j.medpho.2014.12. 001
- Jo JA, Cheng S, Cuenca-Martinez R, Duran-Sierra E, Malik B, Ahmed B, Maitland K, Cheng Y-SL, Wright J, Reese T (2018) Ous fluorescence lifetime imaging (FLIM) endoscopy for early detection of oral cancer and dysplasia. In: 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC), pp 3009–3012. https://doi.org/10.1109/ EMBC.2018.8513027
- McGinty J, Galletly NP, Dunsby C, Munro I, Elson DS, Requejo-Isidro J, Cohen P, Ahmad R, Forsyth A, Thillainayagam AV et al (2010) Wide-field fluorescence lifetime imaging of cancer. Biomed Opt Express 1(2):627–640
- Cheng S, Cuenca RM, Liu B, Malik BH, Jabbour JM, Maitland KC, Wright J, Cheng Y-SL, Jo JA (2014) Handheld multispectral fluorescence lifetime imaging system for in vivo applications. Biomed Opt Express 5(3):921–931
- Xing F, Xie Y, Su H, Liu F, Yang L (2018) Deep learning in microscopy image analysis: a survey. IEEE Trans Neural Netw Learn Sys 29(10):4550–4568. https://doi.org/10.1109/TNNLS. 2017.2766168
- Chen B, Lu Y, Pan W, Xiong J, Yang Z, Yan W, Liu L, Qu J (2019) Support vector machine classification of nonmelanoma skin lesions based on fluorescence lifetime imaging microscopy. Anal Chem 91(20):10640–10647. https://doi.org/10.1021/acs. analchem.9b01866
- Wang Q, Hopgood JR, Finlayson N, Williams GO, Fernandes S, Williams E, Akram A, Dhaliwal K, Vallejo M (2020) Deep learning in ex-vivo lung cancer discrimination using fluorescence lifetime endomicroscopic images. In: 2020 42nd annual international conference of the IEEE engineering in medicine & biology society (EMBC), pp 1891–1894. IEEE
- He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp 770–778. https://doi.org/10.1109/CVPR. 2016.90
- Huang G, Liu Z, v. d. Maaten L, Weinberger KQ (2017)Densely connected convolutional networks. In: IEEE conference on computer vision and pattern recognition, pp 2261–2269. https:// doi.org/10.1109/CVPR.2017.243
- Huang G, Sun Y, Liu Z, Sedra D, Weinberger KQ (2016) Deep networks with stochastic depth. In: European conference on computer vision, pp 646–661. Springer
- Gao S, Cheng M, Zhao K, Zhang X, Yang M, Torr PHS (2019) Res2net: a new multi-scale backbone architecture. IEEE transactions on pattern analysis and machine intelligence, 1. https:// doi.org/10.1109/TPAMI.2019.2938758
- Erdogan AT, Walker R, Finlayson N, Krstajić N, Williams G, Girkin J, Henderson R (2019) A CMOS SPAD line sensor with per-pixel histogramming TDC for time-resolved multispectral imaging. IEEE J Solid-State Circuits 54(6):1705–1719
- Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Meas 20(1):37–46
- Gu J, Fu CY, Ng BK, Gulam Razul S, Lim SK (2014) Quantitative diagnosis of cervical neoplasia using fluorescence lifetime imaging on haematoxylin and eosin stained tissue sections. J Biophotonics 7(7):483–491
- Cuenca R, Cheng S, Malik BH, Maitland KC, Ahmed B, Cheng Y-SL, Wright JM, Rees T, Jo JA (2018) Learning methods for

fluorescence lifetime imaging (FLIM) based automated detection of early stage oral cancer and dysplasia (conference presentation). In: Optical imaging, therapeutics, and advanced technology in head and neck surgery and otolaryngology 2018, vol 10469, p 104690. International Society for Optics and Photonics

