arXiv:1807.03401v1 [cs.CV] 9 Jul 2018

High-Resolution Mammogram Synthesis using Progressive Generative Adversarial Networks

Dimitrios Korkinof[*†], **Tobias Rijken**[†], **Michael O'Neill**[†], **Joseph Yearsley**[†], **Hugh Harvey**[†], **and Ben Glocker**[†,§]

[†]Kheiron Medical Technologies Ltd.
[§]Department of Computing, Imperial College London

Abstract

The ability to generate synthetic medical images is useful for data augmentation, domain transfer, and out-of-distribution detection. However, generating realistic, high-resolution medical images is challenging, particularly for Full Field Digital Mammograms (FFDM), due to the textural heterogeneity, fine structural details and specific tissue properties. In this paper, we explore the use of progressively trained generative adversarial networks (GANs) to synthesize mammograms, overcoming the underlying instabilities when training such adversarial models. This work is the first to show that generation of realistic synthetic medical images is feasible at up to 1280x1024 pixels, the highest resolution achieved for medical image synthesis, enabling visualizations within standard mammographic hanging protocols. We hope this work can serve as a useful guide and facilitate further research on GANs in the medical imaging domain.

1 Introduction

The generation of synthetic medical images is of increasing interest to both image analysis and machine learning communities for several reasons. First, synthetic images can be used to improve methods for downstream detection and classification tasks, by generation of images from a particularly sparse class, or by transforming existing images in a plausible way to generate more diverse datasets (known as data augmentation). Salehinejad et al. (2018) and Frid-Adar et al. (2018) show the benefits of this approach as applied to chest X-ray and liver lesion classification, respectively. Costa et al. (2017) successfully use generative adversarial networks (GANs) in an image-to-image translation setting to learn a mapping from binary vessel trees to realistic retinal images.

Second, GANs can be used in domain adaptation, in which a model trained on images of one domain is applied to images of another domain where labels are scarce or non-existent. Images across related modalities can have significantly different visual appearance, such as in the cases of CT and MRI, or across different hardware vendors or even when using different imaging protocols. As a result, transferring a model across domains can severely degrade its performance. To that end, Kamnitsas et al. (2017) used adversarial training to increase the robustness of segmentation in brain MRI and Lafarge et al. (2017) in histopathology images.

Third, image-to-image translation using GANs has achieved impressive results in several applications, such as image enhancement (i.e. denoising (Yi and Babyn, 2018), super-resolution (Ledig et al., 2017) etc) and artistic style transfer (i.e. (Ulyanov et al., 2017)). Especially the former, has been shown to be successful in enhancing images from low-dose CT scans so that they become comparable with high-dose CTs, as shown in Wolterink et al. (2017) and Yi and Babyn (2018).

[*]Corresponding author: dimitrios@kheironmed.com

Preprint. Work in progress.

Finally, in semi-supervised learning, an adversarial objective can help to leverage unlabeled alongside labeled data in order to improve classification or detection performance. We refer to Lahiri et al. (2017) for an example of semi-supervised learning as applied to retinal images.

In recent years GANs have lead to breakthroughs in a number of different non-medical applications involving generation of synthetic images, such as single-image super-resolution (Ledig et al., 2017), image-to-image translation (Isola et al., 2017) and the generation of artistic images (Elgammal et al., 2017) to name a few.

GANs manage to ameliorate many of the issues associated with other generative models. For instance, auto-regressive models (Van Oord et al., 2016) generate image pixels one at a time, conditioned on all previously generated pixels, by means of a Recurrent Neural Network (RNN). These methods have shown promise, however have not yet been able to scale to high image resolutions. Additionally, the computational cost of generating a single image does not scale favorably with its resolution. With Variational Auto-encoders (VAEs) (Kingma and Welling, 2013), restrictions on the prior and posterior distributions limit the quality of the drawn samples. Furthermore, training with pixel losses exhibits an averaging effect across multiple possible solutions in pixel space, which manifests itself as blurriness (discussed in more detail in Ledig et al. (2017)). In contrast, GANs are able to produce samples in a single shot and do not impose restrictions on the generating distribution in a process similar to sampling from the multitude of possible solutions in pixel space, which generally leads to sharper and higher quality samples.

The framework for training generative models in an adversarial manner was first introduced in the seminal work of Goodfellow et al. (2014). This framework is based on a simple but powerful idea: the generator neural network aims to produce realistic examples able to deceive the discriminator which aims to discern between original and generated ones (a 'critic'). The two networks form an adversarial relationship and gradually improve one-another through competition, much like two opponents in a zero-sum game (see Fig. 1). The main disadvantage is that training these models requires reaching a Nash equilibrium, a more challenging task than simply optimizing an objective function. As a result, training can be unstable, susceptible to mode collapse and gradient saturations (Arjovsky et al., 2017).

