Synthesis of Mammogram From Digital Breast Tomosynthesis Using Deep Convolutional Neural Network With Gradient Guided cGANs

Gongfa Jiang, Jun Wei, Yuesheng Xu, Zilong He, Hui Zeng, Jiefang Wu, Genggeng Qin, Weiguo Chen, and Yao Lu

***Abstract*— Synthetic digital mammography (SDM), a 2D image generated from digital breast tomosynthesis (DBT), is used as a potential substitute for full-field digital mammography (FFDM) in clinic to reduce the radiation dose for breast cancer screening. Previous studies exploited projection geometry and fused projection data and DBT volume, with different post-processing techniques applied on re-projection data which may generate different image appearance compared to FFDM. To alleviate this issue, one possible solution to generate an SDM image is using a learning-based method to model the transformation from the DBT volume to the FFDM image using current DBT/FFDM combo images. In this study, we proposed to use a deep convolutional neural network (DCNN) to learn the transformation to generate SDM using current DBT/FFDM combo images. Gradient guided conditional generative adversarial networks (GGGAN) objective function was designed to preserve subtle MCs and the perceptual loss was exploited to improve the performance of the proposed DCNN on perceptual quality. We used various image quality criteria for evaluation, including preserving masses and MCs which are important in mammogram. Experiment results demonstrated progressive performance improvement of network using different objective functions in terms of those image quality criteria. The methodology we exploited in the SDM generation task to analyze and progressively improve image quality by designing objective functions may be helpful to other image generation tasks.**

***Index Terms*— Breast cancer, digital breast tomosynthesis, image synthesis, deep learning, generative adversarial networks.**

I. Introduction

MAMMOGRAPHY screening is a cost-efficient tool for early detection of breast cancer. Full-field digital mammography (FFDM), a two-dimensional digital imaging technology, has been widely adapted for breast cancer screening. However, FFDM has limited sensitivity and specificity on dense breasts due to tissue overlapping, which can create false negatives when subtle lesions are obscured by complex fibroglandular tissues. Digital breast tomosynthesis (DBT), a new 3D imaging technology, is designed to address the tissue overlapping issue by reconstructing 3D structural information of the breast. Several large-scale clinical studies indicate that the false positive rate is lower and the cancer detection rate is higher for DBT + FFDM, compared to FFDM alone [1]. The advantage of the 2D view is that it is easier for radiologists to assess the density symmetry of two breasts, undertake comparison of priors, and assess the overall breast density, and it was reported that FFDM appears to be slightly more sensitive than digital breast tomosynthesis for the detection of calcification [2].

DBT + FFDM, the so-called "combo-mode," has gradually replaced mammography as an early screening tool in developed countries and can combine the advantages of DBT and FFDM. However, imaging with DBT + FFDM nearly doubles

Manuscript received November 8, 2020; revised February 19, 2021; accepted March 30, 2021. Date of publication April 7, 2021; date of current version July 30, 2021. This work was supported in part by the China Department of Science and Technology under Grant 2018YFC1704206 and Grant 2016YFB0200602; in part by the NSFC under Grant 81971691, Grant 81801809, Grant 81830052, Grant 81827802, Grant U1811461, and Grant 11401601; in part by the Science and Technology Program of Guangzhou under Grant 201804020053; in part by the Science and Technology Innovative Project of Guangdong Province under Grant 2016B030307003, Grant 2015B010110003, and Grant 2015B020233008; in part by the Science and Technology Program of Guangzhou under Grant 201906010014; in part by the Department of Science and Technology of Jilin Province under Grant 20190302108GX; in part by the Guangdong Provincial Science and Technology under Grant 2017B020210001; in part by the Guangzhou Science and Technology Creative under Grant 201604020003; in part by the Guangdong Province Key Laboratory of Computational Science Open under Grant 2018009; in part by the Construction Project of Shanghai Key Laboratory of Molecular Imaging under Grant 18DZ2260400; in part by the US National Science Foundation under Grant DMS-1912958; and in part by the Guangdong Province Key Laboratory of Computational Science at Sun Yat-sen University under Grant 2020B1212060032. *(Corresponding authors: Yao Lu; Weiguo Chen.)*

Gongfa Jiang is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China (e-mail: jianggfa@mail2.sysu.edu.cn).

Jun Wei is with Perception Vision Medical Technology Company Ltd., Guangzhou 510275, China, and also with the Department of Radiology, University of Michigan, Ann Arbor, MI 48109 USA (e-mail: jvwei@med.umich.edu).

Yuesheng Xu, retired, was with the Guangdong Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou 510275, China. He is now with the Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529 USA (e-mail: y1xu@odu.edu).

Zilong He, Hui Zeng, Jiefang Wu, Genggeng Qin, and Weiguo Chen are with the Department of Radiology, Nanfang Hospital, Southern Medical University, Guangzhou 510515, China (e-mail: long3415@smu.edu.cn; zh491157591@163.com; 1425221889@qq.com; zealotq@smu.edu.cn; chenweiguo1964@21cn.com).

Yao Lu is with the School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510275, China, and also with the Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, Guangzhou 510275, China (e-mail: luyao23@mail.sysu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2021.3071544.

Digital Object Identifier 10.1109/TMI.2021.3071544

the radiation dose compared to using FFDM alone [3]. More radiation exposure may increase the risk of radiation-induced breast cancer [4]. One of the solutions to reduce the radiation dose in DBT screening is to synthetically reconstruct a 2D FFDM-like mammogram, called synthetic digital mammography (SDM), from DBT data to replace the FFDM, thereby reducing the radiation dose while maintaining the detection performance of DBT to match that of DBT + FFDM [5].

There are several SDM methods in the literature. In [6], maximum intensity projection (MIP) and average projection approaches are used to generate thick slices from DBT. The approaches are only tested on phantom data, and the results are not comparable to FFDM. Study [7] used a computer-aided detection (CAD) system to determine relevant points in a DBT volume and then rendered a mammogram from the intersection of a surface fitted through these points. This method heavily relies on a CAD system that requires a large set of manually annotated data [8], and it may ignore suspicious microcalcification (MC) clusters and benign lesions which make it clinically inapplicable. Research [9] regarded the gradient value of the DBT volume as the degree of importance in terms of the conspicuity, and selected voxels with the highest conspicuity in DBT volume to derive SDM, while [10] used weighted averaging in which the weighting function was computed by an edge-detection filter to derive SDM from the DBT volume. Both methods, using only edge and gradient information in DBT, are very likely to miss textural abnormalities and unable to correctly assess the overall breast density. In [11], projection data as an adjunct to DBT slices were used to construct an SDM with enhanced MCs, while the conspicuity of masses on SDM was degraded.

There are several vendors with FDA approved SDM solutions: Hologic, Inc. (C-view and Intelligent 2D, Marlborough, MA, USA), Siemens (Insight, Munich, Germany), Fujifilm Medical Systems USA, Inc. (S-View, Stamford, CT, USA) and GE Healthcare (V-Preview, Chicago, IL, USA). The C-view image is created by re-projecting and filtering central projection data and/or the stack of reconstructed DBT slices, with calcification-like or lesion-like characteristics enhanced [12]. Clinical studies have shown that C-view + DBT have the performance similar to that of standard FFDM + DBT [13], [14]. Intelligent 2D was newly developed by Hologic to further improve the performance of C-view. However, the enhancement may result in false positives due to pseudocalcifications [15]. Besides, the C-view image provides poor overall resolution and noise properties compared to FFDM [16], [17]. More importantly, large-scale clinical studies reported that more breasts are categorized as nondense than dense when using C-view + DBT compared to using FFDM + DBT [18], [19], probably due to inherently different visual appearance between the C-view image and FFDM. Since breast density has both imaging and risk implications [20] and is an important component of mammography reports, the C-view image may result in an inconsistent mammography report to FFDM and unreliable risk assessment.

