Semi-supervised classification of medical ultrasound images based on generative adversarial network

Zhaoshan Liu  
Department of Mechanical Engineering  
National University of Singapore  
Singapore  
e0575844@u.nus.edu

Chau Hung Lee  
Department of Radiology  
Tan Tock Seng Hospital  
Singapore  
chau_hung_lee@ttsh.com.sg

Lei Shen*  
Department of Mechanical Engineering  
National University of Singapore  
Singapore  
mpeshel@nus.edu.sg

ABSTRACT

Medical ultrasound (US) is one of the most widely used imaging modalities in clinical practice. However, its use presents unique challenges such as variable imaging quality. Deep learning (DL) can be used as an advanced medical US image analysis tool, while the performance of the DL model is greatly limited by the scarcity of big datasets. Here, we develop semi-supervised classification enhancement (SSCE) structures by combining convolutional neural network (CNN) and generative adversarial network (GAN) to address the data shortage. A breast cancer dataset with 780 images is used as our base dataset. The results show that our SSCE structures obtain an accuracy of up to 97.9%, showing a maximum 21.6% improvement compared with utilizing CNN models alone and outperforming the previous methods using the same dataset by up to 23.9%. We believe our proposed state-of-the-art method can be regarded as a potential auxiliary tool for the diagnoses of medical US images.

Keywords  Semi-Supervised Learning · Generative Adversarial Network · Convolutional Neural Network · Medical Ultrasound Images · Transfer Learning

*Corresponding Author
1 Introduction

Medical ultrasound (US) has become a widely used screening and diagnostic tool in clinical practice due to its lack of ionizing radiation, high sensitivity, portability, and relatively low cost \[1\]. However, there are limitations. Image quality is easily affected by noise and artifacts, inter-operator variability is considerable, and variability across different US systems is usually high. Due to these, diagnosing medical US images always heavily relies on radiologists. To address the problem, developing an advanced medical US images analysis tool to make medical US diagnosis more objective, accurate, and automatic is essential.

In recent years, deep learning (DL), a branch of artificial intelligence, has emerged as a powerful tool to automate the extraction of useful information from big data. It has enabled ground-breaking advances in numerous computer vision tasks \[2\]. Among all tasks, classification is one of the most classic ones. Convolutional neural network (CNN), which was proposed by Lecun in 1998 \[3\], is one of the most effective and widely used types of DL models for classification. Generally, training CNN needs large datasets, while in the medical field, obtaining large datasets is difficult. Therefore, the transfer learning (TL) technique is usually applied to relieve the data shortage. The principle of the TL technique is to train models from pre-trained models to decrease the difficulty of training, and it has been successful in medical image classification \[4,5\]. However, the previous works usually use thousand-level datasets or larger ones. In many medical scenarios, such as the medical US, the public datasets are much scarce, and training CNN is considered more difficult. For example, it is reported that training CNN with a hundred-level dataset can only reach an accuracy of over 70\% \[6\], even if the TL technique is implemented.

The generative adversarial network (GAN) is a type of DL model proposed in 2014 \[7\] that can relieve the data shortage via synthesizing images. GAN is composed of a generator (\(G\)) and a discriminator (\(D\)), in which \(G\) and \(D\) play an adversarial “game”. In detail, when training GAN, the generator synthesizes some images based on the data distribution it learned from the real images, and the discriminator tries to discriminate whether the images are from real images or not. As training progresses, both generator and discriminator become more powerful because both of them are chased by the other. Compared with traditional data augmentation methods, GAN is a more generic solution. Traditional data augmentation methods usually generate images based on the real images themselves. For example, rotate or flip the real images. This is helpful but the number of the generated images is limited as the permutations of transformations are limited. However, since GAN synthesizes images based on the real data distribution, it can generate more images without obvious restriction. For the division, GAN can be divided into conditional GAN and unconditional GAN, depending on whether the input contains labels. Note that the TL technique can also be applied to GAN to improve its performance. Some previous research \[8,9,6,10\] has shown that the CNN can achieve high classification accuracy much easier with the images synthesized by GAN. However, the previous works have some limitations. First, either the resolution of the synthesized images is low or the quality of the synthesized images is not satisfactory. This may be caused by the structure of GAN or the TL technique is not implemented when training GAN. Second, the quality of the synthesized images is not examined using standard evaluation standards, thus the model collapse may be severe. Finally, the classification accuracy is not extremely high even though relatively big datasets are used.