Stabilizing GAN training becomes even more pertinent as our aim shifts towards high resolution images, such as medical images, where the dimensionality of the underlying true distribution in pixel space can be enormous and directly learning it may be unattainable. A key insight made in Karras et al. (2018) is that it is beneficial to start training at a low resolution, before gradually increasing it as more layers are phased in. This was shown not only to increase training stability at high resolutions, but also to speed up training, since, for much of the training, smaller network sizes are used.

The goal of this paper is to demonstrate the applicability of GANs in generating synthetic FFDMs. Mammograms contain both contextual information indicative of the breast anatomy and a great level of fine detail indicative of the parenchymal pattern. The large amount of high frequency information makes it imperative for radiologists to view these images in high-resolution. For instance, the spiculation of a mass or certain micro-calcification patterns as small as 1-2 pixels in diameter can indicate malignancy and are thus very important to consider. Our aim was to train a generator convolutional neural network (CNN) able to produce realistic, high-resolution mammographic images. For that purpose we attempted to follow Karras et al. (2018) as closely as possible, as our previous attempts generating even low-resolution images did not yield acceptable results.

The rest of the paper is arranged as follows: In Section 2 we summarize the key theoretical underpinnings of GANs, including their progressive training, and various stabilization methods that can be employed. In Section 3 we outline our methodology that builds on previously published literature and discuss the results of our experiments in detail. Finally, in the Appendix, readers can find several visual examples of successes and failures of the developed approach.

2 Generative Adversarial Networks

2.1 Conventional GAN

As described above, the GAN framework consists of a *generator* G tasked with generating samples highly probable under the true data distribution, and a *discriminator* D tasked with distinguishing synthesized samples from original ones. Both the generator and discriminator have cost functions

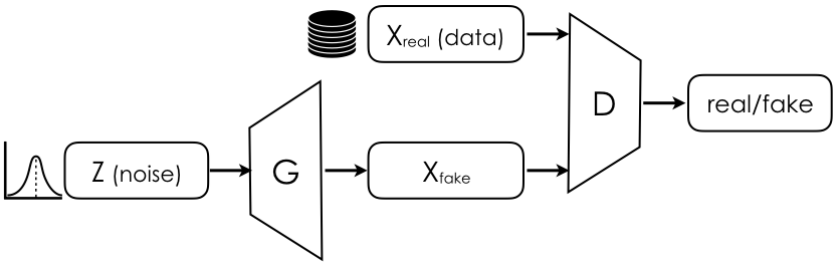


Figure 1: Schematic representation of a generative adversarial network.

which we aim to optimize, but are directly opposing each other, which may be regarded either as an adversarial zero-sum game or as a saddle point optimization problem.

The original cost function for the discriminator is simply the standard binary cross-entropy between two classes, original and generated. The discriminator is trained simultaneously on two batches of data, one batch sampled from the training data and the other sampled from the generator (following the suggestion for batch discrimination in Salimans et al. (2016)).

$$J_D = \mathop{\mathbb{E}}_{\boldsymbol{x}\sim\mathbb{P}_{\boldsymbol{x}}} \left[\log D\left(\boldsymbol{x}\right)\right] + \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathbb{P}_{\boldsymbol{z}}} \left[\log\left(1 - D\left(G\left(\boldsymbol{z}\right)\right)\right)\right] \tag{1}$$

where $\mathbb{P}_{\boldsymbol{x}}$ is the target distribution in pixel space $\boldsymbol{x}$ and $\mathbb{P}_{\boldsymbol{z}}$ is the latent space distribution, selected so that we can easily sample from, i.e. uniform or Gaussian.

The process, if viewed as a zero-sum game where a reduced cost for one player is an increased cost for the other, then we can express the generator objective simply as:

$$J_G = -J_D \tag{2}$$

We can then express the entire training process by the minimax game:

