Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 19 No. 10 (2023) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v19i10.39721

PAPER

Melanoma Classification via Hybrid Saliency and Conditional Random Field with Bottleneck to Optimize DeepLab

Vo Thi Hong Tuyet[1,2,3], Nguyen Thanh Binh[1](✉)

[1]Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and Information Technology (HUFLIT), Ho Chi Minh City, Vietnam

[2]Department of Information Systems, Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), Ho Chi Minh City, Vietnam

[3]Vietnam National University Ho Chi Minh City, Linh Trung Ward, Thu Duc City, Ho Chi Minh City, Vietnam

binh@huflit.edu.vn

ABSTRACT

Neural networks overcome drawbacks of vision tasks by becoming convolutional in a wide range of layers. The salient map is affected by multilevels of strong pixels (superpixels) in global images and that is dependent on the hard threshold for their dividing. Deep neural networks have been established for saliency prediction of segmentation because the feature extraction must be suited to the input data. The convolutional neural network (CNN) also endures conflict between spatial pattern and a likeness of salient objects. Semantic segmentation is one of the approaches to continue classification based on these features. Therefore, upgrading the extraction process can be of use in saliency. In this work, we optimize DeepLab based on an atrous convolutional and a conditional random field (CRF) with a bottleneck in the semantic segmentation method, which serves for classification. The backbone of deep feature extraction is atrous convolution and the bottleneck based on CRF for hybrid saliency in the encoder-decoder system. The classification results are compared with some approaches for saliency prediction of recent deeper methods in an ISIC 2017 dataset. The results give better values not only for saliency prediction for segmentation but also for training and testing for classification.

KEYWORDS

saliency, conditional random field, bottleneck, atrous convolutional, Deeplab, segmentation, classification

1 INTRODUCTION

The structure of layers in a neural network depends on a wide range of parameters. The input of any network depends on not only the size but also the number of kernels. The intensity of the relationship between objects is discerned by color or position in the global image. From these differences, the boundary of each region can be detected exactly. In state-of-the-art methods, classification based on

Tuyet, V.T.H., Binh, N.T. (2023). Melanoma Classification via Hybrid Saliency and Conditional Random Field with Bottleneck to Optimize DeepLab. *International Journal of Online and Biomedical Engineering (iJOE)*, 19(10), pp. 140–155. https://doi.org/10.3991/ijoe.v19i10.39721

Article submitted 2023-03-20. Resubmitted 2023-05-19. Final acceptance 2023-05-19. Final version published as submitted by the authors.

shape is widely applied. Neural networks become one of the best choices for feature extraction because of their advantages. The sensitivity of the output from network models must fully adapt with the prediction. However, strike and the number of features can create many difficulties. This happens when the input data has many features or noise and the contribution levels are different. This limitation the amount of useful information that is necessary for learning. Therefore, being able to detect the boundary of an object can reduce the unimportant areas.

The usefulness of segmentation had been selected for the overcoming ways of feature extraction. The shape of objects in images was solved by automatic laser welding and Sobel-contour [1]. Fausto [2] extracted an object's characterization by V-Net. A convolutional neural network was applied for bio-medical image segmentation in 2018 [3]. The number of models and layers for applications was a topic of interest for Holger and his research team in [4]. These approaches focused on the number of levels for the vital areas.

The problem of multiple levels for object prediction can be surmounted by use of a salient map. The characteristics of objects in a global image include color, the number of objects, context awareness, etc. These distinctions are the vital features for classification algorithms. If any method depends on only the learning model, the results will be limited. From the superpixels concept, saliency prediction applies to the detection of edges and contours. Background in each image also presents problems for distinguishing adjacent regions. Therefore, the neural network and deep learning are popular. This includes shallow and deep convolutional networks [5], level labels [6], and convolutional neural networks [7]. Exploring combined approaches with local parameters [8–14, 20, 24], Lai [15] proposed a saliency-structured learning model with dynamic visual attention. In this, the calculated blocks and fully connected network are built for any structure to find the boundary. The use of a filter matrix as a condition for sliding windows is the idea of [16]. However, the output size with the saliency prediction is hard because of the downsizing of neural layers. For this, the encoder and decoder network is a better construct [17, 18, 22, 23]. Goyzueta [19] combined U-Net, ResU-Net, and DeepLab architectures for image segmentation based on the same structure. Diverse feature extraction using dynamic effects was improved with DeepLab [21]. Feature extraction based on skin lesions has been developed by Li [25], Iqbal [26], Mahbod [28], and Al-Masni [29].

The above solutions optimized the objective functions. Their predictions considered the weights of all network layers. However, the model weights and backbone networks are the key to semantic segmentation. In recent years, convolutional neural networks (CNNs) have been a popular approach for classification and segmentation based on morphology, semantics, and context [30, 31]. In [32], a novel framework for an automatic, real-time, content-based image-retrieval approach was proposed by Malek. The processing includes query features from feature extraction based on vector computing. The manipulation used DeepLab to remove the last fully connected layer to achieve end-to-end output of the last two pooling layers, with the backbone being the VGG16 network. Then, atrous spatial pyramid pooling (ASPP) was used for improving the structure for classification. This reduced the number of parameters and added images upgraded by ResNet-101 from the previous version of DeepLab. A deeper network could replace all convolution and pooling layers, but improvement can start with encoder and decoder stages. This approach should continue to develop in the future.

The proposed method develops DeepLab by atrous convolutional and CRF model for an encoder-decoder system to detect salient maps. Our method includes four periods: a Gaussian filter for smoothing, the downsampling with atrous convolution in a CRF model, bottleneck elimination in the backbone encoder-decoder system with end-to-end dual saliency prediction, and neural layer for classification.

The organization of this paper is as follows: Section 2 introduces the basics of saliency-structured deep-learning methods; section 3 presents the proposed method for saliency prediction clearly; the experiment and results are shown in Section 4; and Section 5 is our conclusions.

2 A SALIENCY-STRUCTURED DEEP-LEARNING METHOD FOR PREDICTION

Almost all saliency-structured deep-learning models depend on the architecture of neural networks. The modular anguylar changes (MAC) block, usually used in prediction, includes a large number of block sizes and layers [28]. This structure is summarized in Figure 1. From the image array of the input stage (360 × 360 spatial resolution), the superpixel-level hand-crafted features focus on the focal stacks from light fields. The feature maps were selected by MAC block 9 × 9. A deep convolutional network based on angular kernels on the upsampled image was applied similarly and continuously in each part (540 × 375 × 64 feature map and variants with kernel size k × k × C and stride S). In the MAC stage, 5 blocks from block 1 to block 5 have two layers—convolutional and max pooling—with the size decreased by half in successive blocks. The atrous spatial pyramid pooling (ASPP) encompassed four sections. The parallel convolutional layers give the fusion for bilinear interpolation.