Most existing studies have proposed different post-processing techniques to enhance structures of interest, such as MCs, masses and important edges, and some CAD techniques have been applied to determine specific structure of interest to be enhanced before post-processing. For some research groups and companies which have access to projection data, projection data and DBT volume were fused along the projection geometry, and post-processing techniques were applied to re-projection data. Projection data played important role for SDM generation in our previous study [11]. For most research groups which have no access to projection data, one possible solution to generate an SDM image is using a learning-based method to model the transformation from the DBT volume to the FFDM image using current DBT/FFDM combo images. Since the FFDM image is the full-dose central projection image in the DBT data acquisition, the FFDM image is essentially identical to the re-projection of the reconstructed DBT volume, when ignoring artifacts and noise on the DBT volume (due to incomplete exposure, narrow angle, low-dose nature, and imperfect reconstruction in DBT acquisition) and post-processing steps used in FFDM acquisition (for optimization of image readability/appearance [21]). Besides, anti-scatter grid is used in FFDM acquisition but not in DBT acquisition. Thus, we consider the transformation from the DBT volume to the FFDM image as a four step process: (1) scatter correction; (2) artifacts reduction and denoising; (3) re-projection of processed DBT volume; (4) post-processing. Noting that DBT volume data used in this work are in cone-beam coordinates [22], every slice of reconstructed DBT is pixel-to-pixel mapped to the central projection image, *i.e.*, the FFDM image. That is, the re-projection process is a pixel-wise process and is spatially invariant, and it can be done by simply taking the average of DBT slices. Thus, the spatially varying re-projection geometry in DBT data has already been eliminated, and the learning-based method only needs to accommodate scatter correction, artifacts reduction, denoising and post-processing.

An end-to-end deep convolutional neural network (DCNN) was proposed to learn the transformation from the DBT volume to the FFDM image using DBT/FFDM combo images. Multi-scale information and 3D information in DBT images were integrated into the architecture of the DCNN. In this work, we mostly focused on progressively improving the performance of the given DCNN by using more complex objective functions and proposing new objective functions to approach similar appearance to FFDM and preserve local structures such as masses and MCs. The mean squared error (MSE) objective function is the baseline of our study and is feasible to minimize the difference of intensity distribution between SDM and FFDM and produce global similar appearance in vision. However, MSE may produce overly smoothed images [23], in which edges of local structures such as MCs and masses may be blurred and some MCs disappeared. To solve blurring issue arising from MSE, a state-of-the-art conditional generative adversarial networks (cGANs) [24], [25], termed pix2pixHD [26] was used to preserve edges, in which a discriminator network is trained to differentiate SDM from FFDM in the form of dissimilar patterns (such as edges) measured by an adversarial loss. However, experiment results showed that pix2pixHD was good at preserving edges of masses and

MC sharpness, but some small-scale structures such as subtle MCs still disappeared in SDM. On possible reason is that small-scale structures may have less weights than large-scale structures in the discriminator network of pix2pixHD, thus small-scale structures contribute less to the objective function. To alleviate this problem, we proposed a gradient guided cGANs (GGGAN) objective function to further enhance weak edges to preserve small-scale structures such as subtle MCs. In GGGAN, the gradient maps of SDM/FFDM images are extracted by the Sobel operators and used as additional input for the discriminator network, which force the discriminator to capture gradient features related to subtle edges. Thus, GGGAN with the gradient guided discriminator is feasible to generate images preserving subtle edges. However, GGGAN may over-enhance edges and result in signal distortion which makes the SDM images visually unpleasant for radiologists. We proposed to use perceptual loss [27] which correlates well with human perceptual judgments [28] to decrease signal distortion and improve perceptual quality of SDM images. In the perceptual loss, the dissimilarity between SDM and FFDM is measured in a feature space generated by a VGG-16 network [29] pre-trained on ImageNet [30] which is a very large natural image dataset. The perceptual loss was used in our study as an additional regularization term for GGGAN objective function, termed GGGAN-VGG.

In the experiments, we trained the DCNN with same architecture by using the four objective functions (*i.e.*, MSE, pix2pixHD, GGGAN, GGGAN-VGG) separately. Pairwise comparison was performed among those DCNNs to show that network performance can be progressively improved from using MSE to using GGGAN-VGG in terms of some important image quality criteria. Automatic mass segmentation and MC detection using in-house CAD software published in our previous studies on mammogram were used to evaluate the performance of preserving edges and local structures such as masses and MCs. The full-width-half-maximum (FWHM) was measured to quantify MC sharpness. A human observer study scoring radiologists' opinion on quality of image characteristics (*e.g.*, skin, glandular tissue, mass, MC) was conducted to evaluate the perceptual quality of images and consistency of breast density category.

This work is a further development based on our preliminary work [31]. The present work adds to the preliminary one in several ways. Firstly, we improve the GGGAN by introducing an additional perceptual loss. Secondly, extensive experiments, including the MC detection and mass segmentation on SDM, human observer study, and a visual comparison with a commercial SDM solution, are conducted. The statistical results provide additional support and insight for the proposed method. Moreover, we present more in-depth discussion and analysis on the proposed framework.

The rest of this paper is organized as follows. The details of network structure of the generator DCNN and the objective functions are described in Section II. The details of the dataset we used for algorithm development and evaluation are described in Section III. Quantitative experiments and human observer study on real patient data were conducted to evaluate the performance of DCNN with different objective functions, shown in Section IV. Finally, limitations of this work and relevant issues are discussed in Section V and a conclusion is drawn in Section VI.

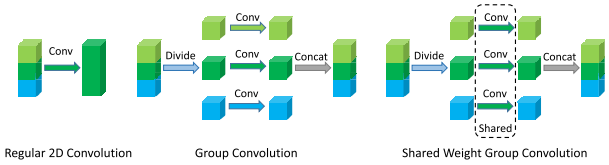


Fig. 1. Cubes with different colours indicate DBT slices/features of different groups. In regular 2D convolution, 3D information in input slices/channels cannot be kept. In group convolution and shared weight group convolution (SWGC), 3D information in input slices/channels can be kept. See Appendix F for visual result of comparison between SWGC, group convolution and regular 2D convolution Best viewed in colour.

II. Methods

In this section, we firstly give the network architecture of the proposed generator DCNN, which is identical for the different objective functions. Then we introduce the objective functions (*i.e.*, MSE, pix2pixHD, GGGAN, GGGAN-VGG) which were used to train the generator DCNN in detail.

A. Network Architecture of the Generator DCNN

We present two kinds of information in DBT that are critical to effectively approaching the target FFDM. Both information were carefully integrated into the network architecture of the generator DCNN.

1) Multi-Scale Information: Mammogram imaging has important multi-scale information. Coarse-scale structures like masses and glandular tissues as well as fine-scale structures like MCs and spiculations are all important components of the image. Thus, multi-scale representation is crucial in bridging DBT and the target FFDM. This motivates the use of a U-net [32] like network as the generator DCNN.

2) 3D Information: DBT input used in this work is noisy due to the incomplete exposure and low-dose nature of DBT acquisition. To help denoise, 3D context information in DBT is used in the generator DCNN. In this work, we use group convolution [33] to explicitly extract 3D structural information in DBT. In group convolution, input slices/channels are divided into several groups, and convolutions are applied separately to each group. Output features are concatenated in original group order to keep 3D information in input slices/channels (Fig. 1).

We proposed a shared weight group convolution (SWGC) structure with the assumption that the feature extraction process of individual groups should be identical. In SWGC, the weights of convolution kernels of individual groups are shared (Fig. 1). In this work, we empirically set 3 slices in a group, and 96 slices of DBT results in 32 groups. We have also tried group convolution with less or more slices in a group and found no significant difference. Thus, we empirically set the number of groups to be 32 in SWGC.

For the proposed generator DCNN, we use a U-net like network. The diagram of architecture is shown in Fig. 2. The DCNN is consist of an encoder which extracts multi-scale features from different scales of input, and a decoder which

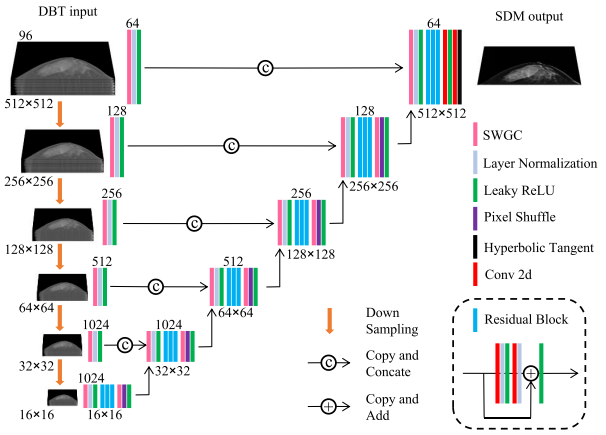


Fig. 2. The diagram of the generator DCNN architecture. In the left-hand side, DBT is used as 96 channels 2D input and downsampled by a factor of 2 for five times using average pooling layer. Multi-scale features with feature size 64, 128, 256, 512, 1024, 1024 are extracted from multi-scale DBT and fused to derive an SDM output. All convolution layers including SWGC have a kernel size of 3 and a stride of 1, except that the last 2D convolution layer has a kernel size of 7. Best viewed in colour.

derive an SDM by fusing multi-scale features. The multi-scale features are extracted from downsampled DBT using the proposed SWGC. We use layer normalization [34] instead of batch normalization, and pixel shuffle [35] layer instead of transpose convolution for upsampling. Besides, we add three residual blocks [36] in each scale of decoder to enlarge network capacity.