To solve the abovementioned problems, we develop semi-supervised classification enhancement (SSCE) structures by combining CNN and GAN, where the TL technique is realized in both CNN and GAN. The major contributions of our works are highlighted below:
We synthesize medical US images with high resolution and high quality using GAN as well as evaluate them quantitatively.

- We develop SSCE structures that can reach a high classification accuracy with a small dataset.
- We propose a new evaluation standard to evaluate the performance of SSCE structures.

We evaluate our SSCE structures on a public dataset, namely the base dataset [11], with 780 images. The results show that our SSCE structures reach a classification accuracy of up to 97.9%. Compared with using CNN models alone, it shows a maximum 21.6% improvement. Compared with the existing methods using the same dataset, it realizes leadership by up to 23.9%.

The rest of this paper is organized as follows: In Section 2, we illustrate the datasets used in this work in detail, followed by the methods shown in Section 3. Section 4 shows the core experiment results and our discussions, and the conclusion together with the outlook for the future is presented in Section 5.

2 Datasets

The base dataset is a breast cancer dataset collected among 600 female patients between 25 and 75 years old in 2018. The system used to collect is LOGIQ E9 ultrasound and LOGIQ E9 Agile ultrasound system. There are 780 images in the base dataset and it is composed of three subsets: benign, malignant, and normal, corresponding to the patients’ breast cancer conditions and each subset has 437, 210, and 133 images, respectively. The average resolution of the images is around $500 \times 500$. Some examples from the base dataset are shown in Figure 1. The white letters and dotted lines in the examples are some medical notations. To keep consistency, all images in the figure of this work are at the resolution of $256 \times 256$, which equals the resolution of the synthesized images.

Besides the base dataset, four large datasets are used as the source datasets of the TL technique. For classification, ImageNet with 14M images is chosen. For synthesize, Flickr-Faces-HQ (FFHQ), Large-Scale CelebFaces Attributes (CelebA), and Large-scale Scene Understanding Challenge (LSUN) DOG with 70K, 200K, and 5M images respectively are selected.
Figure 2: The overall workflow of this study. The base dataset is used to train GAN. Once the GAN is trained well, it is utilized to synthesize images with pseudo labels to compose the extended datasets. Then the extended datasets are added to the base dataset. The acquired merged datasets are then fed into the CNN. By this, the classification accuracy can be improved to a large extent compared with feeding CNN with the base dataset alone, that is, without GAN. The dashed line shows the process of training CNN using the base dataset alone, and the dotted lines show the implementation of the TL technique, where different colors show different source datasets.

3 Methods

Here, we illustrate the implementation details of CNN, GAN, and SSCE. Different evaluation standards are also introduced. The overall workflow of this study can be found in Figure 2. The hyperparameters shown are selected after experiments. All the training is done using Tesla V100 with Python 3.8 and DL framework PyTorch 1.8.0.

3.1 Convolutional neural network

We custom seven CNN models, which are VGGNet [12], ShuffleNet [13], ResNeXt [14], ResNet [15], MobileNet [16], InceptionNet [17], and DenseNet [18] to classify medical US images. The specific version of these models are VGG16, ShuffleNet_v2_x1_0, ResNeXt50_32x4d, ResNet18, MobileNet_v3_large, Inception_v3, and DenseNet161 respectively. The specific way is that we adopt their original feature extraction parts as the backbones and add custom classification layers. The universal structure of the custom classification layers is shown in Figure 3. First, there is a linear layer with 512 out features, followed by a ReLU activation function and a dropout layer with a probability of 50%. Subsequently, another linear layer with 256 out features is added, followed by the same activation function and dropout layer. The last layer is a linear layer with 3 out features, which is the number of subsets.

The base dataset is divided randomly into a training set and a test set with a ratio of 8:2. The reason for not setting the validation set is to prevent further reducing the number of images available for training. The performance of the models on the training set and the test set are observed during training to prevent overfitting. The resolution of real images is preprocessed to $224 \times 224$ or $299 \times 299$ due to different CNN models’ requirements. The traditional data augmentation is present by default. The loss function used is cross-entropy. The Adam is set as the optimizer, where the learning rate equals 0.003 and the weight decay (WD) equals 0.001 if it exists. The number of epochs is set as 30 and the batch size is set as 32. Each CNN model is trained under two groups of settings, depending on whether the WD is
Figure 3: Universal structure of the custom classification layers. $n$ represents the number of in features and ReLU means the rectified linear unit.

