A Two-Stage Neural Network Model for Breast Ultrasound Image Classification

Bining Long, Yanran Guan and Matthew Holden
School of Computer Science
Carleton University
Ottawa, Canada

***Abstract*—Breast cancer continues to be a prominent contributor to female mortality. Ultrasound imaging stands as a widely utilized technique for detecting breast abnormalities. In this paper, we introduce a novel two-stage neural network model to classify breast cancer in ultrasound images. In the first stage, we employ a fully convolutional network (FCN) to perform image segmentation. The FCN learns to predict segmentation masks from the breast ultrasound images, delineating tumor regions. Subsequently, the second stage involves a convolutional neural network (CNN) to classify tumor type, leveraging tumor masks generated by the first stage and the original ultrasound images. Results showcase the added value of the two-stage approach, with our proposed model achieving a classification accuracy of $92.41\,\%$, consistently surpassing the performance of baseline models that rely solely on CNNs for breast ultrasound image classification.**

***Index Terms*—Breast ultrasound images, image classification, image segmentation, two-stage learning.**

I. Introduction

Breast cancer is one of the leading causes of death among women worldwide. Early detection and diagnosis of breast cancer are crucial for effective treatment, which could help reduce the mortality rate through appropriate therapeutic interventions at the right time [1]. One of the most widely used imaging modalities for detecting breast cancer is ultrasound imaging. It can detect suspicious areas or abnormalities in the breast tissue, which may indicate the presence of a tumor. The tumor imaging features, such as size and border shape, can then be used to further classify tumors as normal (i.e., no tumor), benign, or malignant [2].

In prior work on medical image analysis, Pei et al. [3] presented a context-aware deep learning methodology to address various brain tumor analysis tasks using structural multimodal magnetic resonance images. This involved tumor segmentation by a context-aware deep neural network (DNN) model, subtype classification by a regular convolutional neural network (CNN) [4], and prediction of overall survival by a hybrid method of deep learning and machine learning. In a similar context, Samee et al. [5] proposed a computer-aided diagnosis (CAD) system targeting brain tumors in magnetic resonance imaging. They streamlined a U-Net architecture [6] to delineate regions of interest, coupled with a simplified CNN to classify tumors as benign or malignant. Additionally, Anand and colleagues [7] presented a fusion of U-Net and CNN models for skin lesion classification using dermoscopy images. Their technique entailed the multiplication of U-Net-derived masks with preprocessed images to generate segmented images, which were subsequently utilized by the CNN model for further classification. Sudharson and Kokil [8] proposed a strategy to classify kidney ultrasound images by employing an ensemble of pre-trained DNNs using transfer learning. This approach led to performance improvements surpassing those achieved by individual DNNs.

In the realm of breast cancer diagnosis, Xie et al. [9] presented a CAD system for ultrasound images. The system initially classified normal and cancerous samples using a pre-trained ResNet [10] through transfer learning, then performed accurate tumor segmentation via an enhanced mask region-based CNN. Moreover, Mohamed et al. [11] presented an automated system for breast cancer detection from thermograms. It involved a step of segmenting the breast area from the rest of the body, followed by binary classification using a CNN model. Furthermore, a multistage transfer learning approach was introduced by Ayana et al. [12] for breast ultrasound image classification, incorporating medical images during model pre-training.

This paper proposes a novel two-stage neural network model for the classification of breast cancer based on ultrasound images. In our approach, the initial stage employs a fully convolutional network (FCN) [13] to predict segmented tumor masks. The predicted segmentations are then combined with the original ultrasound images in the second stage for tumor classification by a CNN. Our objective is to explore the potential value added by this two-stage model integrating image segmentation and classification techniques from prior research, compared to directly utilizing CNN classification models on ultrasound images.

II. Methods

In this section, we describe the data preprocessing methodology, present the detailed design of the two stages of our model, and discuss the evaluation and validation strategies.