B. Mean Squared Error

We denote the training dataset by $\mathbf{S} = \{(\mathbf{x}_i, \mathbf{y}_i)|1 \le i \le M\}$, where $(\mathbf{x}_i, \mathbf{y}_i)$ is the ith pair of DBT and FFDM in $\mathbf{S}$, and M is the number of pairs in $\mathbf{S}$. To train the generator DCNN, denoted by G, the difference between each output $\hat{\mathbf{y}}_i = G(\mathbf{x}_i)$ and the target FFDM $\mathbf{y}_i$ is measured by an objective function and used to guide the optimization of G's parameters.

In the MSE objective function, the ground truth $\mathbf{y}$ is directly used as the regression target. MSE is the pixel-wise average of the difference between prediction $\hat{\mathbf{y}}$ and target $\mathbf{y}$. The MSE objective function is given by

$$L_{MSE}(G) = \frac{1}{M}\sum_{i=1}^{M} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2, \tag{1}$$

where $\|\cdot\|_2$ is the l_2-norm.

C. Pix2pixHD

In cGANs, to preserve edges and alleviate blurring issue arising from MSE, a discriminator network is trained to differentiate generated images from the ground truth in the form of dissimilar patterns (such as edges) measured by an adversarial loss, and the generator is trained to minimize the adversarial loss. In this work, we used a state-of-the-art cGANs, termed pix2pixHD [26]. In the pix2pixHD framework, a discriminator network D is trained to differentiate between prediction $\hat{\mathbf{y}}_i$ and

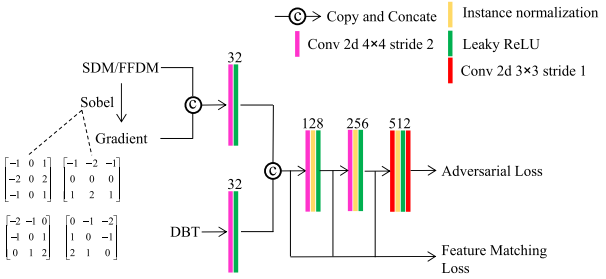


Fig. 3. The diagram of the discriminator of pix2pixHD/GGGAN, except that pix2pixHD don't have Sobel layer and gradient maps in discriminator. In the first layer, the features with a channel size of 32 are extracted from SDM/FFDM and DBT separately and concatenated. The output features of the next 4 convolution layers have a channel size of 128, 256, 512, 1, respectively. The first three convolution layers have a kernel size of 4 and a stride of 2, and the last two convolution layers have a kernel size of 3 and a stride of 1. Best viewed in colour.

the target $\mathbf{y}_i$ with input $\mathbf{x}_i$ as a condition. The loss function used to train D is given by

$$L_{Pix}(D) = \frac{1}{M}\sum_{i=1}^{M}\left(\|\mathbf{1} - D(\mathbf{x}_i, \mathbf{y}_i)\|_2^2 + \|\mathbf{0} - D(\mathbf{x}_i, \hat{\mathbf{y}}_i)\|_2^2\right), \tag{2}$$

where $\mathbf{1} = [1, 1, \ldots, 1]^T$ and $\mathbf{0} = [0, 0, \ldots, 0]^T$, both has the same size as $D(\cdot)$.

Given a trained discriminator D, the adversarial loss and feature matching loss for generator G are given respectively by

$$L_{Adver}(G) = \frac{1}{M}\sum_{i=1}^{M} \|\mathbf{1} - D(\mathbf{x}_i, \hat{\mathbf{y}}_i)\|_2^2, \tag{3}$$

$$L_{FM}(G) = \frac{1}{MT_D}\sum_{i=1}^{M}\sum_{j=1}^{T_D}\frac{1}{N_D^j}\left\|D^j(\mathbf{x}_i, \mathbf{y}_i) - D^j(\mathbf{x}_i, \hat{\mathbf{y}}_i)\right\|_1, \tag{4}$$

where $D^j(\cdot)$ is the the feature of jth layers of $D(\cdot)$, T_D is the total number of layers, N_D^j is the number of elements of the feature in the jth layer, and $\|\cdot\|_1$ is the l_1-norm. The pix2pixHD objective function used to train generator G is given by

$$L_{Pix}(G) = L_{Adver}(G) + \lambda_{FM}L_{FM}(G), \tag{5}$$

where λ_{FM} is a weighting factor to balance between adversarial loss and feature matching loss. The discriminator D and generator G are trained by minimizing (2) and (5) in an alternative way. For the hyperparameters of feature matching loss, we set the number of layers $T_D = 3$, $\lambda_{FM} = 10$, which are the same as that in the original pix2pixHD framework.

The diagram of network architecture of the discriminator of pix2pixHD is shown in Fig. 3 (Sobel layer and gradient maps are not included). The network architecture is the same as that in the original pix2pixHD framework [26] except the first layer is different. In the first layer, the features are extracted from SDM/FFDM and DBT separately, and are concatenated before being processed by the second convolution layer.

Similar to original pix2pixHD, we trained three discriminators with the identical architecture simultaneously to handle multi-scale structures in SDM/FFDM. For the first discriminator, it takes the original SDM/FFDM and DBT as inputs, while the other two discriminators have downsampled SDM/FFDM and DBT with a factor of 2 and 4 as input.

D. GGGAN

To further enhance weak edges to preserve small-scale structures such as subtle MCs, we proposed gradient guided cGANs (GGGAN) to capture gradient features related to subtle edges. Gradient type regularization has been studied in our previous studies and was proven to be useful to preserve subtle MCs in DBT [11], [37], [38]. In GGGAN, the input of the discriminator is augmented with the gradient maps of prediction image $\hat{\mathbf{y}}$ and target image $\mathbf{y}$, denoted by $\hat{\mathbf{y}}'$ and $\mathbf{y}'$ respectively. The gradient maps $\hat{\mathbf{y}}'$ and $\mathbf{y}'$ are concatenated with $\hat{\mathbf{y}}$ and $\mathbf{y}$ to form a multichannel image, denoted by $[\hat{\mathbf{y}}, \hat{\mathbf{y}}']$ and $[\mathbf{y}, \mathbf{y}']$, which together with DBT $\mathbf{x}$ form the new input for the discriminator. The loss function for the new discriminator D' is given by

$$L_{Grad}(D') = \frac{1}{M}\sum_{i=1}^{M}\Big(\big\|\mathbf{1} - D'(\mathbf{x}_i, [\mathbf{y}_i, \mathbf{y}_i'])\big\|_2^2 + \big\|\mathbf{0} - D'(\mathbf{x}_i, [\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i'])\big\|_2^2\Big). \quad (6)$$

Given a trained discriminator D', the adversarial loss and feature matching loss are given by

$$L_{Adver'}(G) = \frac{1}{M}\sum_{i=1}^{M}\big\|\mathbf{1} - D'(\mathbf{x}_i, [\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i'])\big\|_2^2, \quad (7)$$

$$L_{FM'}(G) = \frac{1}{MT_{D'}}\sum_{i=1}^{M}\sum_{j=1}^{T_{D'}}\frac{1}{N_{D'^j}}\Big\|D'^j(\mathbf{x}_i, [\mathbf{y}_i, \mathbf{y}_i']) - D'^j(\mathbf{x}_i, [\hat{\mathbf{y}}_i, \hat{\mathbf{y}}_i'])\Big\|_1. \quad (8)$$

The GGGAN objective function used to train generator G is given by

$$L_{GGGAN}(G) = L_{Adver'}(G) + \lambda_{FM}L_{FM'}(G). \quad (9)$$

The discriminator D' and generator G are trained by minimizing (6) and (9) in an alternative way.

The gradient maps are extracted from SDM/FFDM image by a fixed convolution layer in which kernels are initialized by Sobel operators. We use Sobel operators in four directions: vertical, horizontal, and two diagonal directions. To enlarge the receptive field of the Sobel convolution layer and capture bigger edges in the image, two additional dilated convolution [39] layers with the same Sobel kernels but having a dilation rate of 3 and 5 are used. Then the three gradient maps from three different convolution layers, in which each map has 4 channels (one channel for one direction), are concatenated with SDM/FFDM to be a multichannel image and are used as the input to the discriminator D'. The diagram of the discriminator of GGGAN is shown in Fig. 3.

