RESEARCH ARTICLE

Lung Tumor Localization and Visualization in Chest X-Ray Images Using Deep Fusion Network and Class Activation Mapping

ADE IRMA SURYANI[1], CHUAN-WANG CHANG[2], YU-FAN FENG[1], TIN-KWANG LIN[3,4], CHIH-WEN LIN[4,5], JEN-CHIEH CHENG[6], AND CHUAN-YU CHANG[1,6], (Senior Member, IEEE)

[1]Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Yunlin, Douliou 64002, Taiwan
[2]Department of Computer Science and Information Engineering, National Chin-Yi University of Technology, Taichung 41170, Taiwan
[3]Department of Internal Medicine, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Chiayi 622, Taiwan
[4]School of Medicine, Tzu Chi University, Hualien 622, Taiwan
[5]Department of Medical Imaging, Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, Chiayi 970, Taiwan
[6]Service Systems Technology Center, Industrial Technology Research Institute, Hsinchu 310401, Taiwan

Corresponding author: Chuan-Yu Chang (chuanyu@yuntech.edu.tw)

ABSTRACT Chest X-ray is a radiological clinical assessment tool that has been commonly used to detect different types of lung diseases, such as lung tumors. In this paper, we use the Segmentation-based Deep Fusion Networks and Squeeze and Excitation blocks for model training. The proposed approach uses both wholes and cropped lung X-ray images and adds an attention mechanism to address the problems encountered during lesion identification, such as image misalignments, possible false positives from irrelevant objects, and the loss of small objects after image resizing. Two CNNs are used for feature extraction, and the extracted features are stitched together to form the final output, which is used to determine the presence of lung tumors in the image. Unlike previous methods which identify lesion heatmaps from X-ray images, we use the Semantic Segmentation via Gradient-Weighted Class Activation Mapping (Seg-Grad-CAM) to add semantic data for improved lung tumor localization. Experimental results show that our method achieves 98.51% accuracy and 99.01% sensitivity for classifying chest X-ray images with and without lung tumors. Furthermore, we combine the Seg-Grad-CAM and semantic segmentation for feature visualization. Experimental results show that the proposed approach achieves better results than previous methods that use weakly supervised learning for localization. The method proposed in this paper reduces the errors caused by subjective differences among radiologists, improves the efficiency of image interpretation and facilitates the making of correct treatment decisions.

INDEX TERMS Lung tumor, chest X-ray, Seq-Grad-CAM.

I. INTRODUCTION

X-rays have been widely used clinically to detect lesions in bones or soft tissues of organs to assist in diagnosing diseases. As a result, image recognition technologies are crucial in clinical examinations. Professional training and experience are required to mark possible lesion areas in X-ray images. However, radiologists may misjudge due to insufficient experience and pressure from work, consequently affecting the accuracy

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

of diagnosis and the treatment of patients. In this work, we propose the application of automated or computer-assisted deep learning tools to prevent misjudgments due to the lack of experience, stress, or fatigue among radiologists. The proposed automated image recognition tools accurately detect the location of lesions in X-ray images, assist the doctors in interpreting image data and improve the overall quality of clinical care.

According to a WHO report [1], cancer is the leading cause of death worldwide, accounting for nearly 10 million deaths in 2020. Specifically, later diagnoses of lung cancer

lead to higher mortality rates. In 2020, 2.26 million people were diagnosed with lung cancer, and 1.8 million people died from it. Lung cancer is often related to the patient's lifestyle and environment and can be treated through surgery if it is detected in time. However, early-stage symptoms such as coughing, weight loss, hemoptysis, and sudden fevers are difficult to detect and notice in time. Available data indicate that 80% of patients have already missed the golden treatment period upon diagnosis. Thus, early cancer detection is a great challenge for doctors and medical professionals. While doctors or radiologists undergo professional training and practice to interpret X-ray images correctly, they often suffer from insufficient clinical experience, work pressure, fatigue, and other factors that adversely affect the accuracy of image interpretation.

The use of computer-aided diagnosis (CADx) in the screening and diagnosis of cancer from X-ray images has become a trend with the advancement of information technology. We believe that the development of a tool for computer-aided detection of lung tumors with high sensitivity and low false-positive rates can assist physicians or radiologists in providing positive clinical diagnoses.

In this paper, we combine deep learning models such as the CNN, U-Net, and Seg-Grad-CAM (Semantic Segmentation via Gradient-Weighted Class Activation Mapping) to propose a classification and localization system for detecting lung tumors from chest X-ray images. The main contributions of this paper are as follows:

1. We add an SE block (Squeeze and Excitation block) attention mechanism to improve the performance of lung tumor image classification. The experimental results show that the sensitivity is 99.01%, and the accuracy is 98.51%.
2. We incorporate a Seg-Grad-CAM for lesion visualization, which differs from state-of-the-art methods in that our method outputs more precise tumor locations instead of heatmaps of possible lesions.
3. The developed lung tumor detection software has been clinically implemented in Tzu Chi Hospital to assist doctors and radiologists in interpreting lung tumors in chest X-ray images.
4. Lung tumors can be detected by combining relatively inexpensive X-ray photography and the proposed lung tumor detection software without the need for expensive MRI equipment.

The rest of this paper is organized as follows. First, we briefly describe the related research in Section II. Then the proposed method is described in detail in Section III. Next, we present the experimental results and discuss the results in Section IV, and finally, the conclusions are delivered in Section V.

II. RELATED WORKS
The early detection of lung tumors by reading chest X-ray images is important for the curative treatment of the disease.

In particular, there is a high demand for diagnostic support systems that provide accurate detection of lung tumors to reduce the risk of missed lung tumor diagnoses.

Traditional lung disease detection using image processing techniques has been investigated in detail by Mary et al. [2]. They concluded that computerized classification and detection of lung images consist of five stages: preprocessing, segmentation, feature selection, feature extraction and classification stages.

Abed [3] proposed a system for detecting lung tumors from X-ray images using the principal component analysis (PCA) with a traditional backpropagation neural network (BPNN). The main benefit of using the PCA for feature extraction is to minimize the dimensionality of training images, improve the recognition results of ANNs and reduce the execution time.

In recent years, the combination of artificial intelligence and deep learning technology has provided one of the most popular and effective solutions. Yahyatabar et al. [4] used a deep CNN model called the Dense-Unet to segment regions within the lungs. In this approach, the information flow across the network was increased, and the network parameters were reduced while maintaining the robustness of the segmentation through dense connections between layers.

Ausawalaithong et al. [5] used a 121-layer convolutional neural network, also known as the DenseNet-121, combined with a transfer learning method for lung cancer classification of chest X-ray images. Their proposed model yielded an average accuracy of 74.43±6.01% and an average sensitivity of 74.68±15.33%.

Wang et al. [6] used a weakly supervised learning approach to classify 14 different classes of lung diseases on a large public dataset of chest X-rays - "ChestX-ray14". A heatmap of the lesion area in the X-ray image was detected using a feature visualization technique. They achieved 69.3% lung tumor classification accuracy on "ChestX-ray14" using the ResNet-50 architecture [7]. Rajpurkar et al. [8] achieved 86.8% accuracy in tumor classification using a fine-tuned DenseNet-121 [9] model with a Sigmoid activation function.

Since the approaches in [6] and [7] used the entire chest X-ray image for training, the model loses excessive pixel features in the convolution process, which adversely affects the model's performance. Therefore, Guan et al. [10] added the use of a class activation mapping [11] as an attention mechanism to obtain the image input as a local lesion region and achieved a final tumor classification accuracy of 82.1%.

The three previously described approaches [6], [8], [10] perform class activation mapping directly on the entire chest X-ray image to detect lesion areas in X-ray images. Liu et al. [12] changed the strategy of obtaining local images. First, the U-Net [13] model was used to predict the location of the chest X-ray lung region. A series of post-processing steps were then performed to obtain the local images. The resultant classification accuracy rate of tumors is 81.5%.

Although Liu et al. [12] used whole chest X-ray images and images of lung regions that improved the accuracy of lesion detection, the visualization results of the lesion area

obtained by weakly supervised learning are ambiguous and inaccurate. In this paper, we propose the addition of more accurate pixel-level labels for model learning and use the Seg-Grad-CAM [14] for semantic segmentation to obtain a more accurate visualization interpretation.

Although clinically, when a lung tumor is detected on a chest X-ray image, CT or MRI is used for further confirmation. Li et al. [15] proposed an MRI lung tumor segmentation model consisting of a cross-modal synthesis network and a multi-modal segmentation network (Res-Unet). Based on the principle of GAN, Jiang et al. [16] proposed a joint probabilistic segmentation and image distribution matching generative adversarial network (PSIGAN) for lung tumor segmentation from MRI images. Jiang et al. [17] also proposed a cross-modal technique with segmentation networks called teacher and student combined with image-to-image translation for lung tumor segmentation.