A. Data Preprocessing

The data preprocessing for our model involves resizing input images and generating ground truth (GT) labels for categorical classification. To accommodate GPU memory constraints and ensure runtime efficiency, we resize the original input ultrasound images, which have an average size of 500×500 pixels,

to a smaller dimension of 128×128 pixels. The GT labels for the final classification are derived from the file names of the provided GT images.

B. Tumor Segmentation Using FCN

The ultrasound images contain a multitude of tumor features. However, the presence of redundant information may lead to tedious calculations that take a long computational time, but without significant enhancement in the final classification performance [14]. To mitigate this problem, we opt to extract the essential information that helps improve the classification accuracy. Thus, the first stage of our model employs an FCN for tumor segmentation within the ultrasound images.

We build the FCN based on the U-Net architecture proposed by Ronneberger et al. [6]. The U-Net architecture is renowned as one of the most widespread image segmentation techniques and is applied in various medical image modalities [15]. We have tailored it to our specific task through adjustments in layer structure and meticulous tuning of hyperparameters. The U-Net consists of a contraction path that acts as an encoder and an expansion path that acts as a decoder. Each layer in the expansion path is concatenated with a layer from the contraction path to preserve localization information [6]. In our modified U-Net structure, there are 5 blocks in the contraction path, each consisting of two convolutional layers with a kernel size of 3×3 followed by a batch normalization layer and rectified linear unit (ReLU) activation. Then, a 2×2 max-pooling layer is applied, followed by a dropout layer. In the expansion path, there are 4 blocks, each consisting of a transposed convolutional layer with a kernel size of 3×3 concatenated with a corresponding layer from the contraction path, and a dropout layer followed by two sequential 3×3 convolutions, which are then succeeded by a batch normalization layer and ReLU activation. The final output layer in this U-Net is a convolutional layer with kernel size 1×1 activated by a sigmoid function that outputs single-channel grayscale images representing the predicted tumor masks.

C. Tumor Classification Using CNN

After obtaining the tumor segmentation output from the first stage, we progress to the second stage of our network: a CNN classification model. However, building an accurate image classification model from scratch can be challenging due to the data scarcity problem and the limitation of time and hardware resources [16]. An effective approach is to leverage transfer learning, which has demonstrated the ability to overcome these challenges and has significantly contributed to the field of medical image analysis [16]. Therefore, we implement transfer learning using a pre-trained DenseNet-201 [17] as the feature extractor, followed by 3 trainable 64-unit fully connected (FC) layers activated by ReLU layers and a dropout layer. The output layer is a 3-unit FC layer that performs classification, activated by a softmax function and each entry corresponding to a tumor class.

The DenseNet-201 model is imported from Keras[1]. It is initialized with the pre-trained weights trained on the ImageNet dataset [18]. The convolutional layers are set to be non-trainable. The DenseNet-201 model is originally built to process images of 224×224 pixels and 3 channels. However, since we only use the convolutional layers, the model is able to learn features from images of 128×128 pixels as well. Given that the output from the prior stage consists of single-channel grayscale images, we replicate each of these images twice and concatenate them with the original ultrasound images to create a 3-channel input that aligns with the expectation of the DenseNet-201 model. This concatenated input is then fed into the DenseNet-201 model and it outputs a 3D array representing the extracted features from the input images.

D. Evaluation and Validation Strategies

Our two-stage model is evaluated on the experimental dataset using the hold-out validation strategy. The dataset is randomly shuffled and split into training, validation, and test sets with the ratio $80 : 10 : 10$. Each set of data contains the same proportion of images from the 3 classes to obtain a fair assessment of the model performance.

During training, the U-Net model is evaluated and optimized using the binary cross-entropy loss function. In the classification stage, the learning objective of the CNN model is to minimize the categorical cross-entropy loss. We choose the best-performing U-Net model based on the binary cross-entropy loss on the validation set, and the best-performing CNN model based on the classification accuracy on the validation set.

III. Experimental Setup

In this section, we explain the details of our experimental methodology. The implementation of our model can be found in our code[2].