DeepLab is the concept of deep neural networks (DNNs). That is a method which computes the rank and number of neural layers in networks. DeepLab version 1.0 uses a deep convolutional neural network (DCNN) in each image to calculate a visual score map. The score map is the input of bi-linear interpolation. This system can be adjusted by fully connected CRF from the strong regions of segmentation. The disadvantage of this version is the single neural layer in the network. The parameters are limited and may lack sufficient information. Jiang [15] proposed the second version of DeepLab to overcome this problem by highlighting convolution with upsampled filters in video segmentation. This is the dilated convolution (ConvNet improving) in the deep convolutional neural network by an atrous algorithm. The upgraded structure of prediction in DeepLab version 2.0 is:

- DCNN is calculated by five neural networks from the input images and one of them is the atrous convolution.
- The combination between an object-to-motion convolutional network (OM-CNN) and saliency-structured convolutional long short-term memory (SS-ConvLSTM).
- Upgrading the main structures of objectnes subnets. The aim of this period is the change of size/strike/kernel/ordinal of layers in networks.

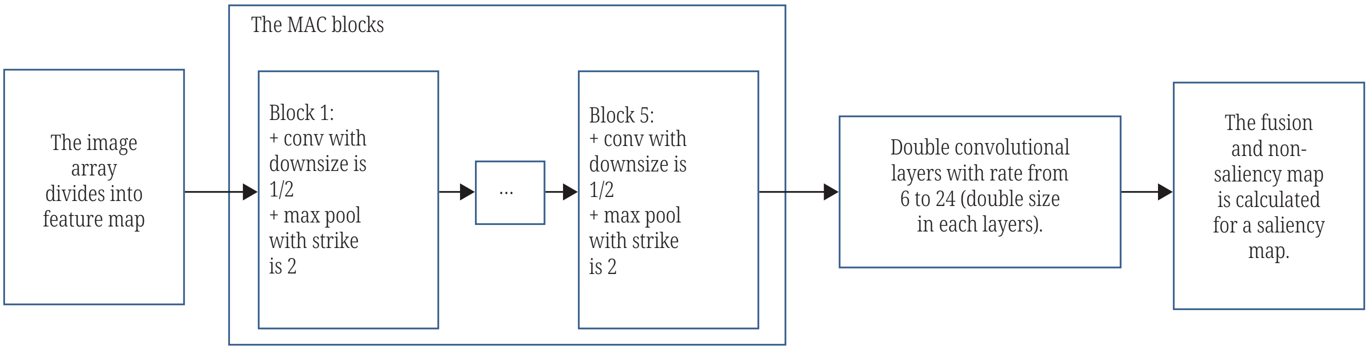

The MAC blocks
The image array divides into feature map
Block 1: + conv with downsize is 1/2 + max pool with strike is 2
...
Block 5: + conv with downsize is 1/2 + max pool with strike is 2
Double convolutional layers with rate from 6 to 24 (double size in each layers).
The fusion and non-saliency map is calculated for a saliency map.

Fig. 1. Architecture of MAC for conversion from a micro-lens image array into a saliency map [14]

The encoder-decoder structure was proposed in [17, 18] for saliency prediction. These systems included deep convolutional neural networks and an encoder-decoder architecture in its backbone. The different parallel mappings in [18] are the full scales and apply local and global contextual features for the prediction period. From the above mentioned, any deep-learning model for saliency-prediction always suffers from limitations of the network parameters and the superpixel level. These results were limited by the ordinal parameters and the reversion of fully connected at each interpolation level.

3 THE PROPOSED METHOD FOR MELANOMA CLASSIFICATION VIA HYBRID SALIENCY

This section presented the method of saliency prediction for classification. The atrous convolution of DeepLab version 2 was applied in 6 blocks of downsampling of the neural network with the same kernel. The results of these blocks provided the features for segmentation. Synthesis of saliency was not done because of the limitation of the segmentation results. Instead, saliency prediction was based on the binary map of each object and smoothing by filter. This is independent of the initial values for the level dividing. In this paper, the saliency method applied morphology for the preprocessing period by Gaussian filter and improved the feature extraction of DeepLab based on the atrous convolution and bottleneck in CRF model (Figure 2).

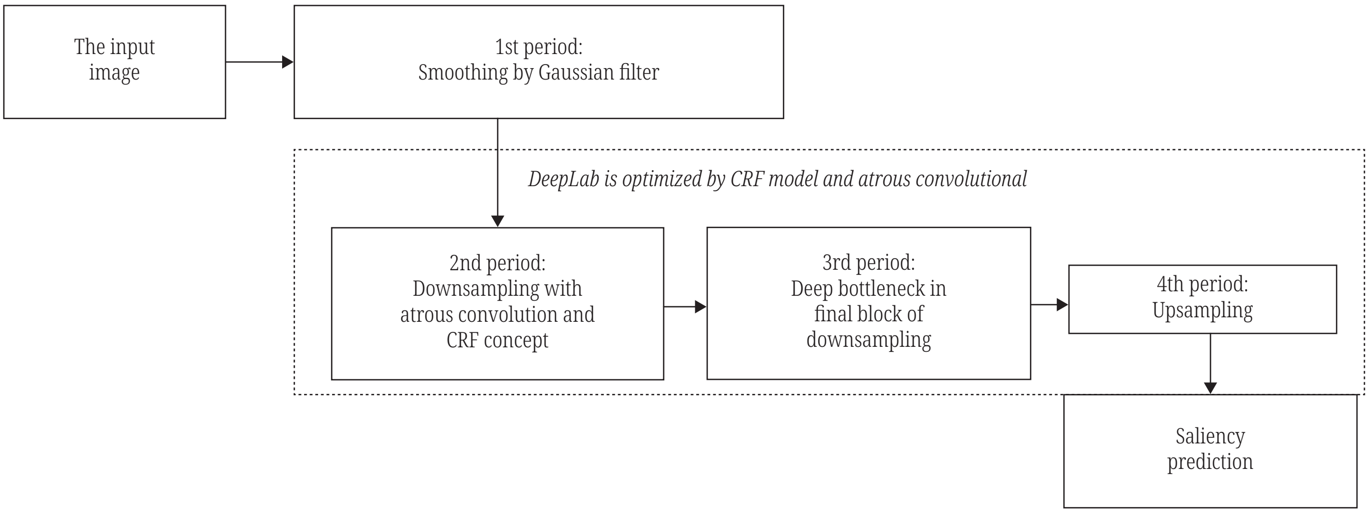

The input image
1st period: Smoothing by Gaussian filter
DeepLab is optimized by CRF model and atrous convolutional
2nd period: Downsampling with atrous convolution and CRF concept
3rd period: Deep bottleneck in final block of downsampling
4th period: Upsampling
Saliency prediction