In this paper, we present a method for detecting lung tumors from X-ray images alone, without the need for expensive CT or MRI equipment, and with high accuracy. This is of great help for the early detection and treatment of lung tumors.

III. METHOD

A. SYSTEM ARCHITECTURE FLOWCHART

We implement a classification and localization system to evaluate the feasibility and effectiveness of the proposed method. The system is comprised of two phases: the classification phase and the localization phase, as shown in Figure 1.

In the classification phase, the whole chest X-ray image and its corresponding lung area image are used as input. The classification CNN model is used to identify whether there is a lung tumor, and a prediction result is provided to the doctor.

In the localization phase, the location of the lesion area is segmented according to any cropped chest X-ray image, and the results predicted by the localization CNN model are visually interpreted using the Seg-Grad-CAM framework. The purpose of using the Seg-Grad-CAM [14] framework is to intercept the convolution of bottleneck from the localization network and its prediction result and obtain the final visualization by the linearly weighted summation of gradient backpropagation.

B. LUNG REGION CROPPING

Chest X-ray images taken by different radiologists from different patients often contain black backgrounds, areas outside of the lungs or tilted images, as shown in Figure 2. However, important features may be lost if the images are directly compressed for training due to the high resolution of chest X-ray images. In addition, if an uncropped chest X-ray image with black background is directly sent to the network for training, the model will not perform well and cannot extract useful features.

In this paper, we extract lung regions from the whole X-ray image for subsequent operations to improve the accuracy of
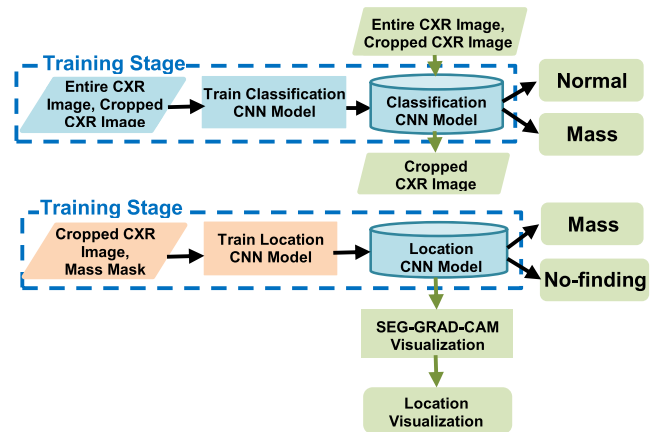

FIGURE 1. Flowchart for X-ray tumor classification and localization.
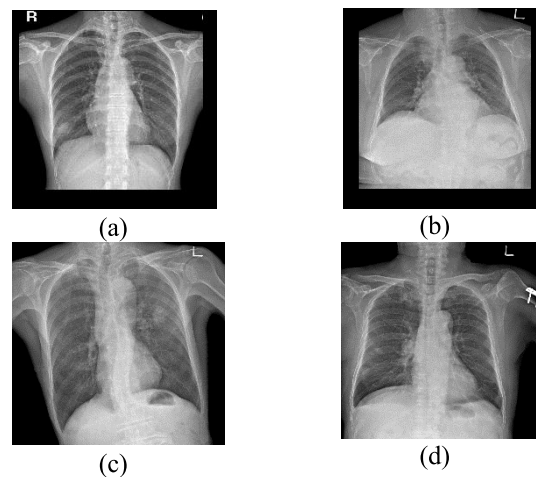

FIGURE 2. (a) (b) black borders in X-ray images, (c) a tilted X-ray image and (d) the presence of black borders and staples in the image.

lung tumor identification. We input the whole X-ray image into a trained lung localization network, and post-processing is performed to obtain a cropped X-ray image containing only the lung region.

1) LUNG REGION CROPPING PROCESS

Figure 3 shows the flowchart to obtain a cropped chest X-ray image containing only the lung area. First, a U-Net model [13] for locating the lung region is trained using a whole chest X-ray image and its corresponding mask of the lung region. After that, an arbitrarily chest X-ray image is sent to the trained U-Net model to obtain the prediction result of its lung area. A series of post-processing is then performed according to the result.

The post-processing process counts the number of all contour regions in the prediction result, calculates their areas respectively, and finds the largest contour among all contour regions. Intestinal gas and lung air may be present in a chest X-ray image, which can cause the U-Net model to predict the wrong contour region. We delete contour regions with areas
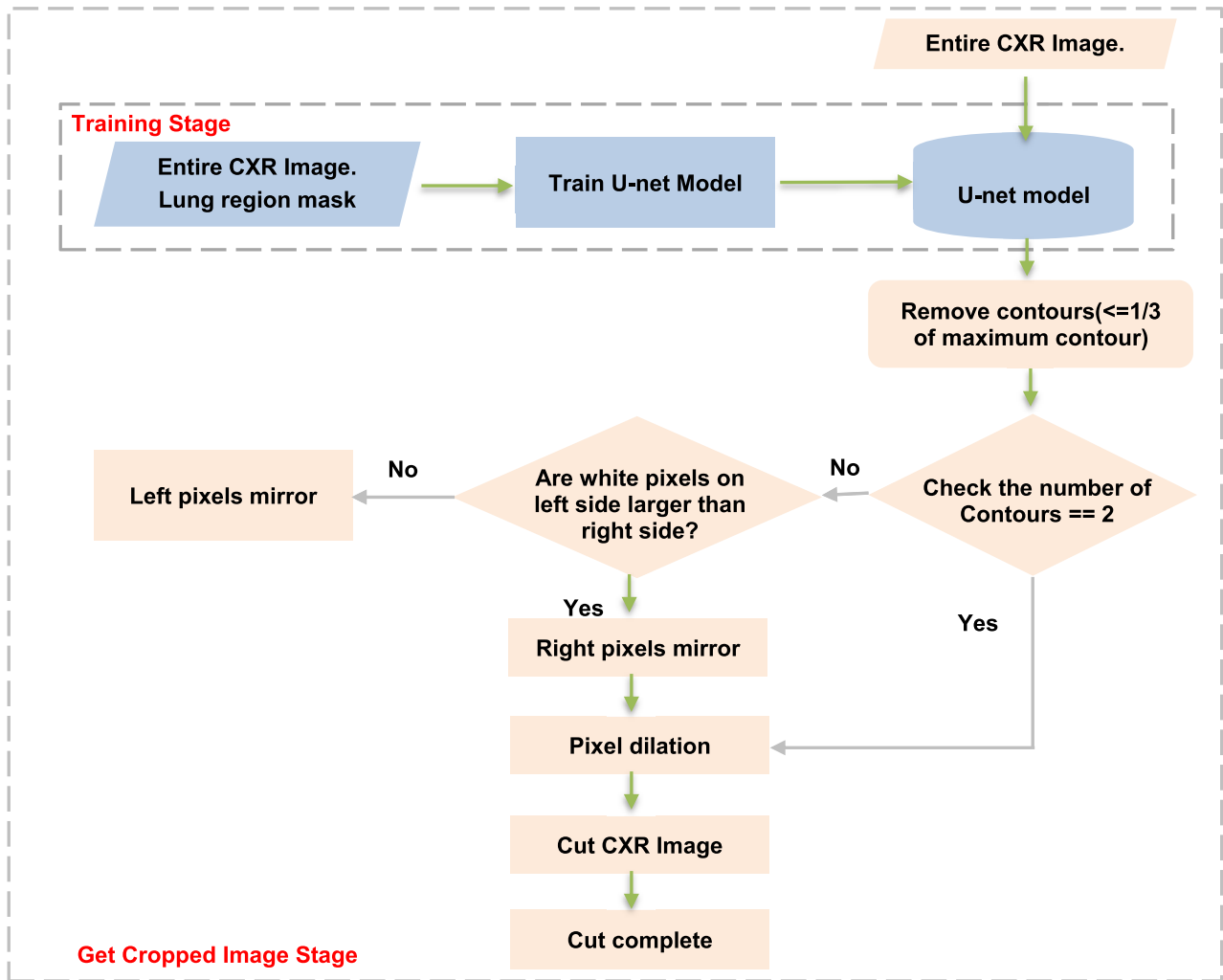
FIGURE 3. Flowchart for cropping the lung region from X-ray images.

less than 1/3 of the maximum contour area and recalculate the number of contours after deletion. If the remaining contour number is less than 2, the original chest X-ray image may not be able to capture the complete left and right lung regions due to poor image quality or lung disease. In general, the left and right lobes of the lungs are symmetrical. We calculate the contour areas of the left and right lungs according to the midline of the chest X-ray image. The side with the larger lung area is selected and mirrored to the left or right to obtain two complete lung regions. If the number of contours is equal to two, the detected left and right lung regions are dilated to obtain the cropped chest X-ray image at the relative position of the original chest X-ray image, according to the maximum and minimum coordinates of the white pixel block.