A. Experimental Data

We used the breast ultrasound images dataset [19] consisting of 780 breast ultrasound images. All the images are in `.png` format with an average size of 500×500 pixels and each pixel contains a grayscale value on the range $[0, 1]$. The dataset is categorized into 3 classes: normal, benign, and malignant, which have 133, 437, and 210 images, respectively. Each image has its corresponding GT segmentation mask. These masks were hand-generated by Al-Dhabyani et al. [19] and reviewed by radiologists from Baheya hospital. Annotations were created and added to the file names of GT images, from which we derived the GT labels.

B. Model Tuning and Configuration

During the tuning process of the U-Net model, we explored various hyperparameters and model structures. These experimental components include the learning rate, number of epochs, dropout rate, selection between bilinear upsampling

[1] https://keras.io/

[2] https://github.com/bininglong/Breast-Ultrasound-Image-Classification

layers or transposed convolutional layers in the expansive path, utilization of either L_1 or L_2 regularization for the parameters, and incorporation of batch normalization layers. The tuning process was conducted by iterating a tuning loop for 5 times. Initially, a combination of the experimental components was pre-defined. Within each iteration, we modified the choice of each component and trained our model with each updated combination for 100 epochs using the Adam optimizer [20]. Throughout this process, we monitored the validation loss. When a lower validation loss is achieved, we update the choice for the corresponding component, maintaining the selections for the other components unchanged. In this manner, we experimentally found an optimal configuration for the U-Net. This configuration involves the following choices: setting a dropout rate of 0.1 for each dropout layer, employing a learning rate of 0.001, integrating transposed convolutional layers in the expansive path, excluding parameter regularization, and incorporating batch normalization layers.

We tuned the CNN model of the second stage based on the outputs of the best-performing U-Net model. Using a similar tuning methodology, we found the following optimal configuration for the CNN model: a learning rate of 0.001, a dropout rate of 0.05, and the use of 3 FC layers, each having 64 neurons.

To find an optimal number of training epochs, we again trained our models for the two stages with the aforementioned configurations for 100 epochs. Observing the U-Net's and the CNN's optimal validation performance typically occurring around 50 epochs and 20 epochs, respectively, we chose to train the two stages for these specific durations.

C. Machine Configuration and Timing

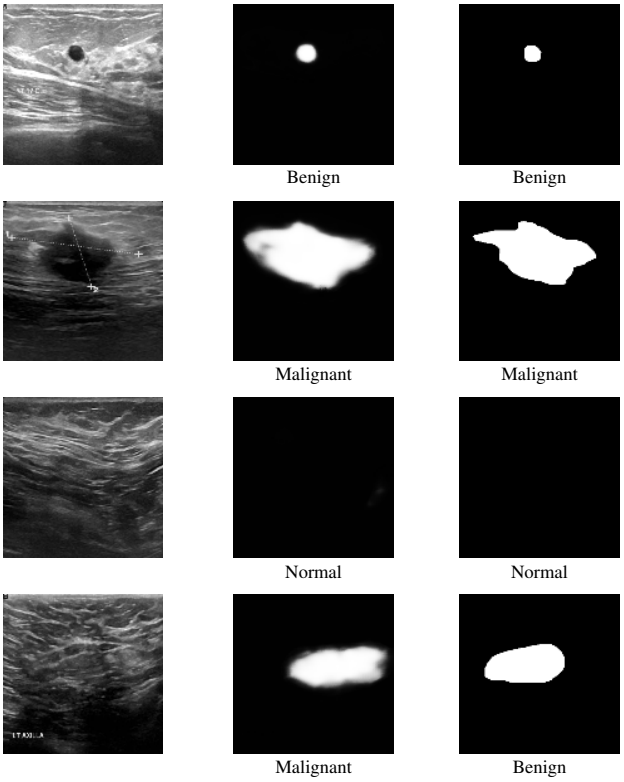
The experiments were carried out on the Ubuntu 18.04 LTS operating system with an NVIDIA GeForce RTX 2080 Ti GPU with $11\,\mathrm{GB}$ of memory and CUDA version 10.1. Training our U-Net model and CNN model requires on average $45.75\,\mathrm{s}$ and $371.90\,\mathrm{s}$, respectively. The inference time of our two-stage model is $4.98\,\mathrm{s}$ per image.