Fig. 2. The proposed method for saliency prediction for classification

Based on Figure 2, the proposed method includes four periods:

- Firstly, the mask for the morphological activities shows the erosion and dilation based on the Gaussian filter.
- Secondly, the model based on atrous convolution and CRF is downsampled. In this period, multi-levels from the first period are the input for the first convolution layer. Changes to the number of blocks, strike, and block organization are performed. Also, the weighting for atrous convolution is applied. The aim is to enhance the extracted parameters of objects in CRF.
- Thirdly, the encoder and decoder with the improvement from CRF are applied. Although the size of the final blocks of downsampling is retained, the calculation of the values for the filter layers of the bottleneck of backbone proceeds.

– Finally, hybrid saliency is based on the similarity with blocks for backbone in upsampling. However, in this network, improving the kernel, weight layer, CRF and deep bottleneck in the encoder-decoder system is the key for feature extraction. The result is the fully connected CNN for classification.

Detailed steps are clarified below.

3.1 Smoothing for parameters based on erosion and dilation of Gaussian filter

This section presents the 1st period of the proposed method. The evaluation for the pixels' quality has been done in previous studies. However, the natural images have a wide range of elements that are given a noise/blur value. This process is a difficult problem for vision tasks. The shape or morphology of objects includes: height, width, and color channel. The size for the square matrix is applied with 512 × 512 for the input image and dividing the RGB for each level. red (the saturated), blue (the contrast between details and background) and green (the high contrast between object area with background) are vital for this task.

From the three levels of the first step, histogram equalization is the transformation for all pixels. Contrast-limited adaptive histogram equalization (CLAHE) redistributes the light values depending on the condition of histogram levels. The average number of pixels is N_{avg}, the gray levels of the dividing areas are N_{gray}, NrX and NrY are the number of pixels for the X-dimension and Y-dimension, respectively. The normalized clip limit (CL) is calculated based on the product of pixels (N_{CL}) and the clipped pixels (N_{avg}). The CL gives values in the range [0, 1] and acts as a condition for clipping of the histogram period (H_{region_clip}). The result of the clipping step is that the number of the final results is $N_{\sum clip}$. The i-th gray level of the original histogram ($H_{region}(i)$) is updated if the synthesis of the $N_{\sum clip}$ / N_{gray} and the original histogram are greater than N_{CL}. Here, the proposed method keeps the number of levels and uses the updated histogram in them.

The top-hat transform is the mask created by the square structure. The extraction concept of erosion and dilation for morphology has small size and details in each element. The white top-hat transform and the black top-hat transform are defined by the input image and the opening/closing of the structuring element. $f: E \mapsto R$ where E is a Euclidean space with grid R. The white top-hat transform is concerned about denoting the opening operation, while the black top-hat transform is the closing operation. The difference between the previous method with the present one is our use of a multi-level in the input images and the continuous processing of the histogram and top-hat transform. Based on the morphology mask in each level, the non-negative values for all pixels are detected.

After creating the mask, the adapter with filter is presented. The distribution as the standard deviation (σ) on horizontal x and vertical y axes is called the distances of them (G). In the other approaches, the Gaussian filter is used because of the relation with neighboring pixels. The multi-scale morphological filter explores this task. In this step, red-green-blue is removed and is changed into binary images. However, the multi-scale for boundary involves operators for edge detection based on the noise artifacts. The binary images that are created from the RGB removal become three windows. If we call f the input scale, the boundary detector will be equation (1):

$$boundary(f, B) = XOR[f, dilate(f, B)] \tag{1}$$

The Gaussians for each direction can be explained as:

$$g(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/(2\sigma^2)} \tag{2}$$

where x and y are the distance from origin with axis x (horizontal) and axis y (vertical), respectively. The Gaussian distribution is shown by standard deviation σ. The merged images are done with RGB of output from the boundary detector with the supercharged. The final step of the pre-processing period is the binary transfer. In this task, the sliding window with height × width is the input for convolution by matrix 5 × 5. The limit threshold for subtraction is C, which is the average of all pixels.

After five steps, the output is high-quality images. The binary images that improve the erosion and dilation are the input for DeepLab in Section 3.2. The merged results from the better values become the parameters for choosing feature extraction with deep learning. The C threshold is applied based on the average value of global images.

3.2 Optimized DeepLab for melanoma classification via hybrid saliency and conditional random field

The original DeepLab produced output shape by atrous convolution, as shown in equation (3):

$$y[i] = \sum_{i=1}^{K} x[i + r.k]w[k] \tag{3}$$

where i is the cell and w is the filter adapting with feature map x. Extraction by atrous has a distance between column and row (zero if distance = 0 and; –1) if distance = number of rate). The benefit of the atrous algorithm is the increased output. This task is useful for the vision of the filter/threshold. The aim is to support the position of the receptive field to extend higher than the cell, from 3 × 3 to 5 × 5. When a viewing area is extended, the local accuracy is preserved.

In this paper, the proposed method contributions improve the DeepLab for saliency prediction as:

- The improvement in pre-processing depends on smoothing with a Gaussian filter.
- The concept of weight layer (a building block) in atrous convolution for improving the identity of the ReLU layer for hybrid saliency.
- The CRF model for the encoder-decoder system in the bottleneck backbone for feature extraction of DeepLab.
- The loops in each level for prediction of hybrid saliency improve the approximation results.

The 2nd period is the downsampling, and atrous convolution is done with 5 blocks in the encoder process. Each block includes 3 layers, which have 3 × 3 for the matrix window (include 1 convolutional layer, 1 batch normalization and 1 ReLU layer). Activating the pooling is then done. Downsampling with 5 blocks with the size of the input image is M × M × 3 kernels. The input layer for the first block is 3 × 3 × 64 filters. The below blocks have to keep this size for two layers downsize for the pooling layer with M/2 and filters × 2. The improvement of this task is the updating for equation (3) and applies for the output of downsampling. The structure of the encoder of the proposed method is presented in Figure 3.