2) LUNG REGION LOCALIZATION MODEL

Figure 4 shows the U-Net [13] model architecture for locating the lung region in a chest X-ray image. A chest X-ray image of size 224×224 pixels is given as the input, and the corresponding lung region mask is given as the label

during training to learn the characteristics of the lung region in a chest X-ray image. The lung region of the image is predicted, and the size of the output image is $2 \times 224 \times 224$. The U-net model performs down-sampling four times during compression and up-sampling four times during expansion to ensure that the features are of the same sizes and can be stitched together.

3) DILATION

In the post-processing process, we use a dilation technique to enlarge the predicted lung area to obtain a complete image of the lung region. Dilation is a basic morphological operation that convolves a selected kernel B based on a part of the image area A to find the local maximum, as given by Eq. (1).

$$A \oplus B = \{z | (\hat{B})_z \cap A \neq \emptyset\}, \qquad (1)$$

where z represents the set of pixel values in the binarized image. The shape of kernel B can be square or circular. When the target, A, is inflated by the kernel, B, the target becomes
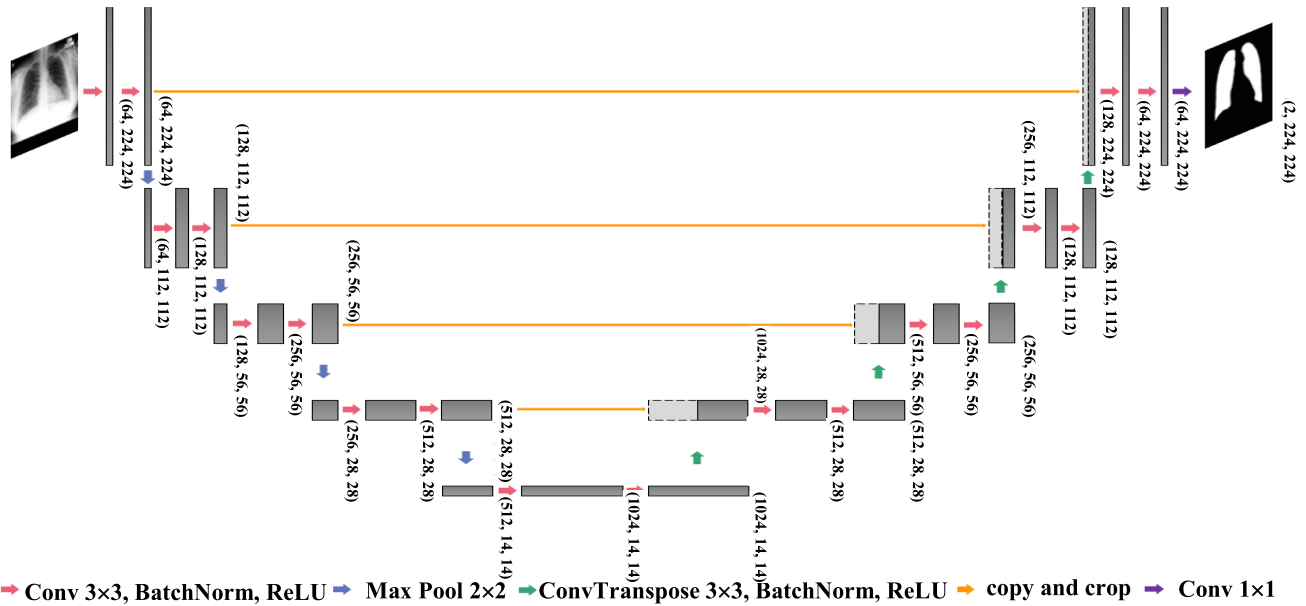
→ **Conv 3×3, BatchNorm, ReLU** → **Max Pool 2×2** → **ConvTranspose 3×3, BatchNorm, ReLU** → **copy and crop** → **Conv 1×1**

FIGURE 4. Diagram of the lung localization U-net model.

larger. Figure 5(a) is an image of the lung region without dilation, and Figure 5(b) is an image after dilations.

4) NEGATIVE LOG-LIKELIHOOD LOSS

In this paper, we use the Negative Log-Likelihood (NLL) Loss to train the lung localization U-Net model, as given by Eq.(2).

$$Loss_{NLL}(t, y) = -\sum_i t_i \log y_i, \qquad (2)$$

where t is the contour of the lung region in the chest X-ray image, y is the result predicted by the U-Net model after a Softmax operation, and i is the category used to distinguish the background from the lung region. The significance of calculating the NLL loss is that the smaller the loss value, the higher the similarity between the prediction result and the original label and vice versa.

C. CHEST X-RAY NORMAL / LUNG TUMOR CLASSIFICATION NETWORK

1) THE SE-SDFN CLASSIFICATION MODEL

In this paper, we propose a more efficient SE-SDFN (SE-Segmentation-based Deep Fusion Network) classification model by integrating the architectures of DenseNet-121 [11] and SDFN [12]. The integrated model provides a quick classification of lung tumors from chest X-ray images and assists physicians to improve the efficiency of diagnosis.

As shown in Figure 6, the proposed SE-SDFN model contains two modified SE-DenseNet-121 networks, and each has 7 SE blocks [18] (indicated by the red blocks). The input of the SE-SDFN model is a whole chest X-ray image and a cropped lung X-ray image. Each of the two images is fed into a modified SE-DenseNet-121 network. The corresponding

(a)   (b)
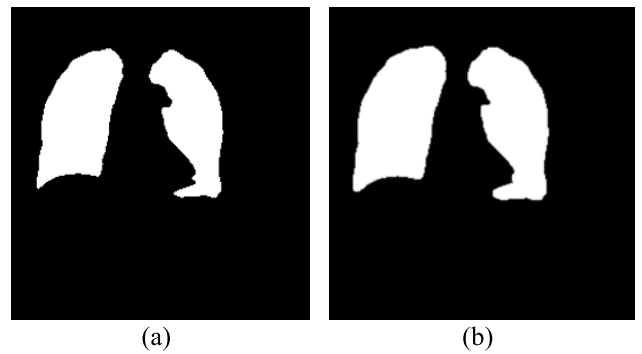
FIGURE 5. Images of the lung region before and after dilation.

lung region image is automatically cropped and generated by the U-Net model and post-processing processes.

The SE blocks added to the classification model play different roles according to their respective positions in the model. The SE block at a higher level extracts the features related to the class, and the SE block at a lower level shares the various features of the class. The proposed model allows the model to perform feature reconstruction based on the information in different feature maps, thereby enlarging major features and ignoring minor features to improve the model's accuracy. Our proposed model avoids focusing on unimportant features and concentrates on extracting effective features in the lung region. Finally, the output from the global average pooling layer of the two SE-DenseNet-121 is concatenated, and the two-class result is obtained through the fully connected layer and the Sigmoid activation function.

2) SE BLOCK

The structure of the SE block (Squeeze and Excitation block) is shown in Figure 7. During the training stage, the SE blocks
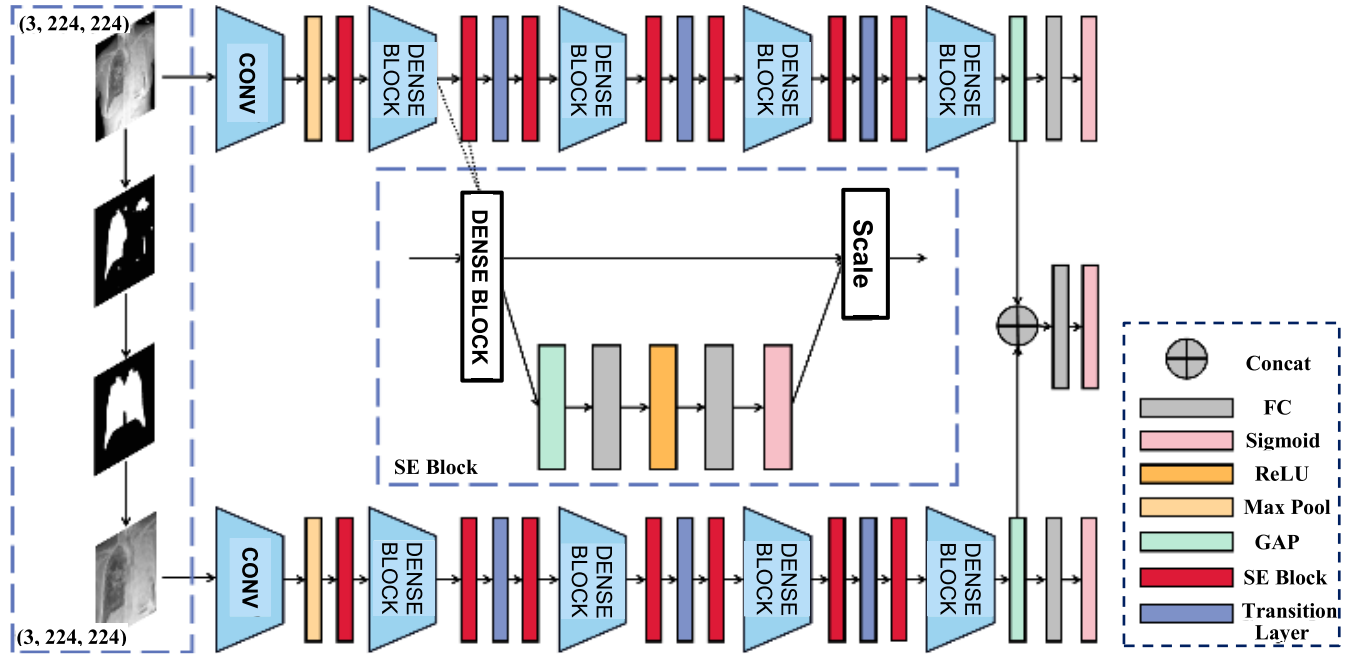
FIGURE 6. Flowchart of the model classification framework, SE-SDFN.

