Accurate segmentation of lungs to assist physicians in computer-aided diagnosis

Chen Yen-Yu[1*], Chen Ching-Cheng [2]

ABSTRACT

The global spread of COVID-19 and influenza over the past few years has made it necessary for first-line clinicians to mark out definite areas of the lungs when making diagnoses on radiographs. In this study, we propose a new method to determine the exact location of the lungs in radiographic images, preserving only the mask of this region to generate the ROI region needed by physicians to assist in diagnosis. Our algorithm consists of three stages, i.e., Depthwise Separable Convolution, Attention Enhancing Block, and Asymmetric autoencoder. Depthwise Separable Convolution can capture the X-ray image with limited computational resources. The Attention Enhancing Block is used to extract the X-ray image features from the three different receptive fields and the fused features are then reduced by the decoder. The Asymmetric autoencoder model is more focused on learning and preserving the precise details of the ROI region masks. We tested our method using lung radiographs collected from the Kaohsiung Medical University (KMU) database, and the simulation results showed that, on the one hand, our proposed method has a better Dice coefficient compared to other segmentation methods; on the other hand, it is able to locate the precise image segmentation of the lung ROI region needed for clinicians' diagnosis. The proposed method can accurately localize the precise image segmentation of the lung ROI region needed for clinician diagnosis.

Keywords: lung, radiograph, ROI region, segmentation, computer-aided diagnosis

I. INTRODUCTION

In recent years, due to global pandemics of infectious diseases such as COVID-19 [1][2] and influenza [3], clinicians have been faced with additional challenges when dealing with the radiographic diagnosis of their patients. Imaging has become an indispensable tool in the management of these diseases, and radiographs are often one of the most common modalities used.

X-rays provide detailed images of the structure of the lungs, which are essential for diagnosing symptoms of respiratory diseases and infections. However, because of the severity and spread of these diseases, doctors need to identify problem areas in a patient's lungs more quickly and precisely.

In this context, machine learning and artificial intelligence technologies have entered the realm of clinical diagnosis [4]. With these advanced technologies, doctors can more easily mark and localize defined areas of the lungs. Machine learning algorithms are able to recognize specific structures and lesions in the image, helping doctors to mark potential areas of disease on the image.

This intelligent labeling system helps improve the speed and accuracy of diagnosis, especially during a pandemic when healthcare resources are under pressure. With this technology, clinicians can more effectively identify possible lesions and quickly implement appropriate treatment measures to ensure timely and effective patient care.

Overall, the application of machine learning and artificial intelligence to clinical X-ray diagnosis provides healthcare professionals with a powerful tool to help meet the challenges posed by the global infectious disease epidemic.

Therefore, this study proposes a new method that incorporates deep learning to label the precise location of the lungs in radiograph images, to accurately segment the lungs and label the ROI regions before further assisting clinicians in computer-aided diagnosis.

II. BACKGROUNDAND MOTIVATION

Recently, deep convolutional neural network (DCNN) based approaches have demonstrated significant potential in various medical diagnostic domains, prompting further exploration in applied research [5]. The utilization of DCNN holds promise for diminishing the reliance on costly computed tomography (CT) and magnetic resonance imaging (MRI) scans. Its automated and precise outcomes alleviate the urgency for clinicians to promptly identify symptoms, thereby reducing their workload.

*Corresponding Author: (E-mail: aicyy@ncut.edu.tw).
[1] Department of Artificial Intelligence and Computer Engineering,, National Chin-Yi University of Technology, Taichung, Taiwan
[2] Department of information Management, National Chung Hsing University, Taichung, Taiwan

Nonetheless, the applicability and effectiveness of DCNN in the detection of lung regions pose challenges.

To the best of our understanding, conventional DCNN-based techniques, including U-Net [6] and DenseNet [7], are employed for segmenting lung regions. However, the identification of lung regions using DCNN encounters difficulties due to subtle variations in the grayscale distribution of these regions in radiographic images. These features may fade away following a series of convolution operations as the network delves deeper into the layers.

III. PROPOSED METHOD

The goal of this study is to be able to segment the X-ray images of the lungs and separate the unnecessary organs. A modified Auto-Encoder is used in the segmentation stage to process the input of lung images [8]. The goal of this stage is to generate a mask that preserves only the lung region from the original image.