IV. Results

In this section, we show the qualitative and quantitative evaluations of our model.

A. Qualitative Results

We visually present in Fig. 1 several results of tumor segmentation and classification achieved by our model, drawn from a random selection of correctly classified cases for each class, as well as misclassified cases. These results feature the following components: the original ultrasound images (Fig. 1a), the predicted masks with the corresponding predicted labels (Fig. 1b), and the GT masks along with their GT labels (Fig. 1c). We can see that the U-Net model, serving as the first stage, is able to effectively mark the boundaries of tumor regions in the original ultrasound images, aiding the second stage of our model in performing an accurate classification. Please note that, in certain rare instances, misclassifications may occur. For instance, as shown in Fig. 1, a benign tumor is misclassified as malignant. This could potentially be attributed to the tumor's relatively large region.



(a) Ultrasound images (b) Predicted masks (c) GT masks

Fig. 1: Examples of our model's segmentation and classification results, with predicted labels and GT labels shown below corresponding masks.

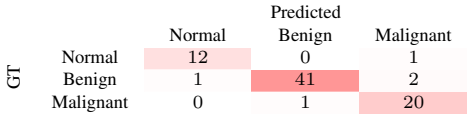
B. Quantitative Results

Using the configuration as explained in Section III-B, we conducted 10 rounds of training and evaluation for the proposed two-stage model. Among all the training rounds, the U-Net model achieved an average validation loss of 0.097 and the CNN model obtained an average validation accuracy of $91.03\,\%$, with the best-performing instance achieving a validation accuracy of $93.59\,\%$. To comprehensively assess the model's strengths and weaknesses, we present the confusion matrices of our model in Fig. 2, derived from both the validation and test sets. Notably, our model consistently avoids misclassifying malignant tumors as normal, which is a positive outcome.

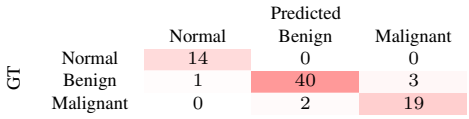
To further evaluate our model's performance, we conducted an ablation analysis on the input of the classification stage. We summarize this evaluation in Table I, where we compare the classification performance on the test set based on 4 types of inputs: the original ultrasound images, the predicted masks derived from our first stage, the predicted masks concatenated with the ultrasound images, and the GT masks concatenated

TABLE I: The overall classification accuracy and the class-wise precision, recall, and F_1-score obtained on the test set across various input types in the second stage of our model.

	Accuracy	Precision			Recall			F_1-score		
		Normal	Benign	Malignant	Normal	Benign	Malignant	Normal	Benign	Malignant
Ultrasound images	0.8734	0.8125	0.9070	0.8500	0.9286	0.8864	0.8095	0.8667	0.8966	0.8293
Predicted masks	0.8861	1.0000	0.8889	0.8000	1.0000	0.9091	0.7619	1.0000	0.8989	0.7805
Predicted masks + ultrasound images	0.9241	0.9333	0.9524	0.8636	1.0000	0.9091	0.9048	0.9655	0.9302	0.8837
GT masks + ultrasound images	0.9873	1.0000	0.9778	1.0000	1.0000	1.0000	0.9524	1.0000	0.9888	0.9756



(a)



(b)

Fig. 2: Confusion matrices of our model on (a) the validation set and (b) the test set.

with the ultrasound images. We use the overall accuracy, as well as the precision, recall, and F_1-score on each tumor class, as the evaluation metrics. We see that concatenating correctly segmented tumor masks can significantly improve the classification accuracy. When using only the original ultrasound images as input, the model achieved an accuracy of $87.34\,\%$. With the GT masks added to the ultrasound images as input, the accuracy was significantly improved to $98.73\,\%$. In our study employing the two-stage model, we observed a performance improvement when utilizing predicted masks. This improvement was further augmented when these masks were concatenated with the original ultrasound images. Specifically, our approach led to a remarkable $5.07\,\%$ increase in overall classification accuracy compared to direct classification solely using ultrasound images, elevating it from $87.34\,\%$ to $92.41\,\%$.