perform feature reconstruction based on the information in different feature maps, enlarging major features and ignoring minor features to improve the model accuracy. When the input feature map, X, is subjected to the operation, F_{tr}, of the convolutional layer, U will be obtained, where $U = \{u_1, u_2, \ldots u_c\}$. The purpose of the SE block is to improve feature extraction through channel recalibration that includes two steps: ***Squeeze*** and ***Excitation***.

a: SQUEEZE

The squeeze operation is performed through F_{sq} operation, wherein the global information of U on the channel is extracted. The purpose of the global average pooling (GAP) is to extract the global information of the feature map and obtain a feature map with a size of $c \times 1 \times 1$, as given by Eq. (3):

$$z_c = F_{sq}(U_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j), \qquad (3)$$

where H, W are the height and width of the feature map, i, j corresponds to the pixel coordinates on the feature map, c is the number of channels, and z_c is the extracted channel descriptor.

b: EXCITATION

A series of excitation operations are performed after the squeeze operation. The purpose is to learn the importance of the channel descriptors obtained by the squeeze operation that also reflects the importance of each channel on the original U. The extracted channel descriptor, z_c, first passes through a linear layer W_1 to compress the feature channel to the original c/r times, where r is equal to 16. The ReLU activation
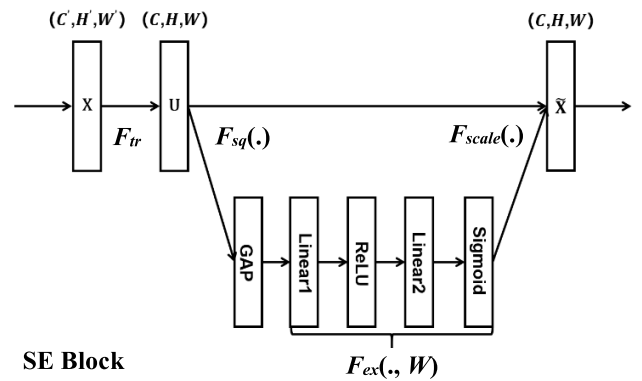
FIGURE 7. The structure of the Squeeze and Excitation block.

function, δ, is then used to activate and restore the number of feature channels through the linear layer, W_2. The Sigmoid function, σ, is used for further activation. In practice, these two linear layers are both FC layers. The entire activation process is given by Eq(4):

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \delta(W_1 z)). \qquad (4)$$

Finally, the output from the SE block is the result of channel-wise multiplication of the feature, s_c, obtained by Excitation and the original input, u_c, as given by Eq.(5).

$$\tilde{x}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c. \qquad (5)$$

3) LOSS FOR THE CLASSIFICATION MODEL

We use the Binary Cross Entropy (BCE) to calculate the loss since the classification network that classifies chest X-ray
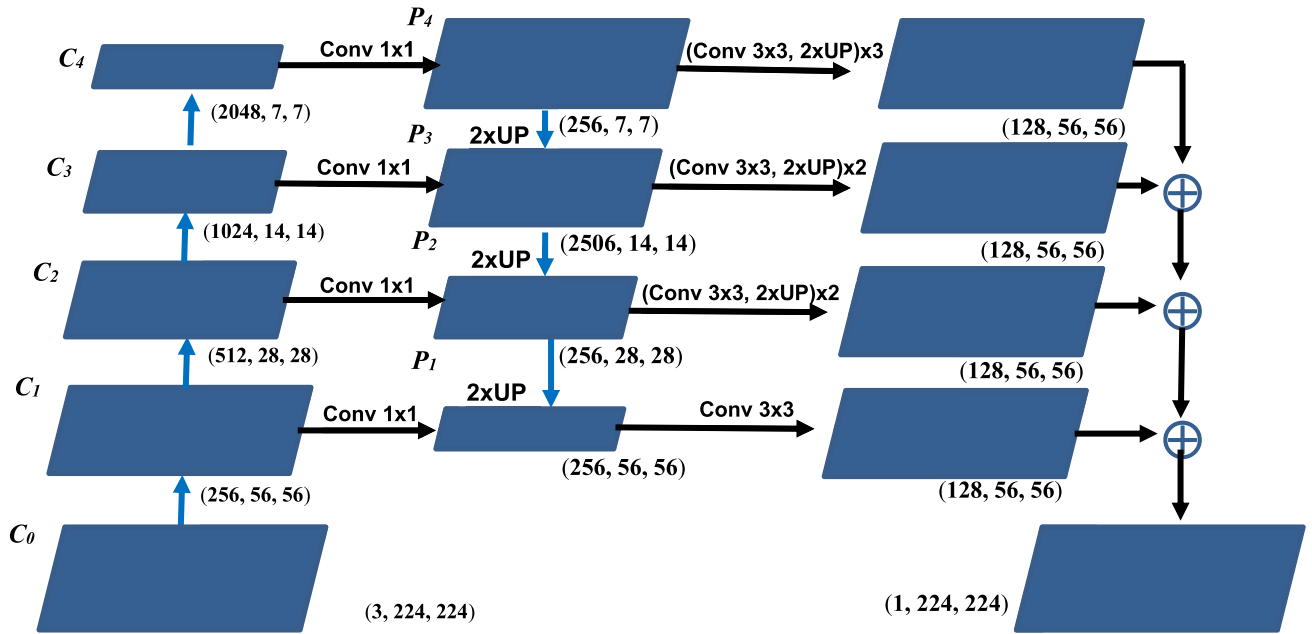
FIGURE 8. Flowchart of the localization model, Semantic-ResNet101-FPN.

images as normal or tumorous is a binary classification task, as given by Eq.(6).

$$Loss_{BCE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log\left(p\left(y_i\right)\right) + (1 - y_i) \\ \cdot log(1 - p\left(y_i\right)), \qquad (6)$$

where N is the total number of samples, i is the ith sample, y is the label of the sample, and $p(y)$ is the predicted probability. We use a threshold of 0.5 to determine the class of the classification model predicts. A larger loss value implies that there is a greater difference between the classification result and the real label, while a smaller loss value implies that there is more similarity between the classification result and the real label.

The proposed classification network, SE-SDFN, consists of three sub-models that use BCE loss during training. We slightly reduce the weight of learning on the whole chest X-ray image to emphasize the classification model's focus on the lung images. The integrated loss function is given by Eq.(7).

$$Loss_{total} = 0.3 \times L_{whole} + 0.6 \times L_{lung} + 0.1 \times L_{fusion}, \quad (7)$$

where L_{entire} is the BCE loss of SE-DenseNet-121 trained on the whole X-ray image, L_{lung} is the BCE loss of SE-DenseNet-121 trained on local lung images, and L_{fusion} is the BCE loss of the fusion layer.

D. CHEST X-RAY LUNG TUMOR LOCALIZATION NETWORK
1) THE SEMANTIC-RESNET101-FPN LOCATION MODEL
In the localization phase, we propose a Semantic-ResNet101-FPN as a chest X-ray lung tumor localization network,

as shown in Figure 8. The proposed network uses the FPN [19] and ResNet-101 [7] as the backbone network and is further designed based on the spirit of the Semantic-ResNet [20] network architecture.

The network takes a $3 \times 224 \times 224$ image as input and outputs the result $\{C_1, C_2, C_3, C_4\}$ for each residual block after feature extraction using the Resnet101. The sizes of the feature maps are 1/4, 1/8, 1/16 and 1/32 times the original input size. These feature maps are connected by the FPN, and the number of channels is reduced to 256 before the semantic information of the upper layer is restored to the size of the next layer through upscaling by a factor of 2. The new feature maps, $\{P_4, P_3, P_2, P_1\}$, containing both the strong semantic information of the upper layer and the high resolution of the lower layer, are then obtained via feature fusion. Subsequently, a 3×3 convolution up-sampling, a Group Norm, a ReLU and two times bilinear interpolation are applied to the new feature map for 3, 2, 2, 1 times, respectively, resulting in a final result of 128 channels. This result is then stitched with the features and restored to its original size using $4\times$ bilinear interpolation to obtain the final predicted result.