Figure 1 shows the deep learning architecture proposed in this study, which mainly consists of encoder, middle layer and decoder. These elements are the key components of the Convolutional Autoencoder model and they work together synergistically to realize the image segmentation task.

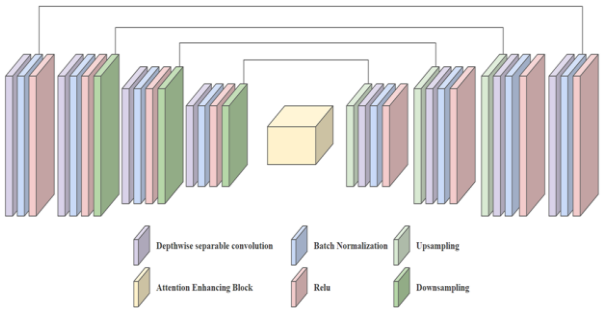


Figure 1. Structure of Autoencoder

Autoencoder is composed of an encoder and a decoder, and in the process of encoding and decoding, the key to connecting encoding and decoding is to obtain the potential representation of the data. The goal of training techniques for self-supervised learning of autoencoders is to make the input equal to the output, but in practice the input will not be exactly equal to the output, instead the input will be approximated by the output, in which case it is necessary to define a reconstruction loss to characterize the difference between the input and the output. The encoder transforms the input image into a compressed representation, the compressed features are intensively computed in the middle layer, and the decoder restores the details to the original image. The encoder in this study differs from the original autoencoder in that it uses a deeply separable convolution technique [9-11] to reduce the computational cost. The Encoder compresses the image into a low-dimensional representation and then restores it. The use of depth-separable convolution is suitable for capturing lung details in radiograph images with limited computing resources, in addition, we use an attention mechanism in the middle layer to allow the model to focus on the lung edge details to increase the accuracy of the segmentation.

Depthwise Separable Convolution

Depthwise Separable Convolution is an efficient convolution operation that is widely used in deep learning to reduce the computational burden and the number of parameters of the model while maintaining a considerable performance. This convolution operation consists of two main steps: Depthwise Convolution and Pointwise Convolution. Before exploring Depthwise Convolution, it is important to explain how it differs from traditional convolution. Conventional convolution operations usually apply multiple convolution kernels on the input feature map, where each convolution kernel spans all input channels, as shown in Figure 2. For example, if we have a feature map with dimensions H × W × D (where H is the height, W is the width, and D is the number of channels), and we use one K × K × D convolution kernel to perform the convolution operation, each convolution kernel will cover all the K × K spatial locations on the D channels. If we use M such convolution kernels, the final output will be a new feature map of H' × W' × M, where H' and W' are the spatial dimensions of the new feature map.

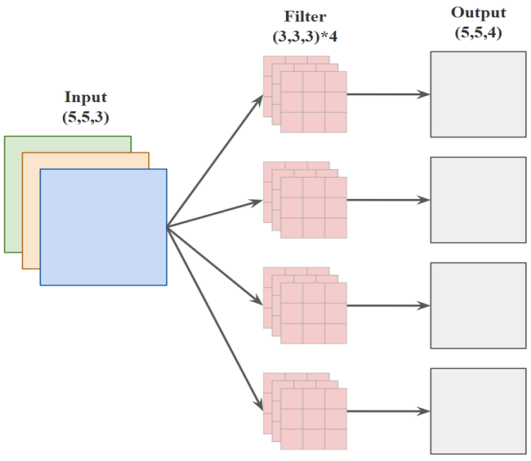


Figure 2. Example of Conventional Convolution Operation

Depthwise Convolution is divided into two steps.

1. Depthwise Convolution

First, a K×K×1 convolution kernel is applied to each input channel independently. For each channel, the convolution is performed only within the channel, and the information is not fused across the channels. This step greatly reduces the computational effort because it involves the convolution of each channel independently instead of using one large convolution kernel to cover all channels.

2. Pointwise Convolution

Next, the output of the depth convolution is convolved using a 1×1×D convolution kernel, a process also known

as pointwise convolution. The purpose of pointwise convolution is to combine and recalibrate the output channels of the depth convolution. In this step, each 1×1 convolution kernel spans all the D channels but operates at only one point in space, thus combining the information of each channel obtained from the depth convolution.