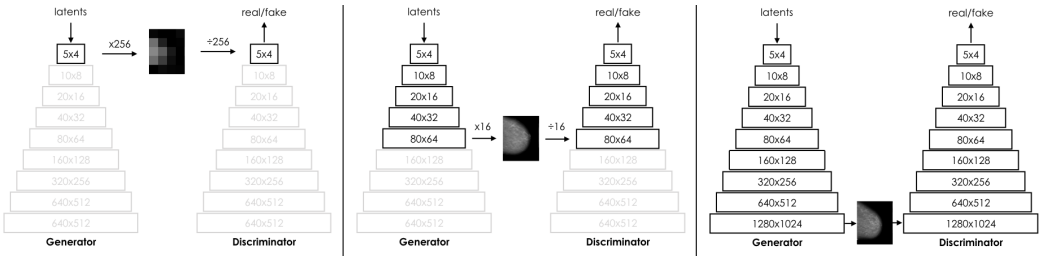
$$\boldsymbol{\theta}_G^* = \arg\min_{\boldsymbol{\theta}_G} \max_{\boldsymbol{\theta}_D} V\left(\boldsymbol{\theta}_G, \boldsymbol{\theta}_D\right) \tag{3}$$

with $V\left(\boldsymbol{\theta}_G, \boldsymbol{\theta}_D\right) = -J_D$ denoting the value function, and $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_D$ parameterize the generator and discriminator networks, respectively.

An issue with the above formulation is that if the discriminator becomes too effective at discerning original from generated examples, the second term of Eq. (1) approaches zero, the gradient of the generator vanishes and it cannot further improve. In an attempt to ameliorate that, the generator can be alternatively trained to minimize $-\log\left(D\left(G\left(\boldsymbol{z}\right)\right)\right)$ as suggested in the original paper of Goodfellow et al. (2014).

2.2 Wasserstein GAN

In Arjovsky et al. (2017), the authors show how the original GAN objective in Eq. (1) is potentially discontinuous with respect to the generator's parameters, which leads to instability during training. They proposed a new objective based on the Wasserstein distance (a.k.a. the earth mover's distance) to remedy this. Intuitively, the Wasserstein distance $W(p, q)$ is the minimum cost of transporting probability mass in order to transform one distribution p into another distribution q, where the cost is the mass multiplied by the transport distance. Under mild assumptions, $W(p, q)$ is continuous everywhere and differentiable almost everywhere, which the authors claim leads to improved stability during optimization.



(a) Scale 0 (b) Scale 4 (c) Scale 8

Figure 2: Illustration of the progressive growth of both networks during training.

Formally, the Wasserstein GAN objective function is defined using the Kantorovich-Rubinstein duality (Villani, 2009) as:

$$\min_{\theta} \max_{\boldsymbol{w} \in \mathcal{W}} \left[\mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathbb{P}_{\boldsymbol{z}}} \left[f_{\boldsymbol{w}} \left(g_{\boldsymbol{\theta}} \left(\boldsymbol{z} \right) \right) \right] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}}} \left[f_{\boldsymbol{w}} \left(\boldsymbol{x} \right) \right] \right] \tag{4}$$

where $f_{\boldsymbol{w}} \left(\cdot \right)$ is the critic function transforming an image to a discriminative latent feature space, as opposed to previously being trained to discern between original and generated images, and $\{ f_{\boldsymbol{w}} \}_{\boldsymbol{w} \in \mathcal{W}}$ is the set of all critic functions that are 1-Lipschitz continuous.

To enforce the Lipschitz continuity on the critic it is sufficient to clip the weights $\boldsymbol{w}$ of the critic to lie within a compact space $[-c, c]$ (Arjovsky et al., 2017). However, as Gulrajani et al. (2017) show, this clipping can lead to optimization problems. Instead, they propose adding a gradient penalty term to the Wasserstein objective as an alternative way to ensure the Lipschitz constraint. Their improved Wasserstein objective used in this work, is formulated as follows:

$$\mathcal{L} = \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathbb{P}_{\boldsymbol{z}}} \left[f_{\boldsymbol{w}} \left(g_{\boldsymbol{\theta}} \left(\boldsymbol{z} \right) \right) \right] - \mathop{\mathbb{E}}_{\boldsymbol{x} \sim \mathbb{P}_{\boldsymbol{x}}} \left[f_{\boldsymbol{w}} \left(\boldsymbol{x} \right) \right] + \lambda \mathop{\mathbb{E}}_{\hat{\boldsymbol{x}} \sim \mathbb{P}_{\hat{\boldsymbol{x}}}} \left[\left(\left\| \nabla_{\hat{\boldsymbol{x}}} f_{\boldsymbol{w}} \left(\hat{\boldsymbol{x}} \right) \right\|_2 - \beta \right)^2 \right] \tag{5}$$

where $\hat{\boldsymbol{x}}$ is a random interpolation between an original and a generated image, $\hat{\boldsymbol{x}} = \gamma \boldsymbol{x} + (1 - \gamma) g_{\boldsymbol{\theta}} \left(\boldsymbol{z} \right)$, $\gamma \sim \mathcal{U}(0, 1)$ and the hyper-parameter β is the target value of the gradient magnitudes, usually selected β=1.