2) THE DICE LOSS FUNCTION
The loss function used in the training of the Semantic-ResNet101-FPN architecture is the Dice loss function. The Dice loss function was proposed by Milletari et al[21] and has been widely used in various segmentation tasks. The mathematical expression is given in Eq.(8):

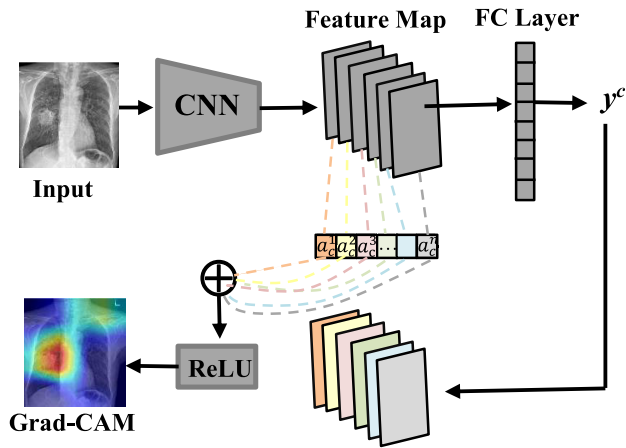$$Loss_{Dice} = 1 - \frac{2\,|X \cap Y|}{|X| + |Y|}, \qquad (8)$$

FIGURE 9. Flowchart of the Grad-CAM framework.

FIGURE 10. Lung localization samples: (left) chest X-ray images and (right) mask images of the corresponding lung areas.

TABLE 1. Images for lung localization.

	Training Images	Testing Images
Chest X-ray images	633	71
Lung area masks	633	71

where $|X|$ and $|Y|$ represent the pixel sets of the lung tumor lesion location and the labeled lung tumor area segmented by the model, respectively, and $|X \cap Y|$ represents the intersection of the lung tumor lesion area segmented by the model and the labeled lung tumor area. Since the overlapping area is counted twice when calculating the Dice loss, it needs to be multiplied by two. A small Dice loss value indicates that the model prediction is more similar to the real label, whereas a large loss value indicates a greater difference between the model prediction and the real label.

E. LUNG TUMOR VISUALIZATION

In this paper, the proposed lung tumor localization network combines the Semantic-ResNet101-FPN and Seg-Grad-CAM [14] and uses the bottleneck of its localization network to generate heatmaps for the final visualization of lung tumor on X-ray images. The Seg-Grad-CAM is a gradient-based interpretation method for semantic segmentation. It is an extension of the widely used Grad-CAM [22] and can be applied locally to generate heatmaps showing the relevance of individual pixels for semantic segmentation.

Figure 9 shows the encoder-decoder architecture of the Grad-CAM framework [22]. The framework averages the gradient of the class score over Z pixels (indexed by i, j) in each feature map and generates a weight to indicate the importance of the feature map. The algorithm is given by the following equation:

$$L_{Seg-Grad-CAM}^c = ReLU\left(\sum_k \alpha_c^k A^k\right)$$

$$with\, \alpha_c^k = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}. \qquad (9)$$

The weight, α_c^k, is linearly summed with the feature maps, A^k, using the ReLU function to zero out the negatively correlated outputs, thus highlighting the regions that contribute positively to class c.

The Seg-Grad-CAM addresses the limitations of the Grad-CAM in image segmentation tasks by replacing $\mathbf{y^c}$ with $\sum_{(i,j)\in M} y_{ij}^c$, where M denotes the set of pixels of the predicted class, and i and j denote the pixel coordinates. As a result, the use of the Grad-CAM is more flexible in the semantic segmentation task. Furthermore, the approach uses the convolutional layers at the bottleneck of the decoder to extract the feature maps.

IV. EXPERIMENTS AND RESULTS

A. DATASET FOR LUNG LOCALIZATION

The image datasets used in this paper for lung localization were collected from the Department of Health and Human Services of Montgomery County (MC), Maryland [23], and the Third People's Hospital of Shenzhen, Guangdong providence, China [24]. There were 704 chest X-ray images, and each image has its corresponding lung area mask, as shown in Figure10. We randomly divide the dataset images into training and testing datasets with a ratio of 9:1, as shown in Table 1.

B. CHEST X-RAY DATA SET FOR CLASSIFICATION

The chest X-ray images used for classification were provided by the Da Lin Tzu Chi hospital, Taiwan. There are 2,004

TABLE 2. The chest X- ray images for lung tumor classification.

	Training Images	Validation dataset	Testing dataset
Tumor	699	202	101
Normal	703	198	101

TABLE 3. Images used with lung tumor markers.

	Training Images	Testing Images
Tumor Samples	548	179
Tumor Masks	548	179

(a)

(b)
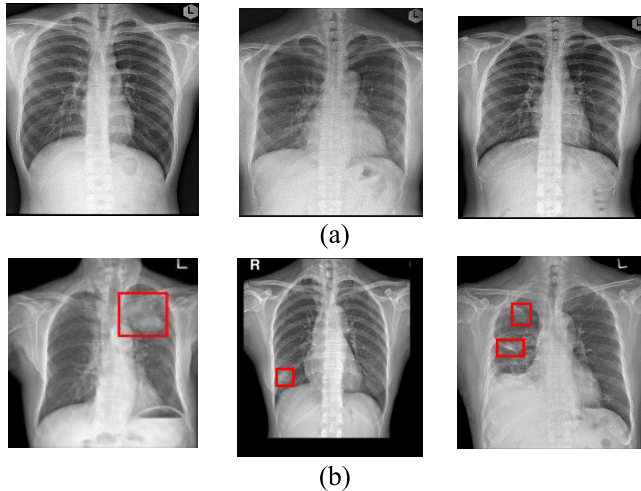
FIGURE 11. (a) Normal X-ray images and (b) X-ray images with lung tumors are highlighted in red.
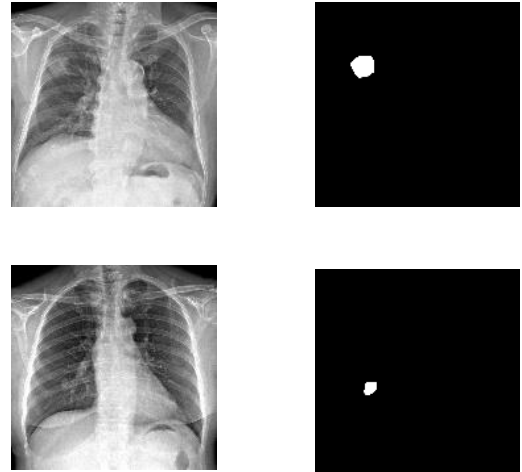
FIGURE 12. Lung tumor images with marked lesion areas.

images in the dataset, including normal chest X-ray images and images with pulmonary tumors. The samples labeled as normal were confirmed by Tzu Chi Hospital's professional physicians through computed tomography (CT) to ensure that the images were indeed free of lung tumors. As shown in Table 2, the images in the dataset were randomly selected for training, validation and testing in a ratio of 7:1:2. Figure 11 shows six image samples from the dataset.

In order to make the experiment process more comprehensive and reduce the bias caused by data selection. We also performed 3-fold cross-validation, and randomly divided the total 2004 labeled images into three groups of 1069 images, 267 images and 668 images, which were used as training, validation and test data respectively.

C. LUNG TUMOR DATASET

In this paper, imaging physicians from Tzu Chi hospital were asked to mark lesion areas of 727 cropped X-ray images containing lung tumors. During training, these images were randomly divided into training and testing datasets in an 8:2 ratio. The numbers of training and testing images in this dataset are given in Table 3. Figure 12 shows two image samples from this dataset and the corresponding lesion area markers.

D. PARAMETER SETTINGS FOR MODEL TRAINING

The input image size is 224 × 224 for training the lung localization network, U-Net. We set the number of iterations to 50, the batch size to 16, and use Adam as the optimizer

with a learning rate of 0.0005. When training the chest X-ray binary classification network, SE-SDFN, we set the number of iterations to 50, the batch size to 16, and use Adam as the optimizer. Since there are three sub-models in the classification network, we adjust the learning rate of the two feature extractors, SE DenseNet-121, to 0.0001 and the fusion layer to 0.001. When training the lung tumor localization model, Semantic-ResNet101-FPN, the input image size is 224×224. Additionally, we set the number of iterations to 100, the batch size to 16, and use Adam as the optimizer with a learning rate of 0.0001.