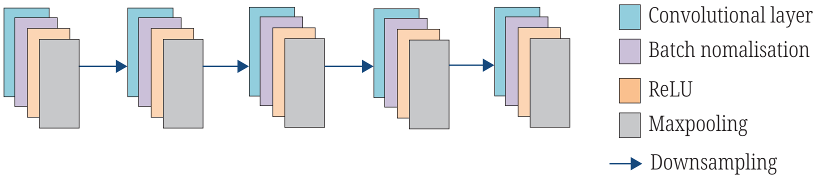


Fig. 3. The downsampling periods of the proposed method

The identity for *F(x)* of atrous convolution based on the size of the input images (in the first blocks of downsampling) becomes the size for adding the result of downsampling. The sliding window with 3 × 3 has been calculated in 10 loops for the approximation with *F(x)* present. The ReLu layer has the same size as the weight layer. This adapting to increasing the number of features had strong values from input. The benefit of atrous convolution for DeepLab lacks the information from parameters based on the sliding window. They were given only from fully connected neural networks. The weight layer is proposed in Figure 4.

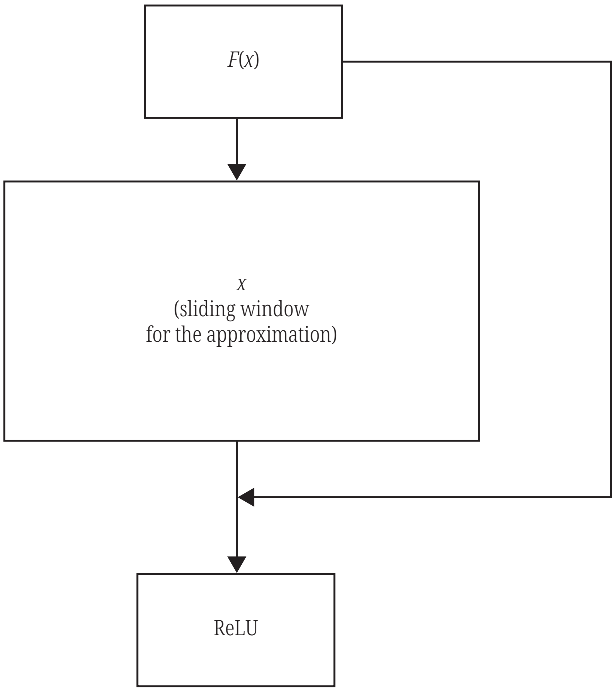


Fig. 4. Improving the atrous convolution based on weight layer

The deep bottleneck, which is the 3rd period, has the same-size input as atrous convolution and is a continuing step for feature extraction. This concept includes batch normalization for training, maintaining size, and decides whether to change window shape for concatenating of upsampling. Two bottleneck blocks have been barriers for upsampling:

- 01 batch normalization for maintaining size.
- 02 batch normalization for training (64 × 16 in the first bottleneck block and 32 × 16 in the second bottleneck block).
- Doubling the size of the window for concatenating.

Conditional random field is a statistical modeling method that does not consider the neighboring values. I_p is the image region proportional of position p of each level.

With the parent p in level $l + 1$ and q is a patch of level l. The edge that connects a patch to a parent path is defined as the below equation:

$$p = argmax_p \frac{\left|I_p \cap I_q\right|}{\left|I_q\right|} \tag{4}$$

This condition is applied before the using the first bottleneck and after the second bottleneck. The aim is the fully connected that pairwise CRF model based on the energy of the x label. The label assignment is defined as equation (5), with the label x_i, x_j for pixels i and j, respectively. $\Psi_p(x_i, x_j)$ is the connected values of two labels, $\Psi_u(x_i)$ being the single components.

$$E(x) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j) \tag{5}$$

In the final downsampling, the feature vectors are applied to the Gaussian kernel to build each potential pair. The minimizing for CRF energy is called *E(x)*, with x as the label assignment. The approximation distribution is *P(X)*, and the simpler version for this distribution is *Q(X)*:

$$Q(X) = \prod_i Q_i(X_i) \tag{6}$$

The final period of *the proposed method is the upsampling periods*, in which 5 blocks that have undergone decoding similar to the encoder block have a similar structure. From the previous periods (deep bottleneck for feature extraction), the invertible layer converts for the first blocks of upsampling. The size of layer is doubled after one block (begining with 16 × 16 to up). The output of the maxpooling layer in each encoder block becomes linked for the decoder. Then, the fully connected layer is used to generate a saliency level. The saliency map for segmentation in this final step is the prediction result from the superpixel map at level K with v iterations. From K levels, the saliency map for classification by the fully connected and CNN is applied. This period is shown in Figure 5.

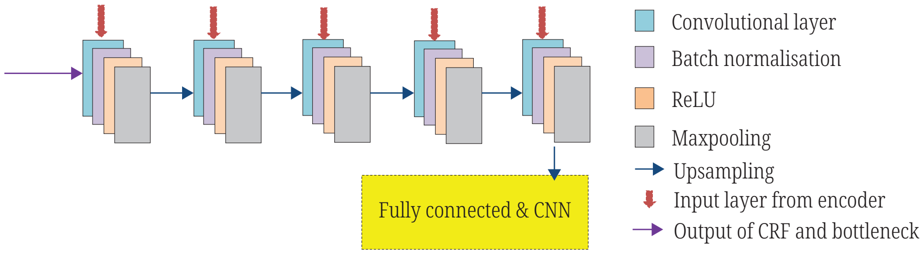


Fig. 5. The upsampling periods of the proposed method

Based on these periods of Figure 2, the image segmentation has improved the object's region exactly. The reasons for this task are the contribution of deep bottleneck and atrous with the weight layer in encoder periods. The combined areas in feature vector growth based on the CRF also have changed. The morphology is based on a Gaussian filter, which is applied to the input images as a smoothing algorithm for parameters of deep-learning models. In the proposed method, the multiple levels of iterations are useful for saliency prediction. That is the improvement for the

DeepLab system. As a result, the accuracy of saliency prediction and classification are improved after upsampling.