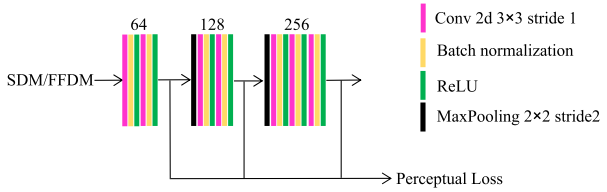


Fig. 4. The diagram of the first 7 convolution layers of VGG-16 network. The VGG-16 network have 13 convolution layers in total, while the rest layers are not used in the perceptual loss and are not shown in this diagram. The output features of the first 7 convolution layers have a channel size of 64, 64, 128, 128, 256, 256, 256, respectively. Maxpooling layers with a stride of 2 are used to downsample the features by a factor of 2. All convolution layers have a kernel size of 3 and a stride of 1. Best viewed in colour.

E. GGGAN-VGG

To improve perceptual quality of SDM images, we proposed to use perceptual loss as an additional regularization. In perceptual loss, the dissimilarity between prediction $\hat{\mathbf{y}}$ and target $\mathbf{y}$ is measured in feature space while the features are extracted by a pre-trained neural network. The commonly used pre-trained neural network is a VGG-16 network [29], pre-trained on ImageNet [30] which is a very large natural image dataset. Given the ImageNet pre-trained VGG-16 network, denoted by V, the perceptual loss is given by

$$L_V(G) = \frac{1}{MT_V}\sum_{i=1}^{M}\sum_{j=1}^{T_V}\frac{1}{N_V^j}\Big\|V^j(\mathbf{y}_i) - V^j(\hat{\mathbf{y}}_i)\Big\|_1, \quad (10)$$

where $V^j(\cdot)$ is the feature of jth layers of $V(\cdot)$, T_V is the total number of layers, and N_V^j is the number of elements of the feature in the jth layer. In this work, we empirically set the number of layers $T_V = 3$.

The diagram of the first 7 convolution layers of VGG-16 network is shown in Fig. 4. We only used the three lower layers (the 2nd, 4th, and 7th convolution layer) of the pre-trained VGG-16 network (which have 13 convolution layers in total) to extract features due to the assumption that natural image only shares low-level feature space with medical imaging.

In the literature, perceptual loss is usually used as an additional regularization term [27] to original objective function. In this work, we also combined perceptual loss (equation (10)) with GGGAN (equation (9)) to stabilize the training. The combined objective function, denoted by GGGAN-VGG in this work, is given by

$$L_{G-V}(G) = L_{Adver'}(G) + \lambda_{FM}L_{FM'}(G) + \lambda_V L_V(G), \quad (11)$$

where λ_V is a weighting factor to balance between adversarial loss and perceptual loss. We empirically set λ_V to 10 in this work.

III. Materials and Training Details

A. Materials

All the data used in this work were retrospectively collected from the Hologic Selenia system. For algorithm development, we collected 1077 cases each of which has malignant lesions (mass and/or MC) in one breast and is normal in

the other breast. We used 977 cases which have both DBT and FFDM for training. The rest 100 cases were reserved for validation. During the training, the number of training iterations is selected to ensure that training loss has converged and the visual appearance of two cases which were randomly selected from validation dataset have no significant change.

For testing, we independently collected 122 cases with mass and 21 cases with MC cluster. We used the 122 cases with mass (265 masses in all FFDM images, all have manually annotated masks) for intensity distortion evaluation, mass segmentation and human observer study. We used the 21 cases with MC cluster for test on MC cluster and MC morphology in human observer study. For MC detection and MC sharpness evaluation, since annotation is very time-consuming on individual MCs, a radiologist selected two cases with different breast density from the 21 cases with MC cluster, and both case has many MCs of various sizes. Then the radiologist manually annotated 128 ground truth MCs on the fatty breast case and 45 ground truth MCs on the dense breast case, and the annotation was reviewed by another radiologist. See Appendix C for the full images and the ROIs of the two selected cases.

B. Data Preprocessing

Before training and testing, since each DBT has different number of slices, all DBTs were padded with all zero slices on one side until each DBT has 96 slices. Similar to deep learning based image processing works in computer vision, the grey level range of both input DBTs (10-bit grey level, *i.e.*0 to 1023) and ground truth FFDMs (12-bit grey level, *i.e.*0 to 4095) were rescaled using linear transform to range from -1 to 1. Then for training data, the DBTs and corresponding FFDMs were cut into patches at 512×512 resolution without overlapping, and the patches with a high percentage of background were discarded for better training convergency. This results in a total of 33,431 patches in the training dataset. Noting that the generator DCNN is a fully convolutional network. Thus, the generator DCNN can use full-sized DBT as input and generates full-sized SDM in the test phase.

C. Training Details

During the training, the Adam [40] solver with a learning rate of 1×10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.9$ was used. The batch size was set to 1 due to the limitations of GPU memory. Horizontal flip augmentation was used for all patches, and vertical flip augmentation was used for patches from the CC-view mammogram. Four DCNNs with the same architecture were trained using four different objective functions, *i.e.*, MSE (equation (1)), pix2pixHD (equation (5)), GGGAN (equation (9)), and GGGAN-VGG (equation (11)) respectively. All DCNNs were trained for 6 epochs (approximately 200,000 iterations). The number of training iterations is large enough to ensure that training loss has converged and the visual appearances of images of the validation cases have no significant change. Training takes about 60 hours for MSE and about 80 hours for pix2pixHD, GGGAN and GGGAN-VGG on an NVIDIA TitanX GPU.

TABLE I
PSNR AND SSIM (MEAN $\pm$ STANDARD DEVIATION) OF MSE, PIX2PIXHD, GGGAN, AND GGGAN-VGG. THE RESULT OF RE-PROJECTION IS USED AS A REFERENCE

	PSNR	SSIM
MSE	$\mathbf{26.70 \pm 4.08}$	$\mathbf{0.708 \pm 0.160}$
Pix2pixHD	25.46 ± 3.98	0.622 ± 0.200
GGGAN	25.77 ± 4.05	0.629 ± 0.196
GGGAN-VGG	25.23 ± 4.16	0.635 ± 0.182
Re-projection	22.07 ± 3.33	0.633 ± 0.147

IV. EXPERIMENTS

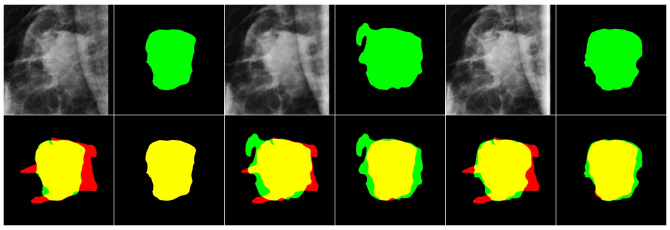
For brevity, we denote the four DCNNs, which have the same architecture and were trained using the four different objective functions (*i.e.*, equation (1), (5), (9), and (11)), by MSE, Pix2pixHD, GGGAN, and GGGAN-VGG respectively in below.

Firstly, we used peak-to-noise ratio (PSNR) and the structural similarity (SSIM) to measure intensity distortion. Then we compared pix2pixHD with MSE by using the mass segmentation task to measure ability to preserve mass edges, and the full-width-half-maximum (FWHM) to quantify MC sharpness. And we compared GGGAN with pix2pixHD by using the MC detection task to measure ability to preserve MCs. We compared GGGAN-VGG with GGGAN by conducting a human observer study to give human opinion scores on quality of characteristics in images (*e.g.*, skin, glandular tissue, mass, MC). In addition, we evaluated consistency of breast density category of different objective functions. For each pairwise comparison, we also provided the results of other objective functions as references in case the audiences are interested in those experiment results.

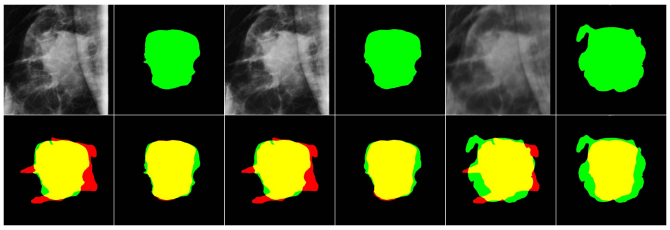
Since plain re-projection of DBT volume is the simplest method to derive an SDM image, we provided the result on re-projection images as a reference for each evaluation. The re-projection of DBT volume can be derived by simply taking the average of DBT slices because DBT volume data used in this work are in cone-beam coordinates [22]. After averaging, the re-projection image is rectified and rescaled using a window level of 540 and a window width of 580, which are provided by Hologic C-view software to make the image have a similar intensity distribution to FFDM. Then the re-projection image is rescaled from grey level of DBT (10-bit, *i.e.*, 0 to 1023) to grey level of FFDM (12-bit, *i.e.*, 0 to 4095). We denote the re-projection method by Re-projection in below.