E. EFFECTIVENESS ASSESSMENT

In this paper, the lung localization model (U-Net) and the lung tumor localization model (Semantic-ResNet101-FPN) are evaluated using Dice and IOU metrics during testing to evaluate the quality of the localization results. Dice and IOU are defined as follows:

$$Dice = \frac{2\,|X \cap Y|}{|X| + |Y|}, \tag{10}$$

$$IOU = \frac{|X \cap Y|}{|X \cup Y|}, \tag{11}$$

where |X| denotes the set of pixels labeled by lung tumors or lung regions, |Y| denotes the set of pixels of lung tumors or lung regions predicted by the model, |X∪Y| denotes the union of labeled region pixels and model-predicted pixels, and |X∩Y| denotes the set of pixels where labeled regions overlap with the region predicted by the model.

In our work, Negative Prediction, Specificity, Precision, Sensitivity, F1 Score, and Accuracy are used to evaluate the

Original	Prediction	Post-processing	Result
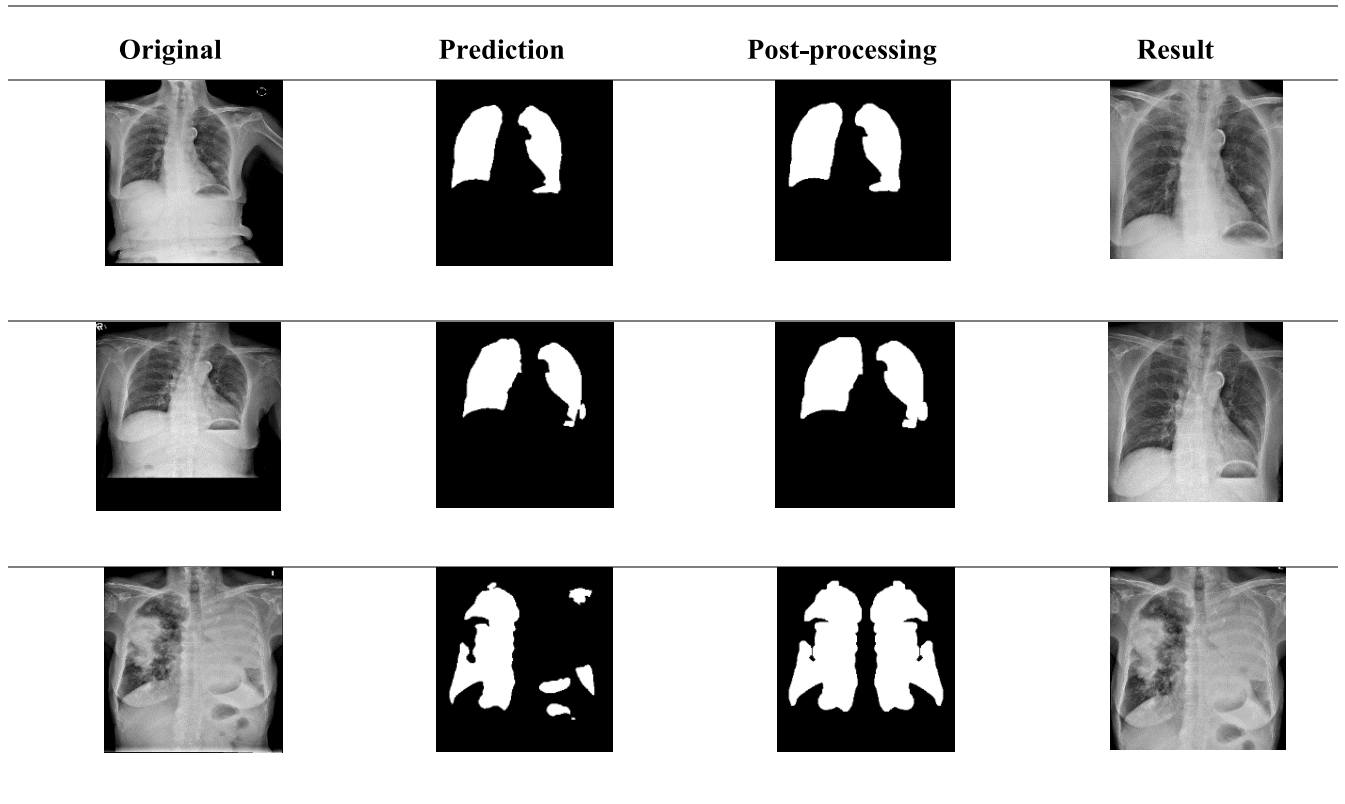
FIGURE 13. From left to right, the original X-ray images are predicted by the U-Net model, post-processed, and then cropped according to the chest position of the original image.

TABLE 4. Comparative results of tumor classification.

	Wang *et al*. [6]	Rajpurka *et al*. [8]	Guan *et al*. [10]	Liu *et al*. [12]	Ours
Precision	97.03%	96.04%	86.14%	**99.0%**	98.04%
Negative Prediction	95.05%	94.06%	72.28%	98.04%	**99.0%**
Sensitivity	95.15%	94.17%	75.65%	98.02%	**99.01%**
Specificity	96.97%	95.96%	83.91%	**99.01%**	98.02%
F1 Score	96.08%	95.10%	80.56%	98.51%	**98.52%**
Accuracy	96.04%	95.05%	79.21%	**98.51%**	**98.51%**

TABLE 5. Comparison of computation time for tumor classification.

	Wang *et al*. [6]	Rajpurka *et al*. [8]	Guan *et al*. [10]	Liu *et al*. [12]	Ours
Training time	3672.33 Sec.	3426.3 Sec.	4751.32 Sec.	4886.82 Sec.	4992.46 Sec.
Testing time	61.50 Sec.	43.88 Sec.	55.59 Sec.	97.40 Sec.	97.88 Sec.

TABLE 6. Comparative results of the confusion matrices and ROC curves.

	Confusion Matrix	ROC Curve
Wang *et al.* **[2]**		
Rajpurka *et al.* **[4]**		
Guan *et al.* **[6]**		
Liu *et al.* **[8]**		
Ours		

TABLE 7. Comparison of tumor classification results by 3-fold cross-validation.

	Wang *et al.* [6]	Rajpurka *et al.* [8]	Guan *et al.* [10]	Liu *et al.* [12]	Ours
Precision	92.11%	93.61%	82.96%	**96.35%**	95.10%
Negative Prediction	94.11%	94.51%	83.86%	95.00%	**96.42%**
Sensitivity	93.39%	92.97%	83.75%	94.93%	**96.38%**
Specificity	92.31%	93.70%	83.25%	**96.41%**	95.09%
F1 Score	93.02%	94.03%	83.27%	95.62%	**95.73%**
Accuracy	93.11%	94.06%	83.38%	95.66%	**95.74%**

TABLE 8. Comparison of different models for lung tumor localization.

	VGG16-CXR-U-NET[25]	VGG19-CXR-U-NET[25]	Semantic-ResNet101-FPN
IOU	66.35%	55.37%	**67.11%**
Dice	77.06%	68.37%	**78.39%**

TABLE 9. Comparison of computation time for Tumor Localization.

	VGG16-CXR-U-NET[25]	VGG19-CXR-U-NET[25]	Semantic-ResNet101-FPN
Training Time	15575.35 Sec.	15348.17 Sec.	16053.38 Sec.
Testing Time	17.30 Sec.	17.84 Sec.	17.42 Sec.

overall performance of the model in classifying chest X-ray images. The equations for these metrics are defined below:

$$Precision = \frac{TP}{TP + FP},\qquad(12)$$

$$Negative\ Prediction = \frac{TN}{TN + FN},\qquad(13)$$

$$Sensitivity = \frac{TP}{TP + FN},\qquad(14)$$

$$Specificity = \frac{TN}{TN + FP},\qquad(15)$$

$$F1\ Score = \frac{2 \times Precision \times Sensitivity}{Precision + Sensitivity},\qquad(16)$$

$$Accuracy = \frac{TP + FN}{TP + TN + FP + FN},\qquad(17)$$

where TP represents the number of cases classified as lung tumors, TN represents the number of cases classified as normal, FP represents the number of normal cases misclassified as lung tumors, and FN represents the number of lung tumor cases misclassified as normal.