- Marsden M, Weyers BW, Bec J, Sun T, Gandour-Edwards RF, Birkeland AC, Abouyared M, Bewley AF, Farwell DG, Marcu L (2020) Intraoperative margin assessment in oral and oropharyngeal cancer using label-free fluorescence lifetime imaging and machine learning. Trans Biomed Eng. https://doi.org/10.1109/ TBME.2020.3010480
- Wang Q, Vallejo M, Hopgood J (2020) Fluorescence lifetime endomicroscopic image-based ex-vivo human lung cancer differentiation using machine learning. TechRxiv Preprint. https:// doi.org/10.36227/techrxiv.11535708.v1
- Wang Q, Hopgood JR, Vallejo M (2021) Fluorescence lifetime imaging endomicroscopy based ex-vivo lung cancer prediction using multi-scale concatenated-dilation convolutional neural networks. In: Medical imaging 2021: computer-aided diagnosis, vol 11597, p 115972. International Society for Optics and Photonics
- Wang Q, Hopgood JR, Vallejo M (2021) Multi-scale aggregateddilation network for ex-vivo lung cancer detection with fluorescence lifetime imaging endomicroscopy. In: 2021 43rd annual international conference of the IEEE engineering in medicine & biology society (EMBC), pp 2918–2922. IEEE
- Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJNL, Išgum I (2016) Automatic segmentation of mr brain images with a convolutional neural network. IEEE Trans Med Imaging 35(5):1252–1261. https://doi.org/10.1109/TMI.2016. 2548501
- Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: IEEE conference on computer vision and pattern recognition, pp 2818–2826
- Setio AAA, Ciompi F, Litjens G, Gerke P, Jacobs C, van Riel SJ, Wille MMW, Naqibullah M, Sánchez CI, van Ginneken B (2016) Pulmonary nodule detection in ct images: false positive reduction using multi-view convolutional networks. IEEE Trans Med Imaging 35(5):1160–1169. https://doi.org/10.1109/TMI.2016. 2536809
- Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: 32nd international conference on machine learning, pp 448–456
- Mou L, Chen L, Cheng J, Gu Z, Zhao Y, Liu J (2019) Dense dilated network with probability regularized walk for vessel detection. IEEE Transactions on medical imaging, 1. https://doi. org/10.1109/TMI.2019.2950051
- Chen L, Papandreou G, Kokkinos IKM et al (2018) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848. https://doi.org/10.1109/ TPAMI.2017.2699184
- Chen L, Bentley P, Mori K, Misawa K, Fujiwara M, Rueckert D (2018) Drinet for medical image segmentation. IEEE Trans Med Imag 37(11):2453–2462. https://doi.org/10.1109/TMI.2018. 2835303
- Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp 1097–1105
- Chollet F (2017) Xception: deep learning with depthwise separable convolutions. In: IEEE conference on computer vision and pattern recognition, pp 1251–1258
- 29. Tan M, Le QV (2019) MixConv: mixed depthwise convolutional kernels. In: 30th british machine vision conference

- 30. Zhang X, Zhou X, Lin M, Sun J (2018) Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: IEEE conference on computer vision and pattern recognition
- Zhang H, Wu C, Zhang Z, Zhu Y, Zhang Z, Lin H, Sun Y, He T, Mueller J, Manmatha R et al (2020) Resnest: split-attention networks. arXiv preprint arXiv:2004.08955
- 32. Liu M, Yin H (2019) Feature pyramid encoding network for realtime semantic segmentation. In: british machine vision conference
- 33. Williams GO, Williams E, Finlayson N, Erdogan AT, Wang Q, Fernandes S, Akram AR, Dhaliwal K, Henderson RK, Girkin JM, Bradley M (2021) Full spectrum fluorescence lifetime imaging with 0.5 nm spectral and 50 ps temporal resolution. Nat Commun 12(1):1–9
- Ballew RM, Demas J (1989) An error analysis of the rapid lifetime determination method for the evaluation of single exponential decays. Anal Chem 61(1):30–33
- Philip J, Carlsson K (2003) Theoretical investigation of the signal-to-noise ratio in fluorescence lifetime imaging. J Opt Soc Am 20(2):368–379

- Ford TN, Lim D, Mertz J (2012) Fast optically sectioned fluorescence HiLo endomicroscopy. J Biomed Opt 17(2):021105. https://doi.org/10.1117/1.jbo.17.2.021105
- 37. Sonka M, Hlavac V, Boyle R (2014) Image processing, analysis, and machine vision. Cengage Learning, Stamford
- Glorot X, Bordes A, Bengio Y (2011) Deep sparse rectifier neural networks. In: 14th international conference on artificial intelligence and statistics, vol 15, pp 315–323
- 39. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: international conference on computer vision, pp 1026–1034
- Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE international conference on computer vision, pp 618–626

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

# **Authors and Affiliations**

Qiang Wang<sup>1</sup> · James R. Hopgood<sup>2</sup> · Susan Fernandes<sup>1</sup> · Neil Finlayson<sup>2</sup> · Gareth O. S. Williams<sup>1</sup> · Ahsan R. Akram<sup>1</sup> · Kevin Dhaliwal<sup>1</sup> · Marta Vallejo<sup>3</sup>  $\bigcirc$ 

☑ Qiang Wang Q.Wang@ed.ac.uk

Marta Vallejo M.Vallejo@hw.ac.uk

> James R. Hopgood James.Hopgood@ed.ac.uk

Susan Fernandes Susan.Fernandes@ed.ac.uk

Neil Finlayson N.Finlayson@ed.ac.uk

Gareth O. S. Williams G.O.S.Williams@ed.ac.uk Ahsan R. Akram Ahsan.Akram@ed.ac.uk

Kevin Dhaliwal Kev.Dhaliwal@ed.ac.uk

- <sup>1</sup> Centre for Inflammation Research, University of Edinburgh, Edinburgh, United Kingdom
- <sup>2</sup> School of Engineering, University of Edinburgh, Edinburgh, United Kingdom
- <sup>3</sup> School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh, United Kingdom