4 EXPERIMENTAL RESULTS

4.1 Material and dataset

To implement the experiment, the International Skin Imaging Collaboration (ISIC) 2017 [33] dataset includes 2750 images. They are lesion images in JPEG format and have 2750 corresponding superpixel masks in PNG format with EXIF data stripped. The training ground truth of this dataset has binary masking in PNG format, dermoscopic feature files in JSON format, and validated lesions for diagnosis. The aim of this ISIC 2017 dataset is the segmentation and classification about the kind of disease of patients (milia-like cyst or streaks in image). Some images of this dataset are shown in Figure 6.

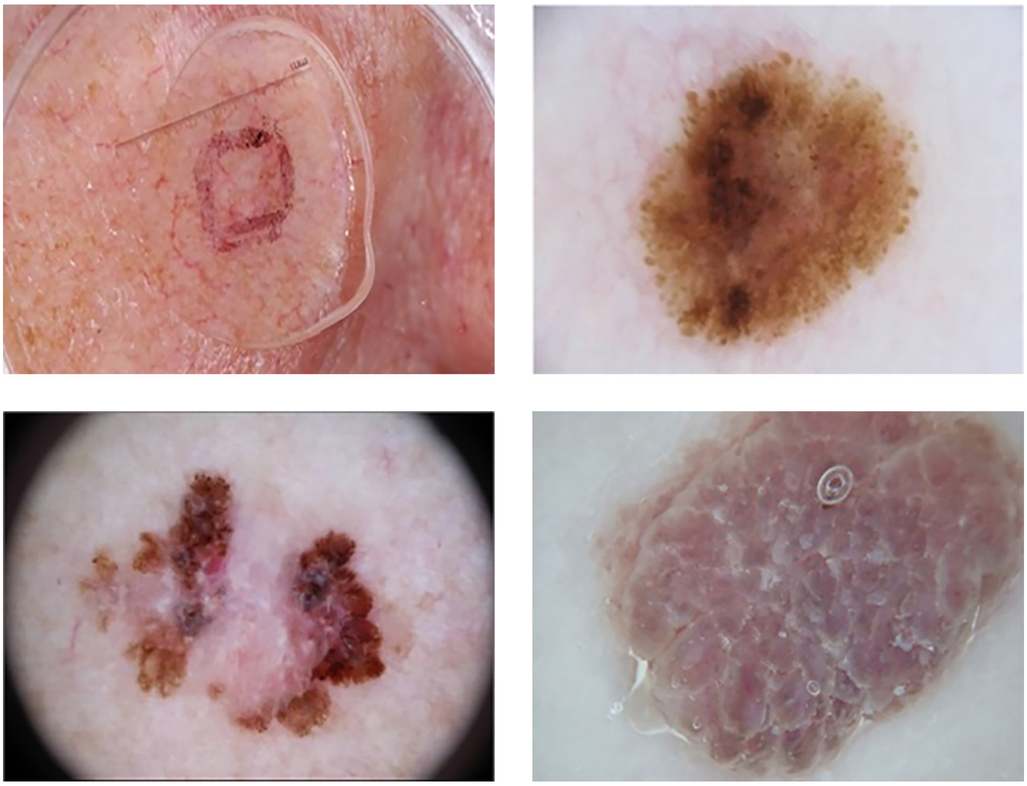


Fig. 6. Some images in ISIC 2017 [33]

The information on patients in each image includes: image id, approximate age, and sex. Each image in this dataset is one of two classes: melanoma (MEL) with 519 images, and benign (BEN) with 2231 images. The BEN includes nevus and seborrheic keratosis. The proposed method divides 2000 images for training, 600 images for testing, and 150 images for validation. The training dataset has 374 images of MEL and 1626 images of BEN. The testing dataset has 116 images of MEL and 484 images of BEN. The validation dataset has 29 images of MEL and 121 images of BEN. Each image for training begins with JPEG format and applies the binary mask based on

JSON. The final results are the classification about the status of patients based on the feature extraction. The results of comparison between the proposed method with others are presented clearly in the next section.

4.2 Results evaluation

The main contribution of our method is dependent on the weight of the encoder stage. The size keeping with deep bottleneck and CRF extracted the better features. The Jaccard Index (JI), a similarity coefficient, evaluated the segmentation results between the proposed method and other methods. This is the division between the number in both sets and the number in either set. The *JI* presented how similar the two sets are in the range from 0 to 1 (0 to 100 percent). The closer to 1, the more similarity between the results and ground truth datasets. As a result, the higher the *JI*, the better. The *JI* value was calculated as shown in equation (3):

$$JI(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{7}$$

where, *A* is the image segmentation, and *B* is the image, which has ground truth.

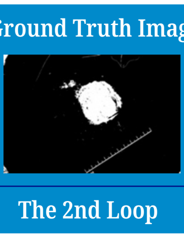
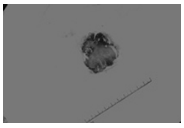
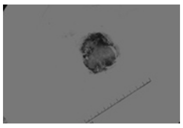
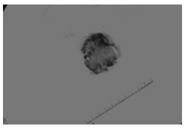
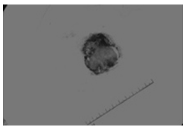
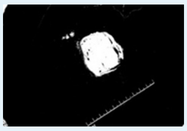
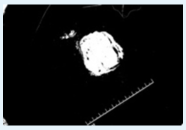
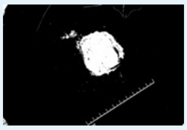
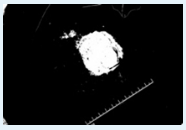
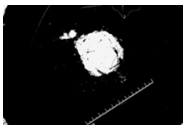
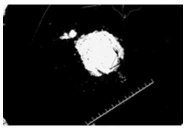
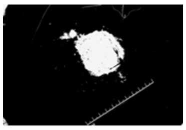
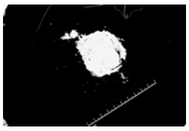
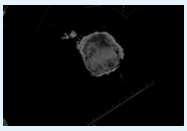
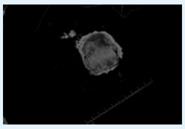
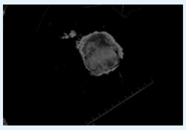
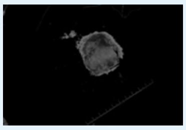
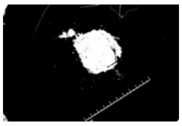
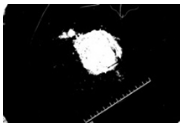
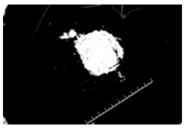
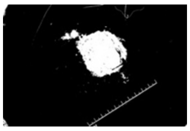
The proposed method compared the segmentation results in 5 loops for saliency prediction with: deep learning [4], DeepVS2.0 [15], bottleneck U-Net [22], and U-Net with atrous algorithm [25]. The JI value is from 0 to 100 percent. The higher JI value is, the better it is. The 5th loop was stopped when it was observed that the values were not increasing.