Moreover, we conducted an extensive comparison between our two-stage model and various CNN baseline models using pre-trained weights on ImageNet. For each type of CNN, we employed the variant with the highest reported performance on ImageNet from Keras Applications[3]. To ensure a controlled experiment, we applied the same configuration and training process to every pre-trained CNN model as we did to our model, involving adding FC layers and a dropout layer. We summarize this comparison in Table II, where we compare the test classification performance using multiple metrics, including overall accuracy, and per-class precision, recall, and F_1-score. We observe that, our model, leveraging the incorporation of predicted segmentation masks and the original ultrasound images, achieves the highest classification performance across all comparative baselines. It attains the highest overall accuracy, the highest precision and F_1-score on all the classes, and the highest recall on normal and malignant classes.

[3]https://keras.io/api/applications/

V. Conclusion, Limitations, and Future Work

We have proposed a novel two-stage model for breast ultrasound image classification. Our model has demonstrated not only a high overall accuracy but also consistently high F_1-scores across all tumor classes. This robust performance highlights our model's proficiency in correctly classifying tumors, even when confronted with an imbalanced dataset. These findings provide substantial evidence that our model can be considered reliable for use in clinical practice, offering decision support for breast cancer detection and diagnosis.

Furthermore, our work has showcased the significance of feature selection and the effectiveness of incorporating a segmentation stage, which extracts and integrates crucial features with the original inputs to enhance the classification performance. These implications may be extended beyond medical image classification, to a wider range of neural network research scenarios.

The limitation and biases in our experiment stem from the utilization of a relatively small and imbalanced dataset and the randomness in the model tuning process, which may produce slightly biased results, posing challenges in selecting the optimal model structure. To mitigate these biases, dropout layers were temporarily removed during specific tuning stages.

As a better segmentation stage can substantially boost the classification performance, we suggest leveraging a U-Net with attention gates [32] in future work. There is also room for enhancing data preprocessing and refining the evaluation metrics to fine-tune the model architecture. To address the challenge posed by imbalanced data, it could be beneficial to employ resampling techniques such as upsampling the minority class and downsampling the majority class. If a larger GPU memory becomes available, implementing a k-fold cross-validation can help mitigate model overfitting, especially when dealing with a limited dataset. Further potential lies in exploring the applications of our model in real-world clinical scenarios and adapting it for diverse tasks in medical image analysis.

References

[1] D. B. Kopans, "Breast-cancer screening with ultrasonography," *The Lancet*, vol. 354, no. 9196, pp. 2096–2097, 1999.

[2] Z. Rezaei, "A review on image-based approaches for breast cancer detection, segmentation, and classification," *Expert Systems with Applications*, vol. 182, p. 115204, 2021.

TABLE II: The overall classification accuracy and the class-wise precision, recall, and F_1-score obtained on the test set from different models.