The efficiency and lightweight nature of Depthwise Separable Convolution represent its notable advantages. In terms of computational efficiency, it significantly reduces the required number of multiplication operations. This is illustrated in Figure 3 through a straightforward example. Consider a conventional 3 × 3 × 3 convolution kernel, which involves 27 multiplication operations on a single channel. In contrast, the depth convolution reduces this to 9 operations. Assuming there are 3 channels, the total number of depth convolution operations remains 27. With 3 such convolution kernels, the total number of pointwise convolution operations becomes 9, resulting in a total of 36 convolution operations. In comparison, a conventional convolution operation would entail 27 times the number of channels. Therefore, when the number of channels exceeds 1, Depthwise Separable Convolution achieves significant computational savings.

The reduction in the number of operations is attributed to the separate handling of spatial and depth-wise convolutions in Depthwise Separable Convolution, leading to a more efficient computation process. This not only streamlines the computational load but also enhances the overall efficiency of the model, particularly in scenarios involving multiple channels.

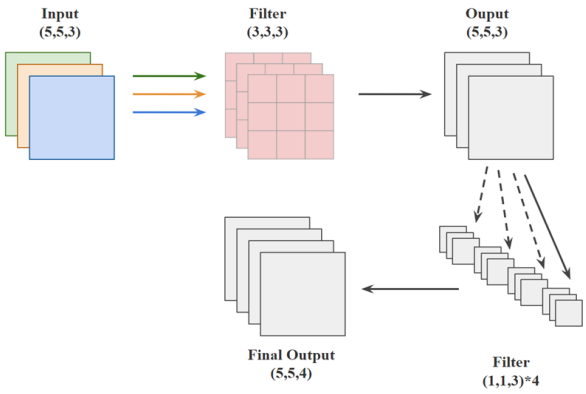


Figure 3. Example of Depth Separable Convolution

Attention Enhancing Block

In this study, the depth separable convolution is used because of the specially designed attention-enhancing block in the middle layer as shown in Figure 4. In order to avoid the high computational cost caused by the Attention Enhancement Block, the depth separable convolution is used in the encoder and decoder to balance the computational cost.

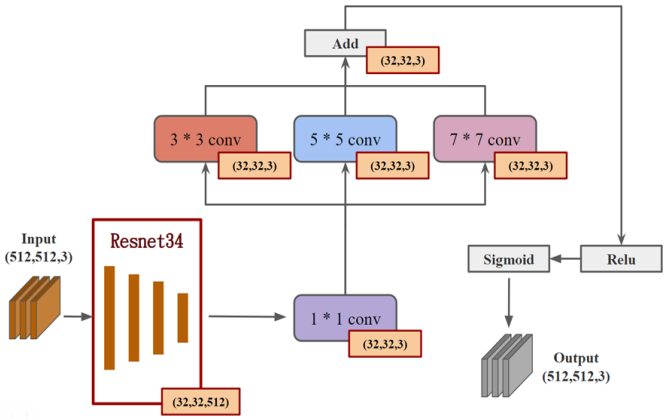


Figure 4. Structure of Attention Enhancement Block

The attention enhancement block in our proposed model primarily employs Resnet 34 [12-13] for multi-scale feature extraction in the intermediate layers. Resnet 34, being a lightweight pre-trained residual network, proves efficient in extracting features from intermediate layers. This efficiency is attributed to the residual structure, which mitigates the issue of gradient vanishing. By overcoming this challenge, the model can be deepened, enabling the extraction of feature layers with an increased number of layers. Utilizing Resnet 34 enhances the model's capability to efficiently capture features at different scales, contributing to the overall effectiveness of the attention enhancement block. The pre-trained nature of Resnet 34 further facilitates the extraction of meaningful features from intermediate layers, as it has been optimized for feature representation in various tasks. This allows our proposed model to leverage the advantages of a well-established architecture, resulting in improved feature extraction and representation. In summary, the attention enhancement block benefits from the implementation of Resnet 34, providing a robust mechanism for multi-scale feature extraction in the middle layers. This choice ensures the model's ability to overcome gradient vanishing issues, allowing for increased depth in the layers of extracted features and, consequently, more effective feature extraction.