2.3 Stabilization Methods

Despite their improved stability, even Wasserstein GANs remain notoriously difficult to train and subject to instabilities when the equilibrium between the generator and discriminator is lost. The problem stems from the fact that the optimal point of the joint GAN objective corresponds to a saddle point, which alternating SGD methods such as those used to train the generator and discriminator networks do not reliably converge to.

A lot of research is being dedicated to stabilizing convergence to this saddle point. Yadav et al. (2018) combine SGD with a 'prediction step' that prevents 'sliding off' the saddle due to maximization with respect to the discriminator overpowering minimization of the generator or vice-versa. Adolphs et al. (2018) exploit curvature information to escape from undesired stationary points and converge more consistently to a desired optimum. Finally, Daskalakis et al. (2018) use Optimistic Mirror Descent (OMD) to address the limit oscillatory behavior known to impede saddle point convergence.

Stable GAN convergence becomes even more elusive when high resolution images are involved. In this setting differences between the high frequency artifacts of original and generated images make it even easier for the discriminator to win out over the generator, destabilizing training. Progressively trained GANs, which we describe next, were developed to tackle this problem.

2.4 Progressive Training of GANs

The research towards using GANs to synthesize ever increasing resolution of images has recently lead to a breakthrough in the work of Karras et al. (2018). The underlying idea is to progressively increase

the resolution of generated images by gradually adding new layers to the generator and discriminator networks. The generator first learns to synthesize the high-level structure and low frequency details of the image distribution, before gradually shifting its attention to finer details in higher scales. The fact that the generator does not need to learn all scales at once leads to increased stability. Progressive training also reduces training time, since most of the iterations are done at lower resolutions where the network sizes are small.

The original work includes several further important contributions. A dynamic weight initialization method is proposed to equalize the learning rate between parameters at different depths, batch normalization is substituted with a variant of local response normalization in order to constrain signal magnitudes in the generator, and a new evaluation metric is proposed (Sliced Wasserstein distance).

2.5 Quantitative Evaluation Metrics

There are two main factors we wish to assess in order to estimate the quality of outputs from the trained generator network. One is how probable the synthesized images are under the true data distribution, and the other is how large is the support of the generated distribution. Neither of these factors are straightforward to quantitatively assess and have been a subject of research since the advent of GANs.

The difficulty in assessing the fidelity to which the generated distribution follows the true data distribution stems from the fact:

- We wish to compare sets of images, as opposed to pairs of images for which most image similarity metrics are designed.
- The comparison is based on conceptual attributes of appearance that are inherently subjective.

A first attempt to the problem was the consideration of the Inception Score (IS). Synthesized images $\boldsymbol{x}$ are presented to an ImageNet trained Inception model to produce a class prediction y, and a score is assigned based on the entropy of $p(y|\mathbf{x})$ and $p(y)$. Intuitively, high fidelity to the true distribution implies low entropy w.r.t. $p(y|\mathbf{x})$ (samples are unambiguous) and high distributional support translates to high entropy w.r.t. $p(y)$ (samples have high diversity).

An alternative to the IS, is the the Frèchet Inception Distance (FID) (Heusel et al., 2017), which instead compares the distributions of the feature maps for original and generated images. The FID directly utilizes the training image dataset and can be more robust to transferring to images that were not used to train the inception model, e.g., facial images, as long as the features are also discriminative in the new domain.

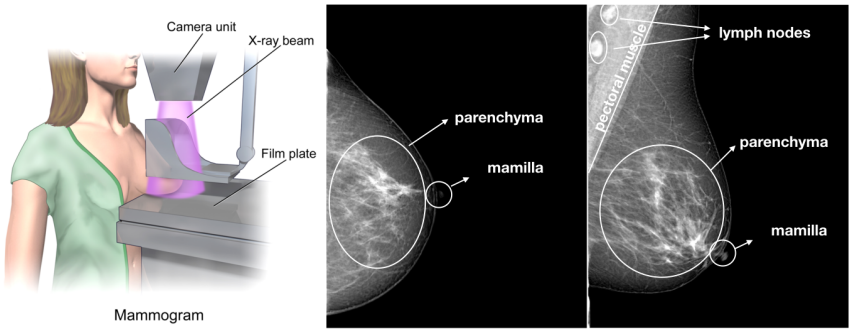
An alternative metric, not requiring the use of a trained model, is the Multi-scale Structural Similarity Index (MS-SSIM) (Odena et al., 2017; Wang et al., 2003). The SSIM was designed to improve upon traditional image quality metrics and has been used as a loss function in deep learning, as it is differentiable (Godard et al., 2017). In order to assess the quality of a trained GAN, it is necessary to randomly pair the original and generated images, compute the SSIM of each set and then compare with within set self-similarities.