The TL technique is set as default when training customed CNN models. Specifically, the models are not trained from scratch, instead, they are initialized by the weights learned from pertaining on ImageNet.

It is worth noting that to ensure the CNN model which performs best can work well on all sizes of datasets and explore more about how the size of the dataset influence the classification accuracy of other CNN models, we consider all CNN models when constructing SSCE structures.

3.2 Generative adversarial network

We implement four GAN models, which are deep convolutional GAN [19] (DCGAN), Wasserstein GAN (WGAN) [20], WGAN-GP [21], and StyleGAN2-ADA [22] to synthesize medical US images. The reason why we implement DCGAN, WGAN, and WGAN-GP, which are the most widely used GAN models for synthesizing, in this work is that we want to compare their performance with that of StyleGAN2-ADA, which is a newly proposed GAN model and has not been used. DCGAN is one of the most famous works that combine GAN and CNN, which trains $D$ and $G$ once in each epoch. WGAN introduces Wasserstein distance into GAN and uses RMSprop as the optimizer instead of Adam. WGAN-GP is a well-known modified model based on WGAN by adding gradient penalty and applying layer norm [23] in $D$. StyleGAN2-ADA combines the StyleGAN2 [24] with adaptive data augmentation (ADA) and has been proved to work well on small datasets. The first three GAN are unconditional GAN while the last one can be both conditional and unconditional, depending on the input. Our implementations are based on unconditional GAN because it has been reported that utilizing unconditional GAN can get better performance compared with using the conditional GAN [25].

GAN models are trained using three subsets of the base dataset respectively. The resolution of the real images is preprocessed to $256 \times 256$ and the resolution of synthesized images is also set as $256 \times 256$ to balance the image quality and time consumption. The optimizer for GAN models is Adam, RMSprop, Adam, and Adam, for DCGAN, WGAN, WGAN-GP, and StyleGAN2-ADA respectively, and the learning rate is set as 0.0002, 0.00005, 0.0001, and 0.0025 separately. The number of iterations is determined as 4000, and the batch size is chosen as 32. DCGAN, WGAN, and WGAN-GP are trained under one or two groups of settings, depending on whether the TL technique is implemented, which is determined by whether it can work normally at this resolution, as shown in Section 4.2. StyleGAN2-ADA is trained under four groups of settings, including training from scratch and using the TL technique from three source datasets to show the influence of the TL technique. The reason why we ensure it can work normally at this resolution is
that it has been proved to work well at the resolution of 1024 × 1024 [22]. The TL technique here is similar to that in Section 3.1 where the weights are learned from pertaining on FFHQ, CelebA, and LSUN DOG.

3.3 Semi-supervised classification enhancement

We develop our SSCE structures by combining seven CNN models and one GAN model to classify medical US images. We first use the GAN model selected in Section 4.2 to synthesize medical US images. Once the GAN model is trained well, that is, the synthesized images perform well in both qualitative and quantitative evaluation, we can endow the synthesized images with pseudo labels that are the same as the real images as the GAN model has learned the data distribution of the real images. Specifically, the images synthesized by the GAN trained with the benign subset are given benign as the pseudo labels, and the same operation is done for the malignant subset and the normal subset. Then, we use the synthesized images to compose the extended datasets. The proportion of the three subsets in the extended datasets is the same as that of the base dataset. The size of each extended dataset is an integer ($\gamma$) multiple of the base dataset, while its maximum value shows the number of the extended datasets. The maximum value of $\gamma$ is determined experimentally based on our proposed evaluation standard, the training efficiency index (TEI). The specific procedure is: (1) Set the initial value as 1, (2) calculate the TEI for all SSCE structures, (3) increase the value by 1, (4) calculate new TEI, (5) compare new TEI with previous ones, (6) if any TEI increases, repeat (3), (4), and (5) until TEI for all structures stops increasing. Next, we add the extended datasets to the base dataset to get the merged datasets. Finally, we train seven CNN models on the merged datasets.

The division and the preprocessing of the merged datasets are the same as that of the base dataset. Most of the hyperparameters remain unchanged except the number of epochs is doubled to 60 due to the increment of the number of images. For a given $\gamma$, each SSCE structure is trained under two groups of settings, depending on whether the WD is chosen or not. The setting of the TL technique on GAN is determined by the result obtained in and 4.2.