Table 1. Average Jaccard Index results using the proposed method and others

Number of Loops	Deep Learning [4]	DeepVS2.0 [15]	Bottleneck U-Net [22]	U-Net with Atrous [23]	Proposed Method
1	70.13	83.04	89.33	75.72	92.03
2	75.33	87.37	91.84	75.80	93.99
3	77.08	89.58	92.09	75.89	94.85
4	77.10	92.02	93.11	75.95	96.27
5	76.92	90.95	92.97	75.90	95.90

In Table 1, the results of methods have the best value at the 4th loop. The JI values using the proposed method increased with the number of loops, peaking at 96.27 at loop 4, with ground truth of the dataset. The average of JI in our approach was always better than with other methods. The drop at the 5th loop can be explained by the saturation in updating. The next comparison of the proposed method builds on the 4th-loop data from Table 1. Table 2 shows segmentation results at 4 loops for deep learning [4], DeepVS2.0 [15], Bottleneck U-Net [22], U-Net with atrous [23] and the proposed method. The evaluation was based on JI value and updated continuously after loops. The solution in this paper also has stable results. All of the methods in Table 2 were applied using the equivalent method described in Section 4.1.

Table 2. The segmentation result of the proposed method with others in 4 loops to upgrade the saliency prediction

<table>
<tr><th colspan="5">Ground Truth Image
</th></tr>
<tr><th>The Methods</th><th>The 1st Loop</th><th>The 2nd Loop</th><th>The 3rd Loop</th><th>The 4th Loop</th></tr>
<tr><td>Deep learning [4]</td><td>
JI = 71.02</td><td>
JI = 76.17</td><td>
JI = 77.60</td><td>
JI = 78.33</td></tr>
<tr><td>DeepVS2.0 [15]</td><td>
JI = 83.55</td><td>
JI = 88.14</td><td>
JI = 89.97</td><td>
JI = 93.08</td></tr>
<tr><td>Bottleneck U-Net [22]</td><td>
JI = 90.11</td><td>
JI = 92.03</td><td>
JI = 92.75</td><td>
JI = 93.26</td></tr>
<tr><td>U-Net with atrous [23]</td><td>
JI = 74.98</td><td>
JI = 75.34</td><td>
JI = 75.62</td><td>
JI = 75.81</td></tr>
<tr><td>The proposed method</td><td>
JI = 93.06</td><td>
JI = 93.86</td><td>
JI = 94.72</td><td>
JI = 96.30</td></tr>
</table>

From Table 2, the first column is the ground truth for this case. The JI values are shown from 1st to 4th loops because the best result was 4th loop. The final column is the result of this paper: the saliency prediction based on atrous and CRF to improve DeepLab for semantic segmentation has the best results. Then, the classification was done based on the results of the 4th loop of the segmentation. The results were compared with U-Net with atrous [23], the novel deep CNN with dermoscopic images [24], deep learning network based on skin lesion [25], skin lesions through deep CNN with dermoscopic [26], fusing fine-tuned of skin lesion [27], skin lesion with ensembles of deep CNN [28] and multiple skin lesions via integrated deep CNN [29]. The accuracy of classification used the below equations:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall} = \frac{TP + TN}{TP + FN} \tag{10}$$

where:

+ *TP* is the number of lesion samples correctly classified as MEL.

+ *TN* is the number of lesion samples correctly classified as BEN.

+ *FP* is the ratio of samples incorrectly classified as MEL.

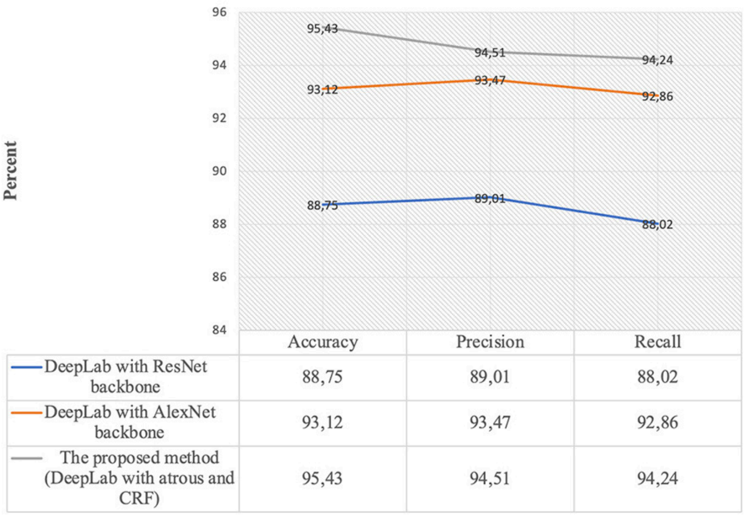
+ *FN* is the number of images determined to be BEN when they are MEL (mistake for classification).

These evaluation values (accuary, precision, and recall) are given in Table 3. The skin lesions through deep CNN with dermoscopic [26] also have the high accuracy for classification in ISIC 2017 dataset. However, the U-Net, atrous after in two periods, did not match the parameters because the weight layer was done after the decoder. Therefore, the results of [23] are not valid.

Table 3. Comparison of results about classification between the proposed method and others

Methods	Accuracy	Precision	Recall
U-Net with atrous [23]	79.88	79.92	78.86
The novel deep CNN with dermoscopic images [24]	88.23	78.55	87.86
Deep learning network based on skin lesion [25]	85.70	72.90	49.00
Skin lesions through deep CNN with dermoscopic [26]	93.25	93.97	93.25
Fusing fine-tuned of skin lesion [27]	87.70	–	87.26
Skin lesion with ensembles of deep CNN [28]	86.60	–	55.60
Multiple skin lesions via integrated deep CNN [29]	81.34	75.67	77.66
The proposed method	**95.43**	**94.51**	**94.24**

The proposed method also gives the best accuracy, precision, and recall values. To further test the proposed method, the backbone of our method was changed with ResNet and AlexNet based on [24]. This change was applied in position with the encoder and decoder of DeepLab, such that ResNet/AlexNet was used with the atrous and deep bottleneck in the last blocks. Figure 7 shows the results of this backbone changing. From this figure, we can conclude that CRF and atrous in bottleneck are a suitable backbone for the proposed method. All methods were adapted based on the 4th loop described above.