F. LUNG LOCALIZATION RESULTS

In this paper, we localize the lung region in the chest X-ray image with the weight of the lowest loss value obtained when training the U-Net model. We use any chest X-ray image to predict the lung area through the U-Net model. The presence of air or other pathological influences in the image may lead to false-positive areas. For this reason, the predicted areas need to be optimized by post-processing procedures and

cropped according to the original image position to obtain the final result. In our work, the Dice and IOU metrics achieve 92.8% and 96.2% accuracy, respectively, when we use the test dataset for validation. Figure 13 shows three original images and the results obtained after the U-Net model prediction, post-processing optimization, and the cropped results according to the original image positions.
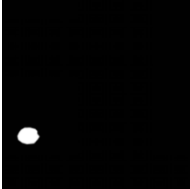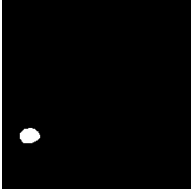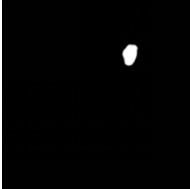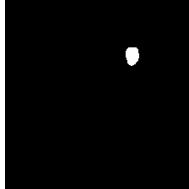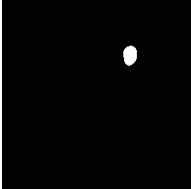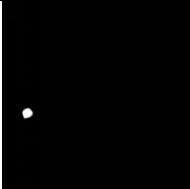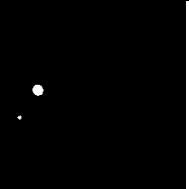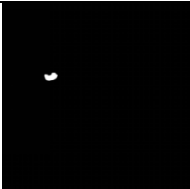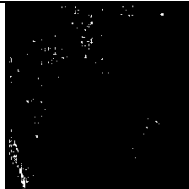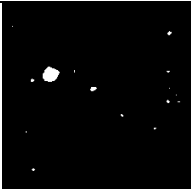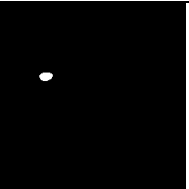
G. COMPARISON RESULTS OF TUMOR CLASSIFICATION

We have conducted an experiment to classify whether or not given chest X-ray images contain lung tumors and compare our method with four other methods proposed by Wang et al. [6], Rajpurka et al. [8], Guan et al. [10] and Liu et al. [12]. The image dataset used for comparison was provided by the Department of Imaging Medicine, Da Lin Tzu Chi Hospital. The comparison results of various effectiveness metrics and computation time are shown in Table 4 and Table 5, respectively.

From these results, we note that the methods proposed by Wang et al. [6] and Rajpurkar et al. [8] use only whole chest X-ray images for training and achieved sensitivities of around 94-95%. Guan et al. [10] crop a discriminative region from the whole chest X-ray image based on an attention mechanism. However, inaccurate CAM cropping may lead to misclassification, resulting in poor recognition rates.

The SDFN method proposed by Liu et al. [12] allows the model to be trained using both the whole and local lung-region X-ray images, thereby improving the model's attention. Our method extends this idea by adding an SE block attention mechanism to the SDFN's feature extractor

TABLE 10. Comparison of lung tumor localization results using different models.

Original	GT	VGG16-CXR-U-NET	VGG19-CXR-U-NET	Semantic-ResNet101-FPN

to improve the model's attention to specific features, enabling our method to achieve better classification results. In Table 5, although the computation time required by our method is a little longer than other methods, better classification results can be obtained. The confusion matrix and ROC curves of the different methods are presented in Table 6, and the results show that our proposed method has better performance.

In order to verify whether the proposed model produces consistent test results for different training data. Table 7 shows a comparison of tumor classification by 3-fold cross-validation with the same dataset. The experimental results show the same trend as Table 4, which indicates that our proposed method has better performance than the state-of-the-art methods.

H. COMPARISON OF LUNG TUMOR LOCALIZATION MODELS

To verify the effect of the proposed chest X-ray lung tumor localization model, Semantic-ResNet101-FPN, we compare the VGG16-CXR-U-Net[25], VGG16/19-CXR-U-Net [25], and the proposed Semantic-ResNet101-FPN models, as shown in Table 8. The proposed Semantic-ResNet101-FPN model using the residual network combined with the feature pyramid achieves higher Dice and IOU metrics than

TABLE 11. Comparison of different lung tumor localization models by 3-fold cross-validation.

	VGG16-CXR-U-NET[25]	VGG19-CXR-U-NET[25]	Semantic-ResNet101-FPN
IOU	51.88%	40.54%	**65.53%**
Dice	59.27%	52.23%	**72.80%**
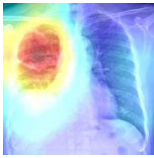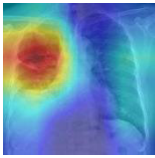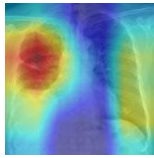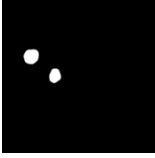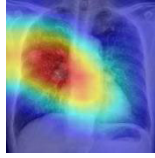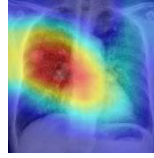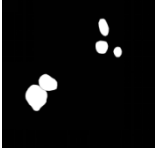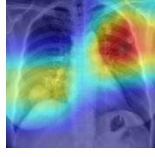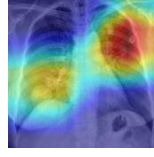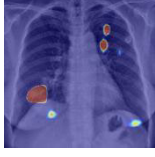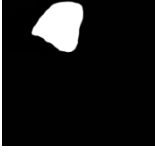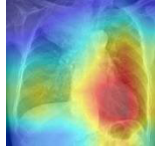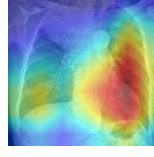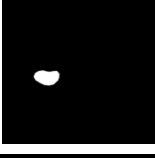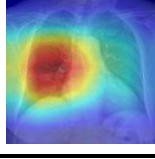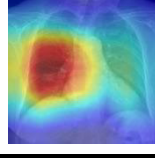
the VGG16/19-CXR-U-NET model. Since the VGG network cannot connect the feature maps between layers as tightly as the residual network when the number of layers is deepened, it eventually leads to excessive loss of target features.

The computational time required for tumor localization by different methods is shown in Table 9. Although it took a little longer to train the model compared to other methods, the time spent on testing was about the same. But our method can achieve better localization results.

Table 10 shows four lung tumor localization results using different models. It can be seen from the experimental results that the proposed Semantic-ResNet101-FPN model is more advantageous in locating small objects, while the VGG16/19-CXR-U-NET model often loses information for small objects and makes false-positive predictions.

In order to verify the generalization of the proposed method, Table 11 shows tumor localization comparisons with

TABLE 12. Comparison of lung tumor visualization by different methods.

the same dataset by 3-fold cross-validation. Each fold randomly selects 484 images from the same dataset for training and 243 images for testing. The same trend can be seen when comparing the experimental results with Table 7. Although IOU and Dice drop slightly, our proposed method significantly outperforms using VGG16-CXR-U-NET versus VGG19-CXR-U-NET.

I. COMPARISON OF LUNG TUMOR VISUALIZATION METHODS

To further verify the effect of lung tumor visualization from X-ray images, we compare the proposed method with three weakly supervised learning models based on the DenseNet-121, which generate heatmaps that correspond to lung tumor lesion regions, namely the CAM [11], Grad-CAM [22] and Ablation-CAM [26].

Our method uses the Semantic-ResNet101-FPN, a lung tumor segmentation network trained with domain-level pixel labels to generate the heatmap corresponding to lung tumor lesion regions. The result is combined with the Seg-Grad-CAM to visualize the heatmap of lung tumor lesion location

obtained at the bottleneck of the Semantic-ResNet101-FPN. Experimental results show that the proposed method has achieved accuracies of 67.11% and 78.39 on the IOU and Dice indices, respectively.

Table 12 shows that the heatmaps of the lesion region generated by the Grad-CAM and Ablation-CAM are roughly the same. In comparison, our method achieves better and more accurate localization and visualization than the other methods.

V. CONCLUSION

This paper integrates clinical, medical image processing and artificial intelligence deep learning technology to develop a lung tumor detection system that can assist doctors or radiologists in clinically interpreting chest X-ray images. Experimental results show that the proposed method has achieved a classification accuracy of 98.51% and a sensitivity of 99.01%, respectively. Unlike conventional methods, which identify lesion areas using a heatmap generated through feature visualization with a classification network, the proposed method uses a semantic segmentation CNN network to localize lung

tumors and an Seg-Grad-CAM to visualize the model predictions. Our chest X-ray lung tumor detection system has assisted doctors in clinical diagnosis at Dalin Tzu Chi Hospital in Taiwan. The proposed computer-aided medical system effectively assists doctors to make diagnosis and treatment decisions efficiently.

REFERENCES

[1] WHO Newsroom. (2022). *Cancer.* [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] I. C. Mary and J. Preethi, "A survey on computerized, quantification and classification of lung disease," *Int. J. Adv. Inf. Commun. Technol.*, vol. 4, no. 3, pp. 631–634, 2017.

[3] I. Abed, "Lung cancer detection from X-ray images by combined backpropagation neural network and PCA," *Eng. Technol. J.*, vol. 37, no. 5A, pp. 166–171, May 2019.

[4] M. Yahyatabar, P. Jouvet, and F. Cheriet, "Dense-Unet: A light model for lung fields segmentation in chest X-ray images," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1242–1245.

[5] W. Ausawalaithong, A. Thirach, S. Marukatat, and T. Wilaiprasitporn, "Automatic lung cancer prediction from chest X-ray images using the deep learning approach," in *Proc. 11th Biomed. Eng. Int. Conf. (BMEiCON)*, Nov. 2018, pp. 1–5.