	Accuracy	Precision			Recall			F_1-score		
		Normal	Benign	Malignant	Normal	Benign	Malignant	Normal	Benign	Malignant
Ours	**0.9241**	**0.9333**	**0.9524**	**0.8636**	**1.0000**	0.9091	**0.9048**	**0.9655**	**0.9302**	**0.8837**
VGG-16 [21]	0.7468	0.6111	0.8571	0.6923	0.7857	0.6818	0.8571	0.6875	0.7595	0.7660
ResNet-152 [10]	0.6582	0.4737	0.8485	0.5556	0.6429	0.6364	0.7143	0.5455	0.7273	0.6250
ResNet-152-V2 [22]	0.8354	0.7500	0.8478	0.8571	0.6429	0.8864	0.8571	0.6923	0.8667	0.8571
Inception [23]	0.8101	0.7647	0.8222	0.8235	0.9286	0.8409	0.6667	0.8387	0.8315	0.7368
Inception-ResNet [24]	0.8228	0.7857	0.8974	0.7308	0.7857	0.7955	**0.9048**	0.7857	0.8434	0.8085
Xception [25]	0.7975	0.7692	0.8222	0.7619	0.7143	0.8409	0.7619	0.7407	0.8315	0.7619
DenseNet-201 [17]	0.8734	0.8125	0.9070	0.8500	0.9286	0.8864	0.8095	0.8667	0.8966	0.8293
MobileNet [26]	0.7975	0.7857	0.8919	0.6786	0.7857	0.7500	**0.9048**	0.7857	0.8148	0.7755
MobileNetV2 [27]	0.8228	0.9167	0.8039	0.8125	0.7857	**0.9318**	0.6190	0.8462	0.8632	0.7027
NASNet [28]	0.8228	0.8000	0.8571	0.7727	0.8571	0.8182	0.8095	0.8276	0.8372	0.7907
EfficientNet [29]	0.7468	0.7500	0.7647	0.7000	0.4286	0.8864	0.6667	0.5455	0.8211	0.6829
EfficientNetV2 [30]	0.6076	0.5000	0.6290	0.5385	0.1429	0.8864	0.3333	0.2222	0.7358	0.4118
ConvNeXt [31]	0.7975	0.7143	0.8780	0.7083	0.7143	0.8182	0.8095	0.7143	0.8471	0.7556

[3] L. Pei, L. Vidyaratne, M. M. Rahman, and K. M. Iftekharuddin, "Context aware deep learning for brain tumor segmentation, subtype classification, and survival prediction using radiology images," *Scientific Reports*, vol. 10, no. 1, p. 19726, 2020.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] N. A. Samee, T. Ahmad, N. F. Mahmoud, G. Atteia, H. A. Abdallah, and A. Rizwan, "Clinical decision support framework for segmentation and classification of brain tumor MRIs using a U-Net and DCNN cascaded learning algorithm," in *Healthcare*, vol. 10, no. 12, 2022, p. 2340.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[7] V. Anand, S. Gupta, D. Koundal, and K. Singh, "Fusion of U-Net and CNN model for segmentation and classification of skin lesion from dermoscopy images," *Expert Systems with Applications*, vol. 213, p. 119230, 2023.

[8] S. Sudharson and P. Kokil, "An ensemble of deep neural networks for kidney ultrasound image classification," *Computer Methods and Programs in Biomedicine*, vol. 197, p. 105709, 2020.

[9] X. Xie, F. Shi, J. Niu, and X. Tang, "Breast ultrasound image classification and segmentation using convolutional neural networks," in *Advances in Multimedia Information Processing*, 2018, pp. 200–211.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[11] E. A. Mohamed, E. A. Rashed, T. Gaber, and O. Karam, "Deep learning model for fully automated breast cancer detection system from thermograms," *PLoS One*, vol. 17, no. 1, p. e0262349, 2022.

[12] G. Ayana, J. Park, J.-W. Jeong, and S.-w. Choe, "A novel multistage transfer learning for ultrasound breast cancer image classification," *Diagnostics*, vol. 12, no. 1, p. 135, 2022.

[13] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[14] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014.

[15] R. Azad, E. K. Aghdam, A. Rauland, Y. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof, "Medical image segmentation review: The success of U-Net," *CoRR*, vol. abs/2211.14830, 2022.

[16] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Medical Imaging*, vol. 22, no. 1, p. 69, 2022.

[17] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[19] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, p. 104863, 2020.

[20] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations*, 2015.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations*, 2015.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 630–645.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2818–2826.

[24] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, Inception-ResNet and the impact of residual connections on learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 4278–4284.

[25] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1800–1807.

[26] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017.

[27] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[28] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8697–8710.

[29] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural network," in *Proceedings of the International Conference on Machine Learning*, 2019, pp. 6105–6114.

[30] ——, "EfficientNetV2: Smaller models and faster training," in *Proceedings of the International Conference on Machine Learning*, 2021, pp. 10 096–10 106.

[31] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 966–11 976.

[32] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning where to look for the pancreas," *CoRR*, vol. abs/1804.03999, 2018.