	Accuracy	Precision	Recall
DeepLab with ResNet backbone	88,75	89,01	88,02
DeepLab with AlexNet backbone	93,12	93,47	92,86
The proposed method (DeepLab with atrous and CRF)	95,43	94,51	94,24

Fig. 7. The other backbones for the proposed method

As a result, the proposed method has some contributions for saliency prediction: morphology to improve the sharpness in pre-processing with a Gaussian filter, DeepLab with a backbone that is CRF, and atrous in bottleneck, which is identified by downsampling with the weight layer in atrous convolution and bottleneck for upsampling, calculating for the approximate saliency. The original values can be changed to adapt training by pre-processing, which includes top-hat transform and Gaussian filter. Smoothing in the first step, as it keeps the strong pixels for deep bottleneck and atrous convolution in the encoder process. Feature extraction in this research fits with a sliding window for the approximately in ReLu layer and deep extraction from the encoder stage. The continuous updates bring the predicted results closer to the actual results. The synthesis of features is a vital parameter for segmentation and final classification. These contributions improve the results of the proposed method.

5 CONCLUSIONS

DeepLab, which is a deep convolutional neural network, synthesises features from decomposition and reversion. The parameters at each stage have a wide range of distinct features to classify. With the CRF model, these are the same for the encoder and decoder stages. The proposed method proposed on this approach by adding a the weight layer. Here, the atrous convolutional and deep bottleneck are implemented for saliency prediction. Comparison with the other recent methods verifies that our proposed method is suitable. The results were obtained by continuous updating in 4 loops for saliency prediction. The backbone was changed with ResNet and AlexNet to check the accuracy of the proposed method. In the future, the detection automation for the number of blocks is essential.

6 ACKNOWLEDGEMENT

We acknowledge the support of time and facilities from Ho Chi Minh City University of Foreign Languages and Information Technology and Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, for this study.

7 REFERENCES

[1] M.A.A. Halim, N.I.R. Ruhaiyem, E.R.I. Fauzi, M.S.C. Jamil and A.S.A. Mohamed (2016). Automatic laser welding defect detection and classification using Sobel-contour shape detection. Journal of Telecommunication, Electronic and Computer Engineering, 8(6): 157–160. https://jtec.utem.edu.my/jtec/article/view/1266

[2] Fausto Milletari, Nassir Navab and Seyed-Ahmad Ahmadi (2016). V-Net: Fully Convolutional Neural Networks for Volumetric medical image segmentation. 3D Vision, Fourth International conference on, IEEE. https://doi.org/10.1109/3DV.2016.79

[3] M. Vardhana, N. Arunkumar, Sunitha Lasrado, Enas Abdulhay and Gustavo Ramirez-Gonzalez (2018). Convolutional neural network for bio-medical image segmentation with hardware acceleration. Cognitive Systems Research, 50: 10–14. https://doi.org/10.1016/j.cogsys.2018.03.005

[4] Holger R. Roth, Chen Shen, Hirohisa Oda, Masahiro Oda, Yuichiro Hayashi, Kazunari Misawa and Kensaku Mori (2018). Deep learning and its application to medical image segmentation. Medical Imaging Technology, 36(2), 63–71. https://doi.org/10.11409/mit.36.63

[5] Junting Pan, Kevin McGuinness, Elisa Sayrol, Noel O'Connor and Xavier Giro-i-Nieto (2016). Shallow and deep convolutional networks for saliency prediction. Computer Vision and Pattern Recognition. https://arxiv.org/abs/1603.00845

[6] S.J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz and B. Schiele (2017). Exploiting Saliency for Object Segmentation from Image Level Labels. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 5038–5047. https://doi.org/10.1109/CVPR.2017.535

[7] G. Zeng (2017). Fruit and Vegetables Classification System Using Image Saliency and Convolutional Neural Network. IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC), Chongqing, China, 613–617. https://doi.org/10.1109/ITOEC.2017.8122370

[8] Yue Zhu, Baochen Hao, Baohua Jiang, Rui Nian, Bo He, Xinmin Ren and Amaury Lendasse (2017). Underwater Image Segmentation with Co-Saliency Detection and Local Statistical Active Contour Model. In Proceedings of the OCEANS 2017 – Aberdeen, IEEE, 1–5. https://doi.org/10.1109/OCEANSE.2017.8084742

[9] Wenguan Wang, Jianbing Shen, Xingping Dong and Ali Borji (2018). Salient Object Detection Driven by Fixation Prediction. IEEE/CVF Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/CVPR.2018.00184

[10] Sen He and Nicolas Pugeaulty (2018). Salient region segmentation. Computer Vision and Pattern Recognition, 1–6.

[11] H. Bi, H. Tang, G. Yang, H. Shu and J.-L. Dillenseger (2018). Accurate image segmentation using Gaussian mixture model with saliency map. Pattern Analysis and Applications, 21: 869–878. https://doi.org/10.1007/s10044-017-0672-1

[12] Bashir Ghariba, Mohamed S. Shehata and Peter McGuire (2019). Visual saliency prediction based on deep learning. Information, 10(8): 257. https://doi.org/10.3390/info10080257

[13] E. Iqbal, A. Niaz, A. A. Memon, U. Asim and K. N. Choi (2020). Saliency-Driven Active Contour Model for Image Segmentation. In IEEE Access, 8: 208978–208991. https://doi.org/10.1109/ACCESS.2020.3038945

[14] Jun Zhang, Yamei Liu, Shengping Zhang, Ronald Poppe and Meng Wang (2020). Light field saliency detection with deep convolutional networks. IEEE Transactions on Image Processing, 29: 4421–4434. https://doi.org/10.1109/TIP.2020.2970529

[15] Lai Jiang, Mai Xu, Zulin Wang and Leonid Sigal (2021). DeepVS2.0: A Saliency-structured deep learning method for predicting dynamic visual attention. International journal of computer vision, 129: 203–224. https://doi.org/10.1007/s11263-020-01371-6

[16] Hemraj Singh and Maheep Singh (2021). Salient object detection on Matrix disintegration using convolutional neural network. Indian Journal of VLSI Design (IJVLSID), 1(1): 8–14.