Finally, an interesting alternative to the aforementioned metrics, proposed in Karras et al. (2018), is the Multi-scale Sliced Wasserstein metric. The concept is to compare the sorted sets of descriptors extracted from original and generated images. In order to make this metric computationally efficient, the authors have used descriptors that correspond to random projections of image patches.

3 Mammogram Synthesis

3.1 Clinical setting and data

Mammograms are relatively low-dose soft tissue X-rays of the breast. Acquisition is performed after each breast in turn has been flattened using two plastic paddles, as illustrated in Fig. 3a. Conventionally both left and right breasts are imaged using two standard views, the cranial-caudal (CC) and the mediolateral-oblique (MLO), which are shown in Fig. 3b. This results in a total of four 7-10 megapixel images per patient.



(a) Acquisition of CC image (Blausen, 2014).

(b) CC and MLO view examples.

Hanging protocols are the series of actions performed to arrange images on a screen to be shown to the radiologist. Hanging protocols are designed to work across hardware and clinical sites. In mammography, this defines how to setup and present the images for the reader, including preferred windowing of image intensities and image size.

We have acquired a large number of images (>1,000,000) from our partners which we used for the purpose of this work. From this proprietary dataset we excluded images containing post-operative artifacts (metal clips, etc.) as well as large foreign bodies (pacemakers, implants, etc.). Otherwise, the images contain a wide variation in terms of anatomical differences and pathology (including benign and malignant cases) and the dataset corresponds to what is typically found in screening clinics.

3.2 Training

We used a simple preprocessing method that preserves both the original aspect ratio of each image and the hanging protocol. More specifically, we down-sampled by the largest factor to match one of the desired dimensions and padded the other dimension with zeros. The final image size is 1280x1024 pixels which (to the best of our knowledge) is the highest image resolution generated by a GAN thus far.

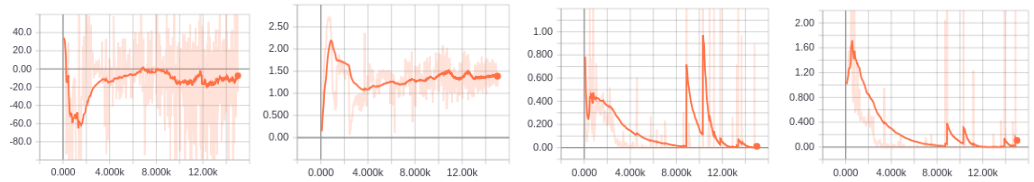
Despite using progressive training, we still had to overcome significant stability issues, due to the high resolution. We took several steps to maximize the probability of a successful run outlined in the following.

First, we increased the number of images used for training, from an initial 150k to 450k. This inevitably introduces more variation, along with some noise due to images that are erroneously included in the training set - some examples are shown in 16c of the Appendix. Nevertheless, we argue the extra information to be leveraged is beneficial for training.

Second, as suggested in Salimans et al. (2016), we added some supervised information. More specifically we conditioned on the view, namely CC and MLO, which is highly relevant as it has significant impact on the visual appearance of the images.

Finally, we slightly decreased the learning rate from the one originally used in Karras et al. (2018), from 0.002 to 0.0015 and gradually increased the discriminator iterations, from 1 to a maximum of 5 discriminator updates for each generator update. Even with these modifications, we had to often restart training and artifacts periodically appeared, but the network was able to recover in most cases. An example of the training progress for a successful run is shown in Fig. 4e.

We performed our training on an NVIDIA DGX-1, with 8 V100 GPUs, 16GB GPU memory each. We initially trained until the network was presented with 15 million images, which is equivalent to 33 epochs which took about 52 hours. Then we resumed training for an additional 5 million images and selected the best network checkpoint based on the Sliced Wasserstein Distance (Karras et al., 2018).