A. MSE and Intensity Distortion

We used PSNR and SSIM to measure intensity distortion. The result is shown in Table I. MSE had a high PSNR and a high SSIM. While comparing with GAN based objective functions (*i.e.*, Pix2pixHD, GGGAN, and GGGAN-VGG), MSE is better at preserving intensity distribution.



(a) FFDM (b) **MSE** (c) **Pix2pixHD**



(d) GGGAN (e) GGGAN-VGG (f) Re-projection

Fig. 5. Representative visual results of mass segmentation on FFDM or SDMs derived from the five methods. For each result, four images are shown. Top left: FFDM or SDMs derived from the five methods. Top right: predicted mask of FFDM or SDMs. Bottom left: false negative (red), true positive (yellow) and false positive (green) regions of the predicted mask when taking the manually annotated mask as ground truth. Bottom right: false negative (red), true positive (yellow) and false positive (green) regions of the predicted mask when taking the predicted mask of FFDM as ground truth. Best viewed in colour.

B. Compare MSE With Pix2pixHD

1) Mass Segmentation: To show that Pix2pixHD can alleviate blurring issue arising from MSE and is better at preserving mass edges than MSE, we performed mass segmentation on the synthesized images to measure ability to preserve mass edges. In the mass segmentation task, a U-net was trained on an independently collected in-house dataset that includes 673 masses and corresponding manually annotated masks (see Appendix A for more details of U-net training). A representative visual result is shown in Fig. 5. In the representative visual result, the predicted mask of MSE had an irregular boundary and over-segmentation, while the predicted mask of Pix2pixHD was much more consistent with the predicted mask of FFDM.

The mean dice scores of predicted masks of FFDM and SDMs when taking the manually annotated mask as ground truth are shown in Table II. Besides, we proposed a semantic similarity score, which takes the predicted mask of FFDM as ground truth to calculate a dice score to directly evaluate the similarity of masses between SDMs and FFDM, also reported in Table II.

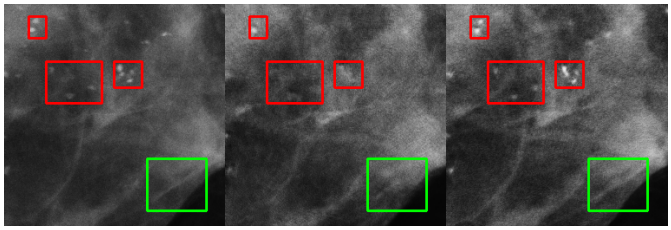
As can be seen, MSE and Pix2pixHD had similar dice score compared to FFDM ($p > 0.05$). However, considering the semantic similarity score, MSE was inferior to Pix2pixHD. The results indicate that Pix2pixHD is better at preserving mass edges than MSE.

2) MC Sharpness: To show that Pix2pixHD can alleviate blurring issue arising from MSE, we compared Pix2pixHD with MSE in terms of MC sharpness. To quantify sharpness of MCs, FWHM was used in this work. FWHM is the distance between points on the line profile at which the signal reaches half its maximum value. A lower FWHM indicates that the MC has a lower standard deviation and is sharper. Two MCs that appear in SDMs of all methods (pointed by the yellow arrows in Fig. 6 (g)) were selected to give a representative result of FWHM (see Appendix B).

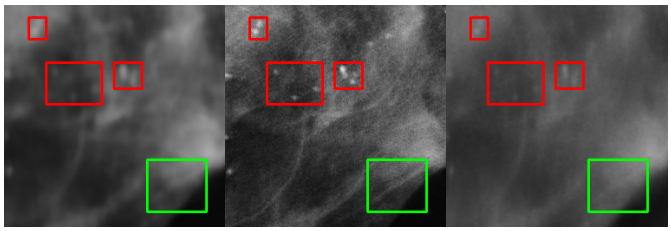
TABLE II

DICE SCORE (MEAN ± STANDARD DEVIATION) OF MSE AND PIX2PIXHD ON THE MASS SEGMENTATION TASK. SIGNIFICANCE (P-VALUE) OF DIFFERENCE BETWEEN THE RESULTS OF FFDM AND SDM ARE LISTED. SEMANTIC SIMILARITY IS THE DICE SCORE (MEAN ± STANDARD DEVIATION) BETWEEN PREDICTED MASK OF SDM AND PREDICTED MASK OF FFDM. THE RESULTS OF GGGAN, GGGAN-VGG, FFDM, AND RE-PROJECTION ARE USED AS REFERENCES

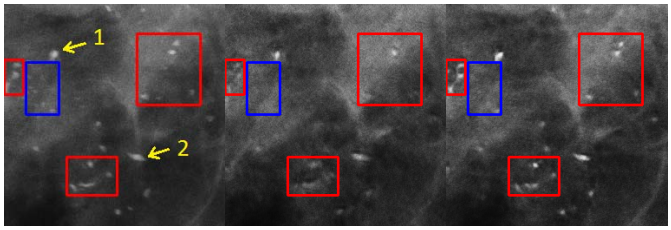
	Dice Score	p-value	Semantic Similarity
MSE	0.7554 ± 0.149	0.3904	0.8256 ± 0.148
Pix2pixHD	0.7639 ± 0.146	0.8429	$\mathbf{0.9074 \pm 0.084}$
GGGAN	0.7535 ± 0.148	0.3135	0.8906 ± 0.094
GGGAN-VGG	0.7621 ± 0.147	0.7347	0.9064 ± 0.082
FFDM	0.7664 ± 0.147		
Re-projection	0.6681 ± 0.198	$< 1 \times 10^{-9}$	0.6739 ± 0.224



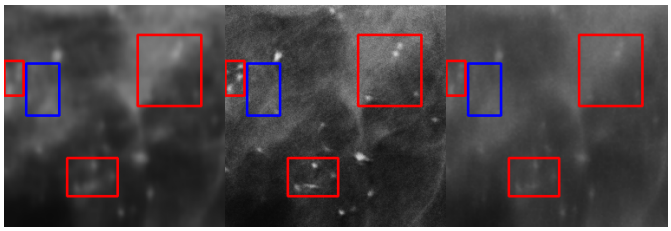
(a) FFDM (b) **Pix2pixHD** (c) **GGGAN**



(d) MSE (e) GGGAN-VGG (f) Re-projection



(g) FFDM (h) **Pix2pixHD** (i) **GGGAN**



(j) MSE (k) GGGAN-VGG (l) Re-projection

Fig. 6. Some visual results of MCs on FFDM or SDMs derived from the five methods. Rectangles outline sharp edges (green) and MCs (red) missed by Pix2pixHD while maintained in GGGAN. Noting that subtle calcifications in FFDM (blue rectangle) are missed by all methods (see Fig. 8 for more details). Best viewed on screen with zoom in.

TABLE III
RATIO (MEAN ± STANDARD DEVIATION) OF STANDARD DEVIATION OF SDM FROM MSE AND PIX2PIXHD, TO STANDARD DEVIATION OF FFDM. P-VALUE OF T-TEST WITH THE NULL HYPOTHESIS THAT THE MEAN OF RATIOS IS EQUAL TO 1. THE RESULTS OF GGGAN, GGGAN-VGG, AND RE-PROJECTION ARE USED AS REFERENCES

	Ratio	p-value
MSE	2.074 ± 1.244	$< 1 \times 10^{-9}$
Pix2pixHD	**1.026±0.537**	0.6455
GGGAN	1.087 ± 0.561	0.0599
GGGAN-VGG	1.1 ± 0.635	0.0436
Re-projection	1.644 ± 0.823	$< 1 \times 10^{-10}$

We statistically compared the MC sharpness of SDM images from MSE and Pix2pixHD on the 173 ground truth MCs. Given an SDM and an MC of gold standard, the line profiles of the MC on FFDM and the SDM of both the x-direction and y-direction were fitted by Gaussian curves. Since the FWHM of Gaussian curve is proportional to the standard deviation of the Gaussian curve, we measured the sharpness of the MC of gold standard on FFDM and the SDM by the standard deviation of Gaussian curve. Thus, the ratio of standard deviation of the SDM to standard deviation of FFDM should be close to 1 if the MC of the SDM has a similar sharpness to the same MC of FFDM. Since the MC sharpness evaluation can only be performed on the MCs present on both FFDM image and SDM image, we selected ground truth MCs that can be detected on both FFDM image and SDM image by using the same MC detection method as in the MC detection experiment. The statistical results of the ratio of MSE and Pix2pixHD are listed in Table III. The results of GGGAN, GGGAN-VGG, and Re-projection are also listed and used as references.