Utilizing Resnet 34 for the extraction of features from three distinct receptive fields, we conduct a multi-scale combination at the conclusion of the process [14-15]. It is noteworthy that, in this context, multi-scale does not imply variations in size; rather, it pertains to the uniformity in size and dimensions of the extracted feature maps. These feature maps can be directly summed across different sensory fields. The amalgamated features are then directed towards the subsequent decoder reduction.

The primary rationale behind opting for Resnet34 for feature extraction preceding multiscale fusion lies in its ability to delineate an approximate contour of the lungs through the advanced features' residual structure. This characteristic contributes to an enhanced segmentation performance of the model. Additionally, Resnet34 is selected for its lightweight nature among residual models, which proves beneficial in scenarios where computational

costs need to be constrained. This study acknowledges the significance of multiscale fusion in achieving a holistic understanding of lung features, and Resnet34's efficiency becomes pivotal in the process. By employing a model that can effectively discern the contour of lungs and provide rich feature representations, we aim to enhance the overall performance of the segmentation model. The choice of Resnet34 aligns with a balanced consideration of segmentation accuracy and computational efficiency within the given constraints.

Asymmetric autoencoder

Converting a traditional symmetric autoencoder to an asymmetric autoencoder brings significant advantages for specific tasks, such as image segmentation [16-18]. First, asymmetric autoencoders perform well in mask generation. This structure allows the model to focus more on learning and preserving the exact details of the mask, especially critical since in image cutting we need to accurately capture the contours and details of the object. The optimized structure of the decoder helps to generate more accurate masks, improving the quality of the cutting result and making it particularly suitable for applications that require a high degree of precision.

Transforming a conventional symmetric autoencoder into an asymmetric autoencoder yields notable advantages, particularly in specialized tasks like image segmentation [16-18]. The distinctive feature of asymmetric autoencoders shines in mask generation. This architecture empowers the model to prioritize the acquisition and preservation of intricate details within the mask. This emphasis on precision is crucial, especially in image segmentation tasks where capturing accurate contours and intricate object details is paramount.

The optimized design of the asymmetric autoencoder's decoder plays a pivotal role in generating more precise masks. This enhancement translates into improved cutting results, elevating the overall quality of the process. As a consequence, this model configuration proves especially well-suited for applications demanding a high level of precision. The emphasis on learning and preserving intricate details ensures that the model excels in tasks requiring nuanced and accurate representations, making it a valuable tool in image cutting scenarios and similar applications.

Furthermore, the asymmetric autoencoder proves to be an efficient solution for conserving computational and memory resources, particularly advantageous when dealing with extensive image datasets or operating in resource-constrained settings. The simplification of the decoder contributes to a significant reduction in model complexity without compromising performance efficiency. This streamlined architecture positions asymmetric autoencoders as an ideal choice for handling substantial image data, delivering exceptional results even within the constraints of limited computing resources.

Dilation and Erosion

Dilation and Erosion are two basic operations in Mathematical Morphology, commonly used in image processing and computer vision.

Dilation is an operation performed on an image, usually using a structure element to enlarge or magnify a specific area of the image. For each element in a structure element, the corresponding area of the image is set to white, and if one of the elements in the structure element matches an element in the image, that area is marked white.Dilation can emphasize or enlarge bright areas of an image, connect adjacent white areas, and fill in gaps.

Erosion is an operation on an image that also uses a structure element, but it shrinks or erodes specific areas. For each element in the structure element, the corresponding area in the image is set to white, and the area is marked white only if all elements in the structure element match those in the image.Erosion shrinks the white areas of the image, separates neighboring areas, and removes small contiguous areas.

Figure 5 shows an example of Dilation and Erosion operating on an image.

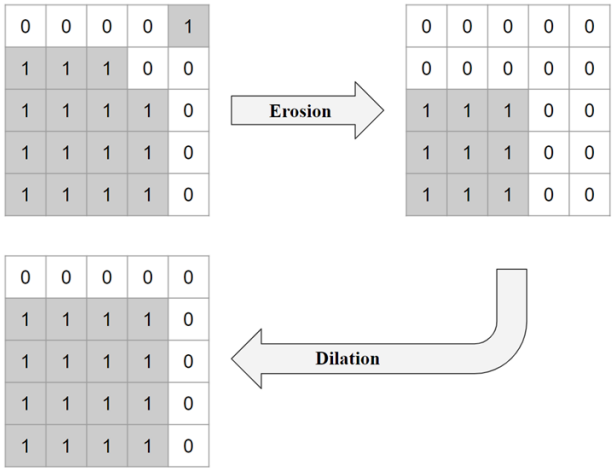


Figure 5. Examples of Dilation and Erosion.