3.4 Evaluation standards

The evaluation standard for CNN models is the classification accuracy ($acc_b$) because the base dataset is small and the time consumption ($t_b$) is within the acceptable range.

To evaluate the quality of images synthesized by GAN model, besides observing directly, two of the most famous evaluation standards, which are inception score (IS) [26], and fréchet inception distance (FID) [27] are chosen.

IS can be calculated via:

$$IS = \exp\left(\mathbb{E}_{x \sim p_g} D_{KL}(p(y \mid x) \parallel p(y))\right)$$  \hspace{1cm} (1)

where $x \sim p_g$ indicates sample $x$ from $p_g$, and $D_{KL}$ represents the KL divergence. The lower the IS is, the worse the result is.

FID can be calculated by:

$$FID = \|m - m_w\|^2 + Tr\left(C + C_w - 2(CC_w)^{1/2}\right)$$  \hspace{1cm} (2)

where $w$ represents the real-world data, $m$ denotes the mean value, $C$ shows the covariance matrix, and $T$ represents the trace. The lower the FID is, the better the model performs.
It is worth noting that when calculating IS, the real images are not taken into consideration, thus the model may still get a high IS via directly copying the real images. Therefore, in this work, we regard FID as the main standard, and take IS as a reference. Both FID and IS are evaluated every 200 iterations.

For SSCE structures, as the number of images increases, the time consumption grows significantly. Thus, besides the classification accuracy, we propose TEI to balance the classification accuracy and the time consumption. The TEI can be calculated using:

$$TEI = \ln \left( t - t_b \right)^{-1} \cdot \left( acc - acc_b \right)$$

where $t$ denotes the time consumption, and $acc$ shows the classification accuracy. The lower the TEI, the worse the structure’s ability to get an ideal accuracy improvement within a limited time.
Figure 5: Records of FID and IS. The solid lines show the results for training from scratch, and the dotted lines illustrate the results for using the TL technique. (a) FID, and (b) IS.

Table 1: Classification accuracy of CNN models. ↑ means the higher, the better and ↓ inverse. Bold numbers show the best results.

| WD  | Standard | VGGNet | ShuffleNet | ResNeXt | ResNet | MobileNet | InceptionNet | DenseNet |
|-----|----------|--------|------------|---------|--------|-----------|--------------|----------|
| ✓   | acc↑     | 81.7%  | 72.2%      | 66.5%   | 70.3%  | 74.7%     | 65.8%        | 69.0%    |
| ✓   | tb↓      | 262.7s | 290.8s     | 300.6s  | 291.0s | 291.8s    | 327.4s       | 305.8s   |
| ×   | acc↑     | 82.3%  | 74.1%      | 69.6%   | 71.3%  | 73.4%     | 65.8%        | 74.7%    |
| ×   | tb↓      | 260.5s | 290.0s     | 300.3s  | 291.8s | 293.2s    | 327.3s       | 306.4s   |

4 Results and discussion

Here, we show the results of the supervised classification using CNN, the unsupervised synthesis using GAN, and the semi-supervised classification using SSCE. The discussions of the results are also presented.

4.1 Supervised classification

The classification accuracy of seven CNN models is shown in Table 1. From the table, it is found that VGGNet gets the highest accuracy with the lowest time consumption, showing a maximum accuracy of 82.3%, and the InceptionNet only reaches a maximum accuracy of 65.8%. The results are not quite satisfactory no matter which model. These results are within our expectations as the size of the base dataset is quite limited. Interestingly, no significant influence of the WD is observed.

4.2 Unsupervised synthesis

Images synthesized by DCGAN, WGAN, and WGAN-GP are shown in App A. It is found that the synthesized images have extremely low quality and diversity, showing that these GAN models cannot work well at this resolution at all. Due to these, quantitatively evaluating these models, using the TL technique on them, or adding their synthesized images are considered meaningless.
Figure 6: Comparison of the classification accuracy across different SSCE structures. For each structure, the first scatter corresponds to $\gamma$ equals 1, similarly for others. The closer to the lower right corner, the better the overall performance. (a) with WD, and (b) without WD.

We show the comparison of FID and IS across different TL experimental groups in App B. As can be seen, TL from FFHQ performs best compared with TL from CelebA and TL from LSUN DOG thus FFHQ is set as the default source dataset of the TL technique for StyleGAN2-ADA.