As can be seen, MSE had much flatter MCs compared to FFDM, while Pix2pixHD had comparable MC sharpness to FFDM. The results indicate that Pix2pixHD can alleviate blurring issue arising from MSE and preserve MC sharpness.

C. Compare Pix2pixHD With GGGAN

To show that GGGAN is better at preserving MCs than Pix2pixHD, we compared GGGAN with Pix2pixHD by using the MC detection task to measure ability to preserve MCs. We performed MC detection on SDM images and FFDM image by using a rule-based MC candidate detection method, which was used in the MC cluster CAD system [41], [42] and our previous work [43]. For each detected MC candidate in FFDM and SDMs, if there exists an MC of gold standard within less than 5 pixels (the radius of the largest MC), it is considered as a true positive and denoted by TP_{detect}. Otherwise, it is considered as a false positive and denoted by FP_{detect}. Then, the precision is calculated as follows:

$$Precision = \frac{TP_{detect}}{TP_{detect} + FP_{detect}}. \quad (12)$$

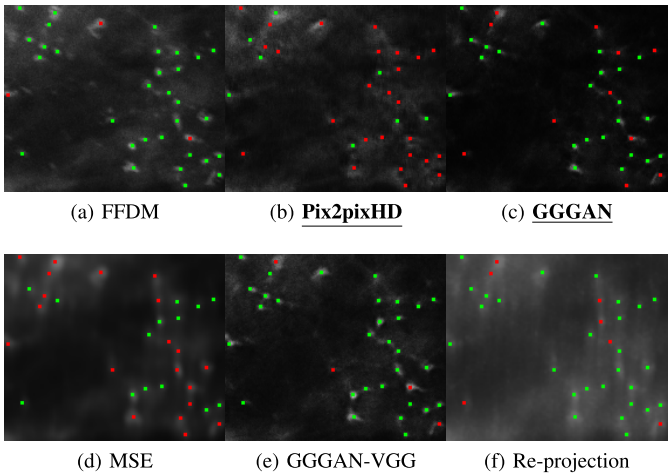
Given an SDM or the FFDM, for each ground truth MC, if there exists an MC candidate in SDM or FFDM within 5 pixels, it is considered as a true positive and denoted by TP_{gt}. Otherwise, it is considered as a false negative and denoted by FN_{gt}. Then, the sensitivity is calculated as follows:

$$Sensitivity = \frac{TP_{gt}}{TP_{gt} + FN_{gt}}. \quad (13)$$

See Fig. 6 for some visual results of preserving MCs and Fig. 7 for some candidate detection results. ROIs in Fig. 6 highlight the MCs and sharp edges that are missed by Pix2pixHD while maintained by GGGAN. Noting that GGGAN and GGGAN-VGG failed to maintain some subtle low-contrast calcifications as shown in Fig. 6 (i) and (k). This is probably due to the low-dose nature of DBT acquisition which makes those subtle low-contrast calcifications too weak in DBT slices (Fig. 8) to be distinguished from noise and recovered from DBT in our method.

Table IV lists the precision and sensitivity of SDMs from Pix2pixHD and GGGAN. The results of FFDM and SDMs from MSE, GGGAN-VGG, and Re-projection are also listed as references. GGGAN significantly outperformed ($p < 0.0001$) pix2pixHD, and had a lower precision ($p = 0.241$) and a lower sensitivity ($p = 0.0017$) compared to FFDM, while Pix2pixHD had a much lower precision ($p < 0.0001$) and a much lower sensitivity ($p < 0.0001$) compared to FFDM, which indicates that the proposed GGGAN objective function is better at preserving MCs than Pix2pixHD.

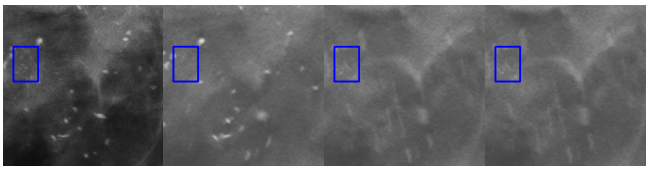
We found that GGGAN had similar performance on fatty breast case to that on dense breast case, and substantially outperformed Pix2pixHD on both case. While Re-projection



(a) FFDM (b) **Pix2pixHD** (c) **GGGAN**

(d) MSE (e) GGGAN-VGG (f) Re-projection

Fig. 7. A part of candidate detection results. Green points indicate true positive TP_{gt} and red points indicate false negative FN_{gt} for FFDM and each SDM. Best viewed in colour.



(a) FFDM (b) 27th slice (c) 35th slice (d) 36th slice

Fig. 8. FFDM and slices of corresponding DBT which contain subtle calcifications missed by GGGAN and GGGAN-VGG shown in Fig. 6. Those subtle calcifications appear on the 35th and 36th slice of DBT.

TABLE IV
PRECISION AND SENSITIVITY OF MC CANDIDATE DETECTION ON SDMS FROM PIX2PIXHD AND GGGAN. THE RESULTS OF FFDM AND SDMS FROM MSE, GGGAN-VGG, AND RE-PROJECTION ARE USED AS REFERENCES

	Precision	Sensitivity
Pix2pixHD	27.67%	44.51%
GGGAN	**44.77%**	**68.21%**
MSE	30.28%	38.15%
GGGAN-VGG	46.83%	75.14%
FFDM	50%	83.24%
Re-projection	43.21%	67.63%

TABLE V
SUBJECTIVE SCORES (MEAN ± STANDARD DEVIATION) OF NORMAL TISSUE QUALITY, MASS CONSPICUITY, DISTRIBUTION CONSISTENCY, AND MORPHOLOGY CONSISTENCY FOR GGGAN AND GGGAN-VGG. THE RESULTS OF MSE, PIX2PIXHD, FFDM, AND RE-PROJECTION ARE USED AS REFERENCES

	Normal	Conspicuity	Distribution	Morphology
GGGAN	3.22 ± 0.31	3.13 ± 0.37	3.46 ± 0.47	2.86 ± 0.42
GGGAN-VGG	$\mathbf{3.61 \pm 0.37}$	$\mathbf{3.5 \pm 0.47}$	3.51 ± 0.53	$\mathbf{3.21 \pm 0.5}$
MSE	1.49 ± 0.2	1.56 ± 0.22	1.25 ± 0.29	1.03 ± 0.1
Pix2pixHD	3.01 ± 0.32	2.91 ± 0.39	2.29 ± 0.37	1.94 ± 0.37
FFDM	4.79 ± 0.23	4.5 ± 0.39	-	-
Re-projection	2.61 ± 0.24	2.57 ± 0.44	2.41 ± 0.65	1.81 ± 0.51

had worse performance on dense breast case than that on fatty breast case, probably because in the fatty breast case, low density makes it easier for the plain re-projection method to maintain those high-contrast MCs.

D. Compare GGGAN With GGGAN-VGG

To show that GGGAN-VGG can improve perceptual quality compared to GGGAN, we conducted a human observer study to give human opinion scores on quality of characteristics in images (*e.g.*, skin, glandular tissue, mass, MC) of GGGAN-VGG and GGGAN. The results of MSE, Pix2pixHD, FFDM, and Re-projection are also reported as references. In addition, we evaluated consistency of breast density category to the show the dependence of breast density category on objective functions.

1) Normal Tissue: We performed a blind reader study on 122 independently collected cases reported with mass in terms of normal tissue quality (including quality of glandular tissue, fatty tissue, lymph node, skin line, nipple, and vessel). For each case, SDMs derived from the five methods (*i.e.*, MSE, Pix2pixHD, GGGAN, GGGAN-VGG, Re-projection) and FFDM were shown to three radiologists in a random order. The MLO-view image and CC-view image of two sides were shown at the same time and the radiologists were free to zoom-in or zoom-out and change window-width and window-level. The three radiologists were asked to independently score each modality on a five-point scale (5=excellent, 1=unacceptable). The three scores were averaged and assigned to the modality on that case. The mean and standard deviation values of the normal tissue quality scores are shown in Table V. It can be seen that GGGAN-VGG had a higher quality on normal tissue, skin line, nipple and vessel than GGGAN ($p < 0.05$).