IV. EXPERIMENTAND ANALYSIS

The experimental results of this study are mainly compared with Dice coefficient, which is a statistical index used to measure the similarity between two sets and is usually used in the fields of image segmentation, natural language processing and data mining. Dice coefficient can evaluate the degree of overlap between two sets, and its value ranges from 0 to 1. The closer the value is to 1, the greater the overlap between two sets and the higher the similarity.

$$Dice(A,B) = (2 * |A \cap B|) / (|A| + |B|), \quad (1)$$

Figure 6 shows the images used for experimental segmentation. In this study, lung X-rays with imaging data collected from Kaohsiung Medical University (KMU) database were used to test our method. Deep Convolutional Neural Network U-net [6] and ResNet [7] were used for comparison.

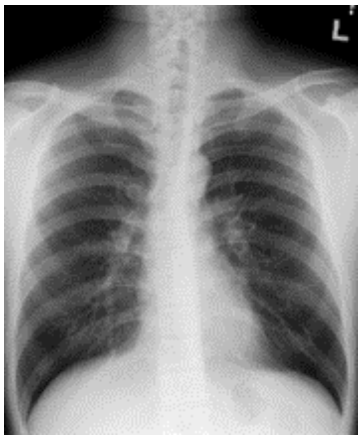


Figure 6. Experimental image for segmentation

Using different number of decoders, a comparison of our method with U-net and ResNet for lung radiographs after segmentation is presented in Figure 7.

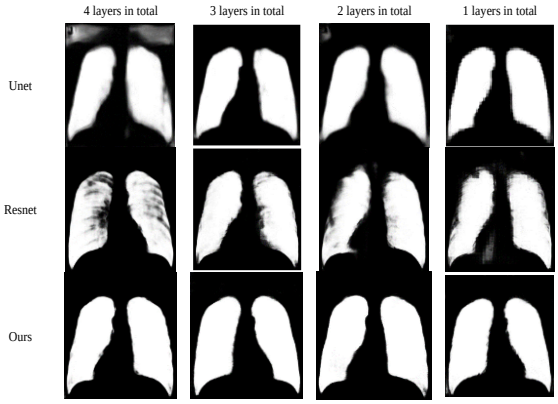


Figure 7. A comparison of our method with U-net and ResNet for lung radiographs.

Table 1 shows the comparison of our method with U-net and ResNet in terms of the number of parameters for lung X-ray radiographs segmentation using different number of decoders.

Table 1. The comparison of our method with U-net and ResNet in terms of the number of parameters for lung X-ray radiographs segmentation

	4 layers in total	3 layers in total	2 layers in total	1 layers in total
Unet	21.2M	20M	19.4M	19.2M
Resnet	8.0M	7.0M	6.3M	5.7M
Ours	25.5M	24.6M	24.3M	24.2M

Table 2 shows the Dice coefficients of our method compared with U-net and ResNet for lung X-ray radiographs segmentation using different number of decoders.

Table 2 shows the Dice coefficients of our method compared with U-net and ResNet

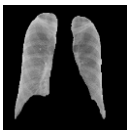
	4 layers in total	3 layers in total	2 layers in total	1 layers in total
Unet	0.9102	0.9248	0.9007	0.9235
Resnet	0.9641	0.9581	0.9582	0.9467
Ours	0.9654	0.9656	0.9625	0.9603

Ablation studies of the asymmetric self-coders performed in this study are presented in Tables 1 and 2. The outcomes of these experiments indicate that a reduction in computational cost and alleviation of over-simulation can be attained by employing fewer decoders to formulate the self-coder while maintaining a constant number of encoders. Throughout the conducted experiments, our approach effectively identifies the optimal number of decoders by systematically adjusting the count of decoders, demonstrating its adaptability to find the most suitable configuration. The results underscore the significance of this approach in achieving a balance between computational efficiency and model performance, providing valuable insights into the strategic utilization of decoders within the self-coding architecture. This experimental exploration contributes valuable findings to the broader understanding of self-coding mechanisms and their parameter configurations, opening avenues for enhanced efficiency in various applications leveraging self-coding architectures.