Some medical US images synthesized by StyleGAN2-ADA are shown in Figure 4. As can be seen, the quality of the images has a huge improvement compared with the images synthesized by DCGAN, WGAN, and WGAN-GP shown in App A, and the StyleGAN2-ADA can even mimic the medical annotations in the base dataset. Some obvious yellow flaws are found in Figure 4a without the TL technique, while the problem is solved after using it.

The records of FID and IS are shown in Figure 5. It is found that without the TL technique, the FID is high and the IS is low. Moreover, the change of FID and IS is frequent and drastic. These show that the performance and the stability are both poor. The problem is solved after using the TL technique. Some anomalous values are observed from the records. For instance, for the normal subset, we observe a better IS for training from scratch compared with using the TL technique at the 400th iteration, showing the limitation of IS in some cases.

In a summary, no matter for direct observation or quantitative assessment, using the TL technique can improve StyleGAN2-ADA’s performance to a large extent. For more details about how StyleGAN2-ADA learns from the real images, see App C which shows the images synthesized by StyleGAN2-ADA in the early training stages and the final training stage.

4.3 Semi-supervised classification

In Section 4.2, we found that the TL technique can largely improve the performance of StyleGAN2-ADA. Thus we set the TL technique as the default for it when developing SSCE structures. The classification accuracy of SSCE structures is shown in Table 2. Note that the maximum value of $\gamma$ is 8. By comparing it with that of using CNN models alone, it is found that the combination of VGGNet and StyleGAN2-ADA without WD (SSCE*) reaches the highest accuracy of 97.9%, showing a 15.6% accuracy improvement. For the combination of VGGNet and StyleGAN2-ADA with WD,
Table 2: Classification accuracy of SSCE structures. S stands for StyleGAN2-ADA. For each structure, \( \text{acc} \), \( t \), and TEI are obtained or calculated according to corresponding \( \gamma \), and \( m\text{acc} \) shows the maximum accuracy across all \( \gamma \). \( \uparrow \) means the higher, the better, and bold numbers show the best results.

| WD  | Standard | VGGNet + S | ShuffleNet + S | ResNeXt + S | ResNet + S | MobileNet + S | InceptionNet + S | DenseNet + S |
|-----|----------|------------|----------------|-------------|------------|---------------|-----------------|--------------|
| ✔️  | \( \text{acc} \) (\( m\text{acc} \uparrow \)) | 97.2% (97.3%) | 91.6% (91.6%) | 84.5% (84.5%) | 85.4% (85.4%) | 94.9% (94.9%) | 81.7% (82.7%) | 90.0% (90.4%) |
| ✔️  | \( t \) | 1480.8s | 1458.2s | 1467.5s | 1457.6s | 1460.3s | 1385.6s | 1263.7s |
| ✔️  | \( \gamma \) | 7 | 7 | 7 | 7 | 7 | 4 | 5 |
| ✔️  | \( \text{TEI} \uparrow \) | 2.19 | 2.75 | 2.55 | 2.53 | 2.86 | 2.29 | 3.06 |
| ✗  | \( \text{acc} \) (\( m\text{acc} \uparrow \)) | 97.8% (97.9%) | 95.0% (95.0%) | 81.2% (81.2%) | 86.6% (86.6%) | 94.9% (95.0%) | 82.4% (82.4%) | 88.8% (89.1%) |
| ✗  | \( t \) | 1478.4s | 1456.5s | 1251.9s | 1455.5s | 1460.3s | 1175.0s | 1260.1s |
| ✗  | \( \gamma \) | 7 | 7 | 5 | 7 | 7 | 3 | 5 |
| ✗  | \( \text{TEI} \uparrow \) | 2.19 | 2.97 | 1.69 | 2.18 | 3.04 | 2.45 | 2.06 |

we obtain an accuracy of 97.3% with the same increase in accuracy. These results illustrate that the VGGNet works well on all sizes of datasets. Though the structures composed of VGGNet and StyleGAN2-ADA reach the highest accuracy, some SSCE structures can obtain a higher accuracy improvement using lower time consumption, or in other words, have a higher TEI. For instance, the combination of DenseNet and StyleGAN2-ADA with WD reaches the TEI of 3.06, getting an accuracy improvement of 21.4%. The combination of MobileNet and StyleGAN2-ADA without WD obtains the TEI of 3.04, showing a 21.6% accuracy improvement. Besides four highlighted SSCE structures, significant accuracy improvement is observed for all SSCE structures, illustrating that our method is suitable for all CNN models and no cherry-picking operation is implemented.