2) Mass Conspicuity: The blind reader study in terms of mass conspicuity was conducted on the same data and the same study protocol as that of evaluating normal tissue quality (except that for scoring, 5=highly conspicuous and 1=highly inconspicuous). The mean and standard deviation values of the mass conspicuity scores are shown in Table V. It can be seen that GGGAN-VGG had a higher mass conspicuity than GGGAN ($p < 0.05$).

3) MC Cluster and MC Morphology: In addition, we independently collected 21 cases with MC cluster. For each case, SDMs derived from the five methods were shown to the three radiologists in a random order. The three radiologists were asked to independently score each modality in terms of consistency of distributional characteristic of MC cluster and consistency of morphology of MCs when taking the MC cluster of FFDM as reference on a five-point scale (5=exactly the same, 1=totally different). The three scores are averaged and assigned to the modality on that case. The mean and standard deviation values of the consistency scores are shown in Table V. It can be seen that GGGAN-VGG is better at preserving MC morphology than GGGAN ($p < 0.05$).

4) Breast Density Category: The reader study in terms of breast density category was conducted on the same data and the same study protocol as that of evaluating normal tissue quality. The three radiologists were asked to assign breast density category to each modality according to the 5th edition of BI-RADS. If at least two of the three readers agreed on a density category, the density category is assigned to the modality. If the three readers assigned three different density categories, the median density category is assigned to the modality. The consensus percent agreement and Cohen kappa coefficient between BI-RADS density categories assigned for FFDM and SDM derived from each method are listed in Table VI. In addition to four-category scale, the percent agreement and Cohen kappa coefficient calculated based on two-category scale, in which heterogeneous and extremely dense are combined into a dense category and fatty and scattered breast densities are combined into a nondense category, are also listed in Table VI. It can be seen that all three GAN based methods achieve higher agreement with FFDM on breast density assessment than MSE on both four-category scale and two-category scale. It may be because GAN based methods are

TABLE VI
PERCENT AGREEMENT AND COHEN KAPPA COEFFICIENT OF BI-RADS DENSITY CATEGORY FOR THE FOUR METHODS. THE RESULT OF RE-PROJECTION IS USED AS A REFERENCE

	Four Category		Two Category	
	Percent	Kappa	Percent	Kappa
MSE	82.8%	0.622	87.7%	0.655
Pix2pixHD	94.3%	0.881	95.9%	0.895
GGGAN	95.9%	0.912	96.7%	0.917
GGGAN-VGG	95.1%	0.898	95.9%	0.897
Re-projection	81.1%	0.623	89.3%	0.723

better at preserving glandular tissues and can provide a more accurate glandular tissue percentage assessment, *i.e.*, breast density category.

In comparison with FFDM in terms of breast density categories, C-view was reported to have a percent agreement of 80.3% and a Cohen kappa coefficient of 0.73 on four-category scale, and have a percent agreement of 91.9% and a Cohen kappa coefficient of 0.83 on two-category scale [44], which are lower than the results of GGGAN in our human observer study. A direct comparison between GGGAN and C-view in terms of breast density consistency is needed in future study.

V. Discussion

In this work, to achieve better image quality for SDM images derived from the given DCNN network, we mainly focused on using more complex objective functions to train the DCNN network without changing the network architecture. We used pix2pixHD to alleviate blurring issue arising from MSE and preserve edges. In the experiment of mass segmentation, pix2pixHD had an average semantic similarity of 0.9074, while MSE had an average semantic similarity of 0.8256. In MC sharpness evaluation, Pix2pixHD had comparable MC sharpness to FFDM, while MSE had much flatter MCs compared to FFDM. Pix2pixHD can alleviate blurring issue arising from MSE and is better at preserving edges than MSE. Then we proposed a GGGAN objective function to preserve MCs lost by pix2pixHD. In MC detection, GGGAN had a precision of 44.77% and a sensitivity of 68.21%, while pix2pixHD had a precision of 27.67% and a sensitivity of 44.51%. The proposed GGGAN objective function is better at preserving MCs than pix2pixHD. Then we combined GGGAN with perceptual loss (*i.e.*, GGGAN-VGG) to improve perceptual quality. In human observer study, GGGAN-VGG had average human opinion scores of 3.61, 3.5, 3.51 and 3.21 in terms of normal tissue quality, mass conspicuity, MC cluster distribution consistency, and MC morphology respectively, while GGGAN had average scores of 3.22, 3.13, 3.46, and 2.86. Radiologists prefer images derived from GGGAN-VGG than GGGAN. GGGAN-VGG has higher perceptual quality than GGGAN. Besides, pix2pixHD, GGGAN, and GGGAN-VGG achieved higher agreement with FFDM on breast density assessment than MSE. The experiment results showed that the performance of the DCNN is progressively improved in terms of various image criteria, especially preserving masses and MCs which are important in mammogram, from using MSE to using GGGAN-VGG.

One major limitation in this work is that reader detection study was not performed. Besides, we used in-house CAD software published in our previous studies on mammogram for mass segmentation and MC detection. Since we have no access to other business CAD software, wide investigation on various CAD software for evaluation was not conducted in this study. In MC detection experiment, since annotation is very time-consuming on individual MCs, we only reported two typical cases. Another major limitation is that we focused our evaluation on four objective functions which motivated our study to achieve clinical requirements on global perceptual quality and local structures. The performance of the DCNN using other objective functions, such as other cGANs based objective functions, is unknown. Besides, in this work, we kept the architecture of the DCNN network fixed. The dependence of network performance on other network architectures and whether the performance gain of using more complex objective function exists for DCNN network with higher/less complexity are unknown.

Another major limitation in this work is that statistical results of comparisons between the proposed method and C-view/Intelligent 2D, such as a comparison in terms of breast density consistency, were not provided. Since both C-view and Intelligent 2D were not approved by FDA of China, we could only collect a limited number of images from one hospital which conducted clinical trial for Hologic C-view, and we can only provide a visual comparison between the proposed GGGAN-VGG method and C-view on a few representative images (see Appendix D). The findings in the visual comparison are preliminary and the conclusions lack significance due to insufficient available cases. However, we do think it is valuable to provide this result in order to make audiences have enough confidence and interests to try our method if they have enough data. We will further quantify the performance of the proposed method compared with commercial SDM solutions when we collect sufficient data in the future.

Some studies reported that visualization of small MCs using C-view is challenging [45]. However, we did not evaluate the detection as a function of MC size in this work. We think that evaluating MC detection as a function of MC size is an interesting topic and requires extensive experiments on additional phantom data or manual annotation which are beyond the scope of this work. We would like to leave it for future study.

There are several directions to improve the performance of the proposed learning-based method. To decrease intensity distortion in GGGAN-VGG, combining MSE with GGGAN-VGG may achieve a good balance between intensity distortion and other image quality criteria. It was shown in this work that using different objective functions to train the DCNN can preserve different characteristics in images. Thus, it is possible to design a new objective function to further enhance specific characteristics in images. For example, a network which has been trained on mass/MC detection task can be used to extract features to compute the loss between predicted image and ground truth (similar to the perceptual loss used in this work). By using this objective function, mass/MC in images may be further enhanced and the SDM image may be similar to C-view image.

Using higher quality data or additional data may also help to improve the performance of the proposed learning-based method, for example, DBT data acquired from Hologic Clarity HD system (newly developed by Hologic to acquire high resolution DBT images), wide-angle scan system (such as Siemens which provides DBT with less artifacts), and other DBT systems using anti-scatter grids (having less artifacts). Our algorithm should be re-trained on those high quality DBT data since those DBT data may have different image

characteristics compared to DBT from current system. Using those high quality DBT data as input may make the training easier and may recover subtle low-contrast MCs (such as MCs shown in Fig. 8). On the other hand, data are cheaper in our task than that in medical image detection/diagnosis task since we don't need manually annotated label to train the network. With much more training data available, it is possible to have a better performance without changing the framework.

In this work, the trained generator DCNN can only be used for data acquired from Hologic system since it only learned the transformation from DBT volume to FFDM image of Hologic system. To investigate the potential capacity of the proposed method to transfer to an unseen machine system, we tested the DCNN trained by GGGAN-VGG on a DBT volume sample acquired from GE system (see Appendix E). However, the transfer method proposed in the test on GE system is preliminary, and the findings lack significance due to insufficient cases being tested. More works are needed in the future to further quantify the cross-vendor potential of the proposed method.