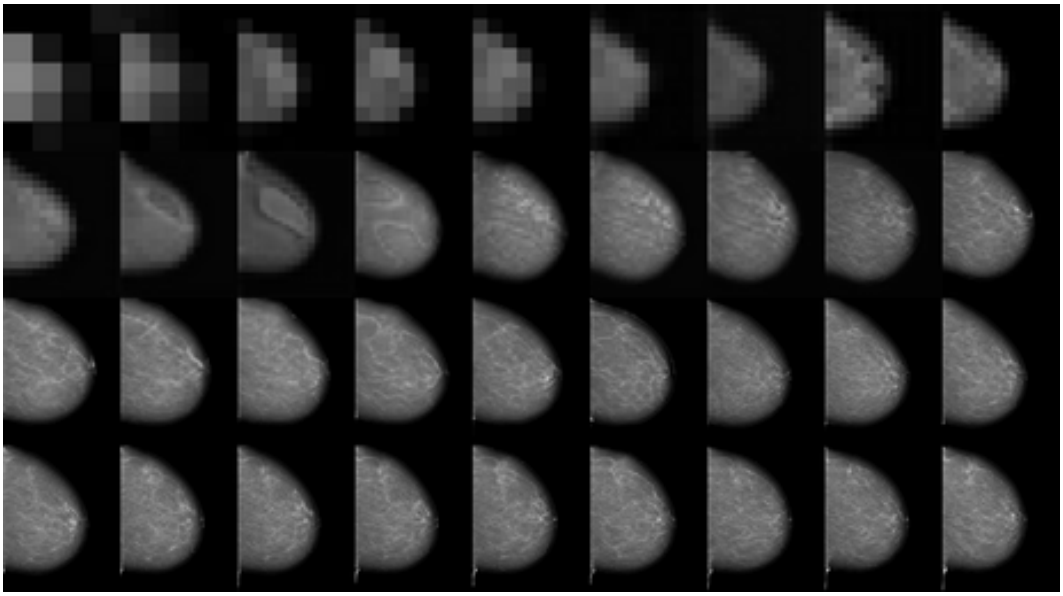


(a) Discriminator binary cross entropy.

(b) Gradient magnitudes contributing in Eq. (5).

(c) Label cross entropy for original images.

(d) Label cross entropy for generated images.



(e) The training progression of a successful run.

Figure 4: Note that artifacts appear after around 4.7 million images have been presented to the network. Training recovers shortly after that, however, as can be seen in the diagnostic plots, this failure is not easily detectable from the curves.

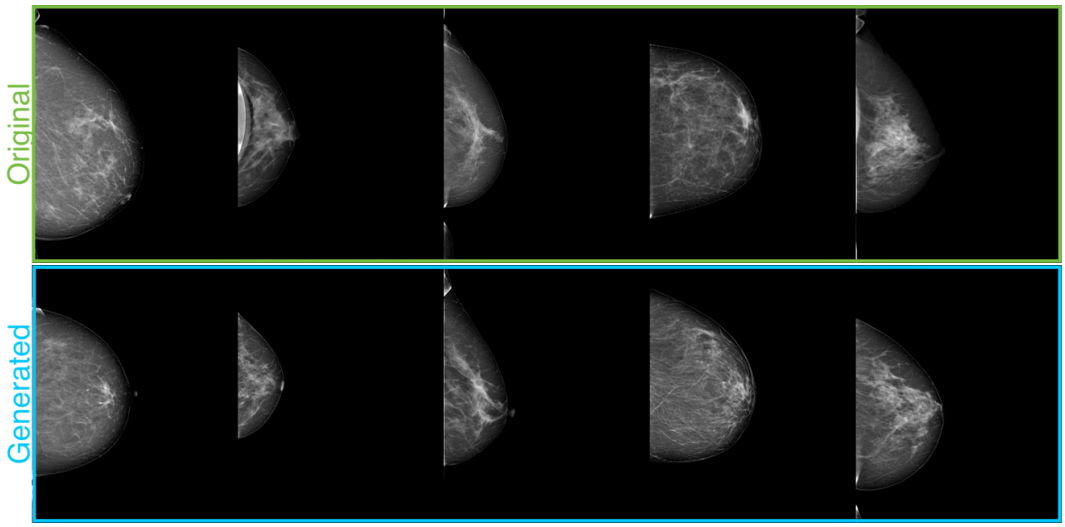
3.3 Results

The final samples drawn from a successfully trained network look very promising. Most of the generated images seem highly realistic with a broad range of inter-image variability, which indicates good representation of the underlying true distribution. However, we also observed some common artifacts and failures, which we discuss below.