[6] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[8] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," 2017, *arXiv:1711.05225*.

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2016, *arXiv:1608.06993*.

[10] Q. Guan, Y. Huang, Z. Zhong, Z. Zheng, L. Zheng, and Y. Yang, "Diagnose like a radiologist: Attention guided convolutional neural network for thorax disease classification," 2018, *arXiv:1801.09927*.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," 2015, *arXiv:1512.04150*.

[12] H. Liu, L. Wang, Y. Nan, F. Jin, Q. Wang, and J. Pu, "SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images," 2018, *arXiv:1810.12959*.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.

[14] K. Vinogradova, A. Dibrov, and G. Myers, "Towards interpretable semantic segmentation via gradient-weighted class activation mapping," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13943–13944.

[15] J. Li, H. Chen, Y. Li, Y. Peng, J. Sun, and P. Pan, "Cross-modality synthesis aiding lung tumor segmentation on multi-modal MRI images," *Biomed. Signal Process. Control*, vol. 76, Jul. 2022, Art. no. 103655.

[16] J. Jiang, Y.-C. Hu, N. Tyagi, A. Rimner, N. Lee, J. O. Deasy, S. Berry, and H. Veeraraghavan, "PSIGAN: Joint probabilistic segmentation and image distribution matching for unpaired cross-modality adaptation-based MRI segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 12, pp. 4071–4084, Dec. 2020.

[17] J. Jiang, A. Rimner, J. O. Deasy, and H. Veeraraghavan, "Unpaired cross-modality educed distillation (CMEDL) for medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 5, pp. 1057–1068, May 2022.

[18] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017, *arXiv:1709.01507*.

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 2016, *arXiv:1612.03144*.

[20] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," 2019, *arXiv:1901.02446*.

[21] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[22] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," 2016, *arXiv:1610.02391*.

[23] S. Jaeger, A. Karargyris, S. Candemir, L. Folio, J. Siegelman, F. Callaghan, Z. Xue, K. Palaniappan, R. K. Singh, S. Antani, G. Thoma, Y.-X. Wang, P.-X. Lu, and C. J. McDonald, "Automatic tuberculosis screening using chest radiographs," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 233–245, Feb. 2014.

[24] S. Candemir, S. Jaeger, K. Palaniappan, J. P. Musco, R. K. Singh, Z. Xue, A. Karargyris, S. Antani, G. Thoma, and C. J. McDonald, "Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration," *IEEE Trans. Med. Imag.*, vol. 33, no. 2, pp. 577–590, Feb. 2014.

[25] S. Rajaraman, L. R. Folio, J. Dimperio, P. O. Alderson, and S. K. Antani, "Improved semantic segmentation of tuberculosis—Consistent findings in chest X-rays using augmented training of modality-specific U-Net models with weak localizations," *Diagnostics*, vol. 11, no. 4, p. 616, Mar. 2021.

[26] S. Desai and H. G. Ramaswamy, "Ablation-CAM: Visual explanations for deep convolutional network via gradient-free localization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 972–980.

ADE IRMA SURYANI received the Diploma IV degree in information technology from the Electronics Engineering Polytechnic Institute of Surabaya (EEPIS), Surabaya, Indonesia, in 2008, and the Master of Information Technology degree (M.T.I.) from the Faculty of Computer Science, University of Indonesia (UI), Jakarta, Indonesia, in 2013. She is currently pursuing the Ph.D. degree with the Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology (YunTech), Yunlin, Taiwan. She has been a full-time Lecturer with the Information Technology Department, Padang State Polytechnic (PNP), Padang, Indonesia, since 2014. Her research interests include computational intelligence and its applications to medical image processing, artificial intelligence, implementation of information systems, and the Internet of Things (IoT) analysis. In 2011, she awarded a Full Scholarship from The Ministry of Research, Technology and Higher Education Republic of Indonesia for her master's program from the University of Indonesia.

CHUAN-WANG CHANG received the M.S. degree in electrical engineering from the National Sun Yat-sen University, Taiwan, in 1995, and the Ph.D. degree in electrical engineering from the National Cheng Kung University, Taiwan, in 2010. He is currently an Assistant Professor at the Department of Computer Science and Information Engineering, National Chin-Yi University of Technology, Taichung, Taiwan. His research interests include application of artificial intelligence, automated optical inspection, multimedia database, and application of the Internet of Things.

YU-FAN FENG received the M.S. degree in computer science and information engineering from the National Yunlin University of Science and Technology, Yunlin, Taiwan, in 2021. His research interests include machine learning and their applications to medical image processing.

TIN-KWANG LIN graduated from the School of Medicine, Kaohsiung Medical University, in 1992. He has been a Cardiologist with the Dalin Tzu Chi Hospital, Buddhist Tzu Chi Medical Foundation, since 2000. Since 2020, he has been an Assistant Professor with the Department of Internal Medicine, School of Medicine, Tzu Chi University, Hualien, Taiwan. He is currently the Director of the Department of Cardiology and the Secretary of Medical Affairs of the Superintendent's Office in the hospital. His specialties are in echocardiography, epithelium function, and cardiac catheterization. He has authored or coauthored more than 30 publications in journals and conference proceedings in his research fields.

CHIH-WEN LIN received the Medical degree from the Medical School, Chung Shan Medical University, Taichung, Taiwan, in 1991, and the Master of Business Administration degree from the Continuing Education Program, Institute of Information Management, National Chung Cheng University, Chiayi, Taiwan, in 2015. Since 2002, he has been the Chief of the Department of Medical Imaging, Dalin Tzu Chi Hospital, Taiwan. His major medical specialties are diagnostic radiology and interventional radiology, especially in hepatoma treatment, lung cancer screening, and lung biopsy. He is currently a Lecturer with the Department of School of Medicine, Tzu Chi University, Hualien, Taiwan. His major research interests include medical image processing and lung cancer diagnosis.

JEN-CHIEH CHENG received the M.S. degree in information and electronic engineering from the National Central University, Taiwan, in 1991, and the E.M.B.A. degree from the National Tsing Hua University, Taiwan, in 2009. He was the Deputy General Director of the Computational Intelligence Technology Center, Industrial Technology Research Institute (ITRI), Taiwan, from 2013 to 2017, and as the Deputy General Director of Information and Communications Research Laboratories, ITRI, from 2017 to 2019. He is currently the Vice President of the ITRI, and the General Director of the Service Systems Technology Center, ITRI. He works as an Executive Board Director of the Information Service Industry Association of Taiwan, as a Consultant of the General Chamber of Commerce of Taiwan, and as a Principal Investigator of the Application and Industrial Development of Sports Technology Program. He holds nine patents and has published 23 international academic articles. He won more than 40 awards in his career includes Top 10 Outstanding I.T. Elite Award, Taiwan, in 2001; and Top 10 Distinguished Engineer Award, Taiwan, in 2008. Under his leadership, his team won research and development 100 Award in 2020, Edison Awards in 2021, National Industrial Innovation Award, MOEA in 2021, and CES Innovation Awards in 2020, 2021, and 2022. These awards across fields of intelligence medical, smart wearable device, blockchain, and AI solutions.

CHUAN-YU CHANG (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the National Cheng Kung University, Tainan, Taiwan, in 2000. From 2009 to 2011, he was the Chair of the Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology (YunTech), Taiwan. From 2011 to 2019, he was the Dean of the Research and Development, and the Director of the Incubation Center for Academia-Industry Collaboration and Intellectual Property, YunTech. He is currently the Deputy General Director of the Service Systems Technology Center, Industrial Technology Research Institute, Taiwan. He is also a Distinguished Professor with the Department of Computer Science and Information Engineering, YunTech. He has authored or coauthored more than 200 publications in journals and conference proceedings in his research fields, which include computational intelligence and their applications to medical image processing, automated optical inspection, emotion recognition, and pattern recognition. He is an IET Fellow, and a Life Member of IPPR and TAAI. He received the National Award for Distinguished Contribution to Industry-Academia Cooperation from Ministry of Education, and the Outstanding Electrical Engineering Professor Award from the Chinese Institute of Electrical Engineering, Taiwan, in 2021. He was the Program Co-Chair of TAAI 2007, CVGIP 2009, 2010–2019 International Workshop on Intelligent Sensors and Smart Environments, and the third International Conference on Robot, Vision and Signal Processing (RVSP 2015). He was the General Co-Chair of 2012 International Conference on Information Security and Intelligent Control, 2011–2013 Workshop on Digital Life Technologies, CVGIP 2017, WIC 2018, ICS 2018, and WIC 2019. From 2015 to 2017, he was the Chair of the IEEE Signal Processing Society Tainan Chapter and the Representative for Region 10 of IEEE SPS Chapters Committee. He is also the President of the Chinese Image Processing and Pattern Recognition Society.

• • •