VI. Conclusion

In this study, we proposed to use a DCNN to learn the transformation from the DBT volume to the FFDM image to generate SDM. We proposed a GGGAN objective function and used the perceptual loss to improve the performance of the proposed DCNN. We used various image quality criteria for evaluation, including preserving masses and MCs which are important in mammogram. We observed progressive performance improvement of network with different objective functions in terms of those image quality criteria. To our best knowledge, this is the first work of learning-based SDM method in the literature. In the future, we will conduct statistical experiments to compare the proposed method with commercial SDM solutions and conduct clinical study to further quantify the potential of the proposed method.

References

[1] R. Hodgson *et al.*, "Systematic review of 3D mammography for breast cancer screening," *Breast*, vol. 27, pp. 52–61, Jun. 2016.

[2] M. L. Spangler *et al.*, "Detection and classification of calcifications on digital breast tomosynthesis and 2D digital mammography: A comparison," *Amer. J. Roentgenol.*, vol. 196, no. 2, pp. 320–324, Feb. 2011.

[3] T. M. Svahn, N. Houssami, I. Sechopoulos, and S. Mattsson, "Review of radiation dose estimates in digital breast tomosynthesis relative to those in two-view full-field digital mammography," *Breast*, vol. 24, no. 2, pp. 93–99, Apr. 2015.

[4] D. L. Miglioretti *et al.*, "Radiation-induced breast cancer incidence and mortality from digital mammography screening: A modeling study," *Ann. Internal Med.*, vol. 164, no. 4, pp. 205–214, Jan. 2016.

[5] S. P. Zuckerman *et al.*, "Implementation of synthesized two-dimensional mammography in a population-based digital breast tomosynthesis screening program," *Radiology*, vol. 281, no. 3, pp. 730–736, 2016.

[6] F. Diekmann *et al.*, "Thick slices from tomosynthesis data sets: Phantom study for the evaluation of different algorithms," *J. Digit. Imag.*, vol. 22, no. 5, p. 519, 2009.

[7] G. van Schie, R. Mann, M. Imhof-Tas, and N. Karssemeijer, "Generating synthetic mammograms from reconstructed tomosynthesis volumes," *IEEE Trans. Med. Imag.*, vol. 32, no. 12, pp. 2322–2331, Dec. 2013.

[8] G. van Schie, M. G. Wallis, K. Leifland, M. Danielsson, and N. Karssemeijer, "Mass detection in reconstructed digital breast tomosynthesis volumes with a computer-aided detection system trained on 2D mammograms," *Med. Phys.*, vol. 40, no. 4, Mar. 2013, Art. no. 041902.

[9] S. T. Kim, D. H. Kim, and Y. M. Ro, "Generation of conspicuity-improved synthetic image from digital breast tomosynthesis," in *Proc. 19th Int. Conf. Digit. Signal Process. (DSP)*, Aug. 2014, pp. 395–399.

[10] H. Homann, F. Bergner, and K. Erhard, "Computation of synthetic mammograms with an edge-weighting algorithm," *Proc. SPIE*, vol. 9412, Mar. 2015, Art. no. 94121Q.

[11] J. Wei *et al.*, "Synthesizing mammogram from digital breast tomosynthesis," *Phys. Med. Biol.*, vol. 64, no. 4, Feb. 2019, Art. no. 045011.

[12] C. Ruth, A. Smith, and J. Stein, "System and method for generating a 2D image from a tomosynthesis data set," U.S. Patent 7 760 924, Jul. 20 2010.

[13] P. Skaane *et al.*, "Two-view digital breast tomosynthesis screening with synthetically reconstructed projection images: Comparison with digital breast tomosynthesis with full-field digital mammographic images," *Radiology*, vol. 271, no. 3, pp. 655–663, Jun. 2014.

[14] F. J. Gilbert *et al.*, "Accuracy of digital breast tomosynthesis for depicting breast cancer subgroups in a UK retrospective reading study (TOMMY Trial)," *Radiology*, vol. 277, no. 3, pp. 697–706, Dec. 2015.

[15] L. Ratanaprasatporn, S. A. Chikarmane, and C. S. Giess, "Strengths and weaknesses of synthetic mammography in screening," *RadioGraphics*, vol. 37, no. 7, pp. 1913–1927, Nov. 2017.

[16] J. S. Nelson, J. R. Wells, J. A. Baker, and E. Samei, "How doesc-viewimage quality compare with conventional 2D FFDM?" *Med. Phys.*, vol. 43, no. 5, pp. 2538–2547, Apr. 2016.

[17] P. Barca, R. Lamastra, G. Aringhieri, R. M. Tucciariello, A. Traino, and M. E. Fantacci, "Comprehensive assessment of image quality in synthetic and digital mammography: A quantitative comparison," *Australas. Phys. Eng. Sci. Med.*, vol. 42, no. 4, pp. 1141–1152, Dec. 2019.

[18] M. P. Aujero, S. C. Gavenonis, R. Benjamin, Z. Zhang, and J. S. Holt, "Clinical performance of synthesized two-dimensional mammography combined with tomosynthesis in a large screening population," *Radiology*, vol. 283, no. 1, pp. 70–76, Apr. 2017.

[19] A. Gastounioti, A. M. McCarthy, L. Pantalone, M. Synnestvedt, D. Kontos, and E. F. Conant, "Effect of mammographic screening modality on breast density assessment: Digital mammography versus digital breast tomosynthesis," *Radiology*, vol. 291, no. 2, pp. 320–327, 2019.

[20] S. V. Destounis, A. Santacroce, and A. Arieno, "Update on breast density, risk estimation, and supplemental screening," *Amer. J. Roentgenol.*, vol. 214, no. 2, pp. 296–305, 2020.

[21] A. P. Smith, B. Chen, and Z. Jing, "Mammography/tomosynthesis systems and methods automatically deriving breast characteristics from breast X-ray images and automatically adjusting image processing parameters accordingly," U.S. Patent 8 170 320, May 1 2012.

[22] B. Ren *et al.*, "A new generation FFDM/tomosynthesis fusion system with selenium detector," *Proc. SPIE*, vol. 7622, Mar. 2010, Art. no. 76220B.

[23] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[24] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[25] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[26] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," 2017, *arXiv:1711.11585*. [Online]. Available: http://arxiv.org/abs/1711.11585

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2016, pp. 694–711.

[28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 586–595.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.

[31] G. Jiang, Y. Lu, J. Wei, and Y. Xu, "Synthesize mammogram from digital breast tomosynthesis with gradient guided cGANs," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent (MICCAI)*. Cham, Switzerland: Springer, 2019, pp. 801–809.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent (MICCAI)*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[33] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep roots: Improving CNN efficiency with hierarchical filter groups," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1231–1240.

[34] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: http://arxiv.org/abs/1607.06450

[35] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.

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

[37] Y. Lu, H.-P. Chan, J. Wei, and L. M. Hadjiiski, "Selective-diffusion regularization for enhancement of microcalcifications in digital breast tomosynthesis reconstruction," *Med. Phys.*, vol. 37, no. 11, pp. 6003–6014, Oct. 2010.

[38] Y. Lu, H.-P. Chan, J. Wei, L. M. Hadjiiski, and R. K. Samala, "Multiscale bilateral filtering for improving image quality in digital breast tomosynthesis," *Med. Phys.*, vol. 42, no. 1, pp. 182–195, Dec. 2014.

[39] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[41] H.-P. Chan, S.-C.-B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," *Med. Phys.*, vol. 22, no. 10, pp. 1555–1567, Oct. 1995.

[42] J. Ge *et al.*, "Computer aided detection of clusters of microcalcifications on full field digital mammograms," *Med. Phys.*, vol. 33, no. 8, pp. 2975–2988, Jul. 2006.

[43] G. Cai, Y. Guo, W. Chen, H. Zeng, Y. Zhou, and Y. Lu, "Computer-aided detection and diagnosis of microcalcification clusters on full field digital mammograms based on deep learning method using neutrosophic boosting," *Multimedia Tools Appl.*, pp. 79, pp. 17147–17167, May 2019.

[44] T. I. Alshafeiy *et al.*, "Comparison between digital and synthetic 2D mammograms in breast density interpretation," *Amer. J. Roentgenol.*, vol. 209, no. 1, pp. W36–W41, Jul. 2017.

[45] L. C. Ikejimba *et al.*, "Assessment of task-based performance from five clinical DBT systems using an anthropomorphic breast phantom," *Med. Phys.*, vol. 48, no. 3, pp. 1026–1038, 2020.