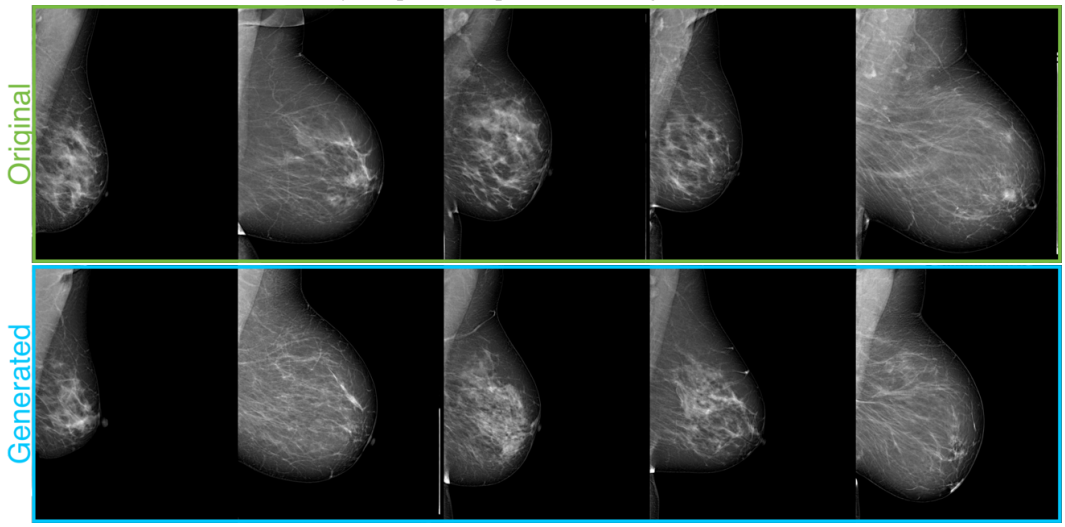
For more visual examples we refer to the Appendix, where we present images in several different format, described as follows:

- 6x5 grids of randomly selected generations from CC and MLO views (Fig. 9 and 10).
- 5x2 grids of randomly selected generations from CC and MLO views, alongside randomly selected real images. In this case, we also indicatively mark the best and worst generations (Fig. 11 and 12).
- 3x5 grids of handpicked convincing results from CC and MLO views (Fig. 13).
- 1x3 grids of handpicked convincing results, alongside real images (Fig. 14 and 15).
- 2x5 grids where we present examples of failures from CC and MLO views, along with images with artifacts from the training set (Fig. 16).

Views The MLO view is evidently the harder one to model, unsurprisingly so, as it exhibits the highest variation and contains the most anatomical information, with the pectoral muscle clearly visible, lymph nodes in some cases, and of course the breast parenchyma (Fig. 3b)



(a) Randomly sampled examples of real and generated CC views.



(b) Randomly sampled examples of real and generated MLO views.

Figure 5: Examples of generated images from the GAN.

Samples from the CC view seem subjectively of higher quality, due to their relative simplicity compared the MLO view.

Calcifications and metal markers Calcifications are caused naturally in the breast from calcium deposition and can vary in size and shape, but appear very bright (white) on the image as they fully absorb passing X-rays. They are important in mammography as certain patterns can be a strong indication of malignancy, while others are benign (e.g., vascular deposits).

External skin markers are frequently used by technicians performing the mammogram to indicate the position of a palpable lesion in the breast for the attention of the radiologists who is going to perform the reading. They also appear very bright, but are distinctively fully circular in shape.

In Fig. 8 we show an example of both calcifications and a marker in the bottom right, appearing in the same image.

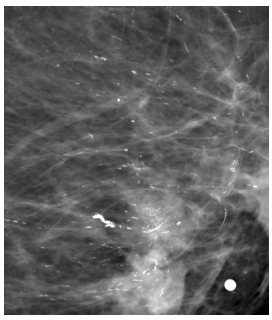


Figure 8: Calcifications and a round marker (bottom right) commonly used by the technician to indicate a palpable lesion.