As depicted in the table, the optimal Dice coefficient value for the segmented image is 0.9656. Consequently, employing an extensive number of decoders tends to overly emphasize details in lung X-rays, leading to substantial noise in the output mask. Conversely, a sparse number of decoders results in subpar image recovery due

to insufficient detailing. Striking a balance by utilizing an appropriate number of decoders yields a mask that closely aligns with the lung contours. This highlights the positive impact of adjusting the decoder count, evident in both the quantitative data and the quality of the output image. The findings underscore the importance of optimizing the decoder count to achieve a harmonious trade-off between detailed feature representation and noise reduction in the segmented lung images.

Table 3 Comparison of post-processing

	Unprocessed X-ray images	X-ray images processed by median filter	X-ray images processed by Dilation and Erosion
Sample 1			
Dice	0.9385	0.9404	0.9661
Sample 2			
Dice	0.9112	0.9133	0.9473

Following the precise segmentation of the lungs, post-processing steps involve the application of Dilation and Erosion to eliminate extraneous noise. Two lung X-rays are subjected to testing, as outlined in Table 3, with the addition of a median filter for reference purposes. An examination of the results in Table 3 reveals that the median filter is primarily adept at addressing irregularities along the lung periphery but exhibits limited efficacy in handling extensive noise during post-processing. Conversely, Dilation and Erosion prove more effective in mitigating ambient noise, showcasing superior results in terms of the Dice coefficient. These morphological operations not only outperform the median filter in noise reduction but also contribute to refining the segmentation output, highlighting their utility in enhancing the accuracy of lung image analysis. The comparative evaluation underscores the significance of selecting appropriate post-processing techniques tailored to the specific characteristics of lung X-rays for optimal segmentation outcomes.

V. CONCLUSIONS

The widespread occurrence of COVID-19 and influenza globally in recent years has underscored the importance for frontline clinicians to precisely identify specific lung regions when diagnosing radiographs. In this study, we introduce a novel methodology aimed at accurately delineating the precise location of lungs within radiographic images. This approach selectively retains only the mask corresponding to the lung region, thereby creating a Region of Interest that proves invaluable for physicians in aiding their diagnostic processes. The proposed method enhances the efficiency and accuracy of diagnostic evaluations by providing clinicians with a focused and clearly defined area for examination within radiographic images.

Our algorithm is structured into three key stages: Depthwise Separable Convolution, Attention Enhancing Block, and Asymmetric Autoencoder. The Depthwise Separable Convolution adeptly captures X-ray images with constrained computational resources. The Attention Enhancing Block plays a pivotal role in extracting features from the X-ray image by utilizing three distinct receptive fields. The resulting fused features are subsequently condensed by the decoder. The Asymmetric Autoencoder model places particular emphasis on the acquisition and preservation of intricate details within the masks of the Region of Interest. This three-stage approach optimally balances computational efficiency, feature extraction, and detailed preservation to contribute to the effectiveness of our algorithm in the context of X-ray image analysis.

Our method underwent testing using lung radiographs sourced from the Kaohsiung Medical University database. The simulation results demonstrate the superior performance of our proposed method, showcasing a higher Dice coefficient compared to alternative segmentation methods. Additionally, our method excels in precisely locating the image segmentation of the lung Region of Interest, a critical requirement for clinicians' accurate diagnosis. The efficacy of our proposed method lies in its ability to not only achieve superior segmentation results but also accurately pinpoint the specific lung ROI, enhancing its utility in medical diagnostics.