To provide guidance for practical applications, we plot the comparison of the classification accuracy across different SSCE structures in Figure 5. It is found that the combination of VGGNet and StyleGAN2-ADA reach the highest accuracy using the lowest time consumption, no matter whether the WD is present or not. This illustrates that the combination of VGGNet and StyleGAN2-ADA should be considered first when deploying the classification task in practice. Unlike the results obtained in Section 4.1, the WD significantly influences the classification accuracy. Without the WD, training of some SSCE structures can become unstable, which is expressed as points A, B, and C in Figure 6.

In Table 3, we compare our SSCE* structure with the existing methods from previous reports using the same dataset, the base dataset. From the table, we can find that the accuracy of 97.9% exceeds all the listed methods, even for the binary classification, which is a much easier task compared with ours. This can show the effectiveness of our proposed structure and turns it into a new state-of-the-art milestone.

5 Conclusions

We propose SSCE structures to improve the classification accuracy of medical US images. The experimental results show that our proposed method is powerful and general for the classification task under the small dataset regime. With the outstanding performance, We believe that our proposed state-of-the-art method can potentially be regarded as an auxiliary tool for real-time medical US image diagnoses.

For future works, besides classification, developing methods for segmentation, detection, and other analysis tasks are necessary. Furthermore, proposing methods to analyze 3D medical US images is also important.
Table 3: Comparison of our SSCE* structure with the existing methods from previous reports. ↑ means the higher, the better. † shows the base dataset is used to execute binary classification tasks, which is much easier compared with ours, and the bold number shows the best results.

| References | Methods                                      | Year | macc↑ |
|------------|----------------------------------------------|------|-------|
| [28]       | Hybrid structure                             | 2021 | 95.6% |
| [6]        | Semi-supervised based on DAGAN               | 2019 | 88.0% |
| [29]       | ResNet                                       | 2021 | 88.9% |
| [30]       | TL, binary gray wolf optimization and machine learning | 2021 | 84.9% |
| [31]       | YOLO                                         | 2022 | 95.3% |
| [32]       | Vision transformer                           | 2021 | 74.0% |
| [33]       | Semi-supervised domain knowledge guided      | 2021 | 81.1% |
| [34]       | Ensemble learning †                          | 2020 | 90.8% |
| [35]       | VGGNet †                                     | 2021 | 84.9% |
| [36]       | CADx system †                                | 2020 | 96.6% |
| [37]       | Radiomics †                                  | 2021 | 97.4% |
| [38]       | Deep representations scale †                 | 2021 | 92.3% |
| Our        | SSCE*                                        | -    | 97.9% |

A Supplementary images

As shown in Figure A for DCGAN, the model collapse is severe, even for the biggest benign subset. For the smaller malignant subset, things become worse, the images become more blur and difficult to distinguish. For WGAN, its synthesized images seem to be composed of noise, and it’s difficult to find any useful information from these images. For WGAN-GP, it performs better compared with WGAN, while the results are still completely unacceptable.

B Transfer learning comparison

From Table B it is found that TL from FFHQ performs best compared with TL from CelebA and TL from LSUN DOG, getting an FID of 62.92, 68.78, and 73.92 for three subsets respectively. However, for the malignant subset, TL from FFHQ obtain a lower IS compared with TL from CelebA. The opposite conclusion proves the shortages of IS in some cases. One point worth denoting here is that the LSUN DOG, which has higher diversity, performs worst in our case. This result doesn’t meet Karras et al.’s conclusion, where the success of the TL technique seems to depend primarily on the diversity of the source datasets instead of the similarity between subjects [24]. We infer that their conclusion may be influenced by the relatively near relationship between dog and cat.

C Learning procedure

By analyzing Table C, we show that the TL technique can improve both the learning quality and the learning speed. For the 32nd iteration, the left images only shows some outlines, while the right images already look like real images. For the 128th iteration, left images occurs with severe model collapse, where right images do not. The remaining iterations show similar results.
Figure A: Images synthesized by DCGAN, WGAN, and WGAN-GP: (a) benign images, DCGAN (b) malignant images, DCGAN (c) benign images, WGAN, and (d) benign images, WGAN-GP.