[17] Alexander Kroner, Mario Senden, Kurt Driessens and Rainer Goebel (2020). Contextual encoder – decoder network for visual saliency prediction. Neural Networks, 129: 261–270. https://doi.org/10.1016/j.neunet.2020.05.004

[18] Ziqiang Wang, Zhi Liu, Weijie Wei and Huizhan Duan (2021). SalED: Saliency prediction with a pithy encoder – decoder architecture sensing local and global information. Image and Vision Computing, 109: 104149. https://doi.org/10.1016/j.imavis.2021.104149

[19] C.A.R. Goyzueta, J.E.C. De la Cruz and W.A.M. Machaca (2021). Integration of U-Net, ResU-Net and DeepLab Architectures with Intersection Over Union metric for Cells Nuclei Image Segmentation. IEEE Engineering International Research Conference (EIRCON), 1–4. https://doi.org/10.1109/EIRCON52903.2021.9613150

[20] Vo Thi Hong Tuyet, Nguyen Thanh Binh, Nguyen Kim Quoc and Ashish Khare (2021). Content based medical image retrieval based on salient regions combined with deep learning. Journal of Mobile Networks and Applications, Springer. https://doi.org/10.1007/s11036-021-01762-0

[21] C. Wang, R. Zhang and L. Chang (2022). A study on the dynamic effects and ecological stress of eco-environment in the headwaters of the Yangtze river based on improved DeepLab V3+ network. Remote Sensing, 14(9): 2225. https://doi.org/10.3390/rs14092225

[22] V.T.H. Tuyet, N.T. Binh and D.T. Tin (2022). A deep bottleneck U-Net combined with saliency map for classifying diabetic retinopathy in fundus images. International Journal of Online and Biomedical Engineering (iJOE), 18(02): 105–122. https://doi.org/10.3991/ijoe.v18i02.27605

[23] M.M. Stofa, M.A. Zulkifley, M.A.A.M. Zainuri and A.A. Ibrahim (2022). U-Net with Atrous Spatial Pyramid Pooling for Skin Lesion Segmentation. Proceedings of the 6th International Conference on Electrical, Control and Computer Engineering, Lecture Notes in Electrical Engineering, Springer, 842. https://doi.org/10.1007/978-981-16-8690-0_89

[24] Ranpreet Kaur, Hamid GholamHosseini, Roopak Sinha and Maria Linden (2022). Melanoma classification using a novel deep convolution neural network with dermoscopic images. Sensors, 22(3): 1134. https://doi.org/10.3390/s22031134

[25] Y. Li and L. Shen (2021). Skin lesion analysis towards melanoma detection using deep learning network. Sensors, 18: 556. https://doi.org/10.3390/s18020556

[26] I. Iqbal, M. Younus, K. Walayat, M.U. Kakar and J. Ma (2021). Automated multi-class classification of skin lesions through deep convolutional neural network with dermoscopic images. Computerized Medical Imaging and Graphics, 88: 101843. https://doi.org/10.1016/j.compmedimag.2020.101843

[27] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot and C. Wang (2019). Fusing fine-tuned deep features for skin lesion classification. Computerized Medical Imaging and Graphics, 71: 19–29. https://doi.org/10.1016/j.compmedimag.2018.10.007

[28] B. Harangi (2018). Skin lesion classification with ensembles of deep convolutional neural networks. Journal of Biomedical Informatics, 86: 25–32. https://doi.org/10.1016/j.jbi.2018.08.006

[29] M.A. Al-Masni, D.H. Kim and T.S. Kim (2020). Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. Computer Methods and Programs in Biomedicine, 190: 105351. https://doi.org/10.1016/j.cmpb.2020.105351

[30] H. Vega-Huerta, R. Villanueva-Alarcón, D. Mauricio, J. Gamarra Moreno, H.D. Calderon Vilca, D. Rodriguez and C.RodriguezH. Vega-Huerta, R. Villanueva-Alarcón, D. Mauricio, J. Gamarra Moreno, H.D. Calderon Vilca, D. Rodriguez and C.Rodriguez (2022). Convolutional neural networks on assembling classification models to detect melanoma skin cancer. International Journal of Online and Biomedical Engineering (iJOE), 18(14): 59–76. https://doi.org/10.3991/ijoe.v18i14.34435

[31] Z. F. Shaaf, Muhammad Mahadi Abdul Jamil and Radzi Ambar (2023). A convolutional neural network model to segment myocardial infarction from MRI images. International Journal of Online and Biomedical Engineering (iJOE), 19(02): 150–162. https://doi.org/10.3991/ijoe.v19i02.36607

[32] Malek AI-Nawashi, Obaida M. AI-Hazaimeh and Mohamad Saraee (2017). A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments. Neural Computing and Applications, 28(Suppl 1): 565–572. https://doi.org/10.1007/s00521-016-2363-z

[33] https://challenge.isic-archive.com/data/ (lass accessed March 12, 2023).

8 AUTHORS

Vo Thi Hong Tuyet received the Bachelor of Pedagogical in Informatics degree from Ho Chi Minh City University of Pedagogy and the Master of Technology degree in computer science from Ho Chi Minh City University of Technology, Vietnam

National University in Ho Chi Minh City (VNU-HCM), Vietnam, in 2011 and 2015, respectively. Now, she is a lecturer at Faculty of Information Technology, Ho Chi Minh City University of Foreign Languages and Information Technology (HUFLIT), Vietnam. She is working as a PhD student at the Faculty of Computer Science and Engineering, the Ho Chi Minh City University of Technology (VNU-HCM), Vietnam. Her research interests include recognition, image processing (email: vthtuyet.sdh19@hcmut.edu.vn, tuyetvth@huflit.edu.vn).

Nguyen Thanh Binh received the Bachelor of Engineering degree from Ho Chi Minh City University of Technology, Vietnam National University Ho Chi Minh City (VNU-HCM), Vietnam, in 2000, and the Master's degree and Ph.D. degree in computer science, both from University of Allahabad, India, in 2005 and 2011, respectively. Now, he is an Associate Professor in the Ho Chi Minh City University of Foreign Languages and Information Technology, Vietnam. He has published one book, one book chapter, and more than 70 research papers. His research interests include recognition, image processing, multimedia information systems, decision-support systems, and time series data (email: binh@huflit.edu.vn).