REFERENCES

[1] Rachna Jain, Meenu Gupta, Soham Taneja, and D. Jude Hemanth, "Deep learning based detection and analysis of COVID-19 on chest X-ray images", Applied Intelligence, 51:1690–1700, Oct. 2020

[2] Shiv Goel, Adam Kipp, Nirmit Goel, and Jingjing Kipp, "COVID-19 vs. Influenza: A Chest X-ray Comparison", Cureus, 14(11): e31794, Nov. 2022

[3] Xi-ming Wang, Su Hu a, Chun-hong Hu, Xiao-yun Hu, Yi-xing Yu, Ya-fei Wang, Jian-liang Wang, Guo-hua Li, Xin-feng Mao, Jian-chun Tu, Ling Chen, Wei-feng Zhao, "Chest imaging of H7N9 subtype of human avian influenza", Radiology of Infectious Diseases, Vol. 1, no 2, pp. 51-56, Mar. 2015

[4] Agata Giełczyk, Anna Marciniak, Martyna Tarczewska, Zbigniew Lutowski, "Pre-processing methods in chest X-ray image classification" PLoS One, 5;17(4):e0265949., Apr. 2022

[5] Kermany D S, Goldbaum M, Wenjia Cai, Carolina C.S, Valentim, Liang H, Baxter S, McKeown A, Yang G, Wu X, Yan F, Dong J, Prasadha M, Pei J, Tin M, Zhu J, Li C, Hewett S, Dong J, Ziyar I, Shi A, Zhang R, Zheng L, Hou R, Shi W, Fu X, Duan Y, Huu V, Wen C, Zhang E, Zhang C, Li O, Wang X, Singer M, Sun X, Xu J, Tafreshi A, Lewis M, Xia H, Zhang K. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. Cell, 172(5):1122-1131.e9., 2018,

[6] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015.

[7] Huang G, Liu Z, Laurens V, Weinberger K. Densely Connected Convolutional Networks. IEEE Computer Society, 2016.

[8] Karimpouli, S., & Tahmasebi, P. Segmentation of digital rock images using deep convolutional autoencoder networks. Computers & geosciences, 126, 142-150, 2019

[9] Chollet, F. Xception, "Deep learning with depthwise separable convolutions", In Proceedings of the IEEE conference on computer vision and pattern recognition , pp. 1251-1258, 2017.

[10] Kaiser, L., Gomez, A. N., & Chollet, F, "Depthwise separable convolutions for neural machine translation", arXiv preprint arXiv:1706.03059. 2017

[11] Bai, L., Zhao, Y., & Huang, X., "A CNN accelerator on FPGA using depthwise separable convolution", IEEE Transactions on Circuits and Systems II: Express Briefs, 65(10), pp.1415-1419, 201.

[12] Koonce, B., & Koonce, B, "ResNet 34. Convolutional Neural Networks with Swift for Tensorflow", Image Recognition and Dataset Categorization, pp. 51-61, 2021

[13] Abedalla, A., Abdullah, M., Al-Ayyoub, M., & Benkhelifa, E, "The 2ST-UNet for pneumothorax segmentation in chest X-Rays using ResNet34 as a backbone for U-Net", arXiv preprint arXiv:2009.02805, 2020

[14] Dong, H., Pan, J., Xiang, L., Hu, Z., Zhang, X., Wang, F., & Yang, M. H., "Multi-scale boosted dehazing network with dense feature fusion", In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 2157-2167, 2020

[15] Wang, G., Gan, X., Cao, Q., & Zhai, Q., "MFANet: multi-scale feature fusion network with attention mechanism", The Visual Computer, 39(7), pp. 2969-2980, 2023

[16] Majumdar, A., & Tripathi, A., "Asymmetric stacked autoencoder", In 2017 International Joint Conference on Neural Networks (IJCNN), pp. 911-918, May 2017

[17] Meng, R., Yin, S., Sun, J., Hu, H., & Zhao, Q., "scAAGA: Single cell data analysis framework using asymmetric autoencoder with gene attention", Computers in Biology and Medicine, 165, 107414, 2023

[18] Kim, J. H., Choi, J. H., Chang, J., & Lee, J. S., "Efficient deep learning-based lossy image compression via asymmetric autoencoder and pruning", In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2063-2067, May 2020

Chen Yen-Yu received his Ph.D. degree in electrical engineering from National Cheng Kung University in 2004. Currently, he is working as an associate professor in the Department of Artificial Intelligence and Computer Engineering at National Chin-Yi University of Science and Technology. His research interests include biomedical signal processing and machine learning.

Chen Ching-Cheng received the B.S. degree in Information Management from Tamkang University and is currently pursuing the M.S. degree in Information Management at National Chung Hsing University. His current research focuses on lung autoencoder segmentation networks and pneumoconiosis detection networks.