Table B: Comparison of FID and IS across different TL experimental groups. In each group, the listed FID and IS are the optimal results calculated in the corresponding number of iterations. Bold numbers represent the best results.

| Group | Subset   | TL   | FID↓ | Iterations | IS↑ | Iterations |
|-------|----------|------|------|------------|-----|------------|
| 1     | Benign   | CelebA | 69.24 | 200 | **3.58** | 2200 |
| 2     | Benign   | LSUN DOG | 102.95 | 600 | 2.92 | 4000 |
| 3     | Benign   | FFHQ  | **62.92** | 1000 | **3.58** | 2200 |
| 4     | Malignant| CelebA | 72.55 | 400 | **2.84** | 2000 |
| 5     | Malignant| LSUN DOG | 89.68 | 600 | 2.25 | 600  |
| 6     | Malignant| FFHQ  | **68.78** | 3200 | 2.82 | 400  |
| 7     | Normal   | CelebA | 79.69 | 600 | 2.19 | 1200 |
| 8     | Normal   | LSUN DOG | 91.63 | 200 | 2.01 | 1400 |
| 9     | Normal   | FFHQ  | **73.92** | 2200 | **2.24** | 2400 |

Data availability statement

The base dataset can be accessed [here](#). The relevant codes, trained models, synthesized images, and other data will be available from the author upon reasonable request.

Acknowledgement

The authors thank Shuting Yu for her helpful discussion. This work was supported by the Singapore Ministry of Education Academic Research Fund Tier 1 Grants No. R265-000-691-114.

References

[1] Aimen Zlitni and Sanjiv S Gambhir. Molecular imaging agents for ultrasound. *Curr. Opin. Chem. Biol.*, 45:113–120, 2018. [https://doi.org/10.1016/j.cbpa.2018.03.017](https://doi.org/10.1016/j.cbpa.2018.03.017).

[2] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.*, 2018:1–13, 2018. [https://doi.org/10.1155/2018/7068349](https://doi.org/10.1155/2018/7068349).
Table C: Images synthesized by StyleGAN2-ADA in the early training stages and the final training stage for both training from scratch and using the TL technique. Only for illustrating the evolution of $G$. $\alpha$ denotes 3200 for training from scratch and 1000 for using the TL technique.

| Iterations | Training from scratch | Using the TL technique |
|------------|-----------------------|------------------------|
| 0          | ![Image](image1)      | ![Image](image2)       |
| 8          | Showing black holes   | Becoming darker        |
| 16         | Becoming darker       | Showing less original outlines |
| 32         | Containing some outlines | Showing no original outlines |
| 64         | Appearing more features | Appearing more details |
| 128        | Showing some details  | Containing more details |
| 192        | Having higher diversity | Showing more details |

$\alpha$
[3] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. In Proceedings of the IEEE, pages 2278–2324, 1998. https://doi.org/10.1109/5.726791

[4] S Deepak and PM Ameer. Brain tumor classification using deep cnn features via transfer learning. Comput. Biol. Med., 111:103345, 2019. https://doi.org/10.1016/j.compbiomed.2019.103345

[5] Shaohua Wan, Yan Liang, and Yin Zhang. Deep convolutional neural networks for diabetic retinopathy detection by image classification. Comput. Electr. Eng., 72:274–282, 2018. https://doi.org/10.1016/j.cpeleceng.2018.07.042

[6] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Deep learning approaches for data augmentation and classification of breast masses using ultrasound images. Int. J. Adv. Comput. Sci. Appl, 10:1–11, 2019. https://doi.org/10.14569/IJACSA.2019.0100579

[7] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv preprint, 2014.

[8] Ali Madani, Mehd Moradi, Alexandros Karargyris, and Tanveer Syeda-Mahmood. Semi-supervised learning with generative adversarial networks for chest x-ray classification with ability of data domain adaptation. In Proceedings of the IEEE International Symposium on Biomedical Imaging, pages 1038–1042, 2018. https://doi.org/10.1109/ISBI.2018.8363749

[9] Ting Pang, Jeannie Hsiu Ding Wong, Wei Lin Ng, and Chee Seng Chan. Semi-supervised gan-based radiomics model for data augmentation in breast ultrasound mass classification. Comput. Meth. Programs Biomed., 203:106018, 2021. https://doi.org/10.1016/j.cmpb.2021.106018

[10] Ibrar Amin, Saima Hassan, and Jafreezal Jaafar. Semi-supervised learning for limited medical data using generative adversarial network and transfer learning. In Proceedings of the International Conference on Computational Intelligence, pages 5–10. IEEE, 2020. https://doi.org/10.1109/ICCI51257.2020.9247724

[11] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. Data Brief, 28:104863, 2020. https://doi.org/10.1016/j.dib.2019.104863

[12] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, pages 1–14, 05 2015.

[13] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6848–6856, 2018. https://doi.org/10.1109/CVPR.2018.00716

[14] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1492–1500, 7 2017. https://doi.org/10.1109/CVPR.2017.634

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016. https://doi.org/10.1109/CVPR.2016.90

[16] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In
[17] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. https://doi.org/10.1109/CVPR.2016.308.

[18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017. https://doi.org/10.1109/cvpr.2017.243.

[19] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint*, 2016.

[20] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the International Conference on Machine Learning*, volume 70, pages 214–223, 2017.

[21] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 5769–5779, 2017.

[22] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 1–37, 2020.

[23] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint*, 2016.

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

[25] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. https://doi.org/10.1016/j.neucom.2018.09.013.

[26] Tim Salimans, Ian J Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 1–10, 2016.

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the Conference on Neural Information Processing Systems*, pages 1–38, 2017.

[28] Yeşim Eroğlu, Muhammed Yildirim, and Ahmet Çinar. Convolutional neural networks based classification of breast ultrasonography images by hybrid method with respect to benign, malignant, and normal using mrrm. *Comput. Biol. Med.*, 133:104407, 2021. https://doi.org/10.1016/j.compbiomed.2021.104407.

[29] Arijit Das and Srinibas Rana. Exploring residual networks for breast cancer detection from ultrasound images. In *Proceedings of the International Conference on Computing, Communication and Networking Technologies*, pages 1–6, 2021. https://doi.org/10.1109/ICCCNT51525.2021.9580160.
[30] Priyanka Khanna, Mridu Sahu, and Bikesh Kumar Singh. Improving the classification performance of breast ultrasound image using deep learning and optimization algorithm. In Proceedings of the IEEE International Conference on Technology, Research, and Innovation for Betterment of Society, pages 1–6. IEEE, 2021. https://doi.org/10.1109/TRIBES52498.2021.9751677

[31] Rakesh Chandra Joshi, Divyanshu Singh, Vaibhav Tiwari, and Malay Kishore Dutta. An efficient deep neural network based abnormality detection and multi-class breast tumor classification. Multimed. Tools Appl., 81(10):13691–13711, 2022. https://doi.org/10.1007/s11042-021-11240-0

[32] Behnaz Gheflati and Hassan Rivaz. Vision transformer for classification of breast ultrasound images. arXiv preprint, 2021.

[33] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Qingfeng Li, Yong Wang, and Shaojie Tang. Dk-consistency: A domain knowledge guided consistency regularization method for semi-supervised breast cancer diagnosis. In Proceedings of the International Conference on Bioinformatics and Biomedicine, pages 3435–3442, 2021. https://doi.org/10.1109/BIBM52615.2021.9669494

[34] Woo Kyung Moon, Yan-Wei Lee, Hao-Hsiang Ke, Su Hyun Lee, Chiun-Sheng Huang, and Ruey-Feng Chang. Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. Comput. Meth. Programs Biomed., 190:105361, 2020. https://doi.org/10.1016/j.cmpb.2020.105361

[35] Zhemin Zhuang, Zengbiao Yang, Alex Noel Joseph Raj, Chuliang Wei, Pengcheng Jin, and Shuxin Zhuang. Breast ultrasound tumor image classification using image decomposition and fusion based on adaptive multi-model spatial feature fusion. Comput. Meth. Programs Biomed., 208:106221, 2021. https://doi.org/10.1016/j.cmpb.2021.106221

[36] Tariq Sadad, Ayyaz Hussain, Asim Munir, Muhammad Habib, Sajid Ali Khan, Shariq Hussain, Shunkun Yang, and Mohammed Alawairdhi. Identification of breast malignancy by marker-controlled watershed transformation and hybrid feature set for healthcare. Appl. Sci., 10(6):1900, 2020. https://doi.org/10.3390/app10061900

[37] Arnab K Mishra, Pinki Roy, Sivaji Bandyopadhyay, and Sujit K Das. Breast ultrasound tumour classification: A machine learning—radiomics based approach. Expert Syst., 38(7):e12713, 2021. https://doi.org/10.1111/exsy.12713

[38] Michal Byra. Breast mass classification with transfer learning based on scaling of deep representations. Biomed. Signal Process. Control, 69:102828, 2021. https://doi.org/10.1016/j.bspc.2021.102828