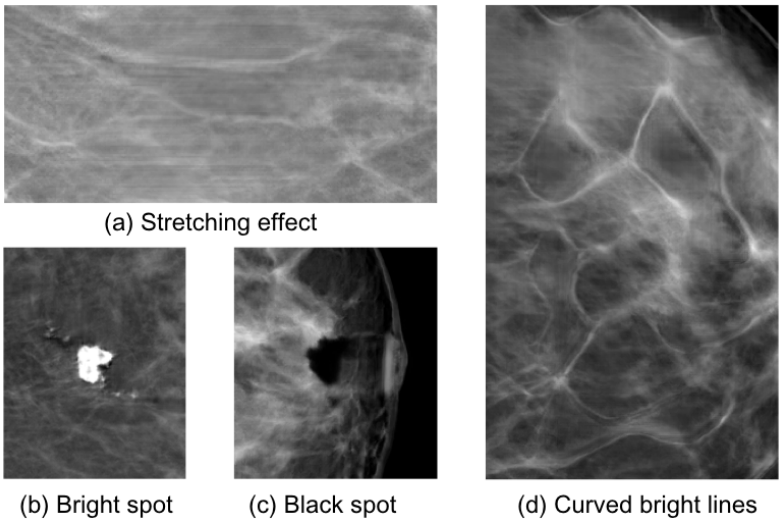


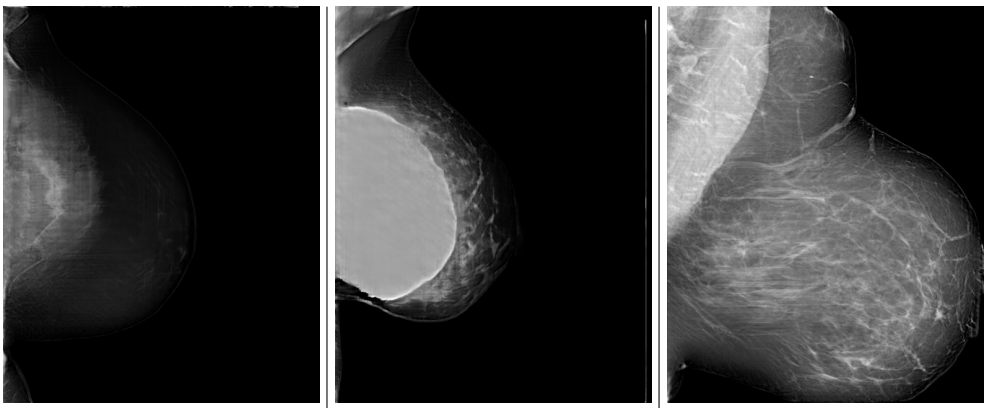
Figure 6: Most commonly seen artifact patterns

We have observed that the generator strongly resists these structures. It is only in very late stages of training that features roughly similar to medium sized calcification may appear in the generations, but they are not very convincing. We assume that the network architecture acts as a strong prior against such features, which do not appear in natural images (as also suggested in Ulyanov et al. (2018)).

Common artifacts We observe several types of failures in the generated images. Some of them are clearly network failures, which indicate that not all possible latent vectors correspond to valid images in pixel space. Others can be attributed to problems in the training set. Examples of such images are shown in Fig. 16c.

4 Conclusion

In this work we present our methodology for generating highly realistic, high-resolution synthetic mammograms using a progressively trained generative adversarial network (GAN). Generative models can be especially valuable for medical imaging research. However, GANs have not so far been able to scale to the high resolution required in FFDM. We have managed to overcome the underlying



(a) Transitioning towards a larger size.

(b) Attempted reproduction of breast implant.

(c) Distorted reproduction near the right hand side border of the image.

Figure 7: Common failures.

instabilities inherent in training such adversarial models and have been able to generate images of highest resolution reported so far, namely 1280x1024 pixels. We have identified a number of limitations, including common artifacts and failure cases, indicating that further research is required but that promising results can already be achieved. We hope this work can serve as a useful guide and facilitate further research on GANs in the medical imaging domain.

Acknowledgments

The authors would like to thank Dr. Andreas Heindl and Dr. Galvin Khara for their valuable inputs and insights, Nicolas Markos for his valuable expertise in high performance computing, as well as the rest of their colleagues at Kheiron Medical Technologies Ltd. for their help and support.

References

Adolphs, L., Daneshmand, H., Lucchi, A., and Hofmann, T. (2018). Local Saddle Point Optimization: A Curvature Exploitation Approach. (2).

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv preprint arXiv:1701.07875 (2017)*.

Blausen, M. (2014). Blausen gallery 2014. *Wikiversity Journal of Medicine*, 1(2).

Costa, P., Galdran, A., Meyer, M. I., Niemeijer, M., Abràmoff, M., Mendonça, A. M., and Campilho, A. (2017). End-to-end adversarial retinal image synthesis. *IEEE transactions on medical imaging*, 37(3):781–791.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. (2018). Training GANs with Optimism. In *Proceedings of the International Conference on Learning Representations*.

Elgammal, A., Liu, B., Elhoseiny, M., and Mazzone, M. (2017). CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. *arXiv preprint arXiv:1706.07068 (2017)*.

Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. (2018). Synthetic data augmentation using gan for improved liver lesion classification. In *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 289–293.

Godard, C., Mac Aodha, O., and Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. pages 6602–6611.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637.

Isola, P., Zhu, J. Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. pages 5967–5976.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al. (2017). Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *International Conference on Information Processing in Medical Imaging*, pages 597–609. Springer.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. In *Proceedings of the International Conference on Learning Representations*.

Kingma, D. P. and Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114 (2013)*.

Lafarge, M. W., Pluim, J. P., Eppenhof, K. A., Moeskops, P., and Veta, M. (2017). Domain-adversarial neural networks to address the appearance variability of histopathology images. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 83–91. Springer.

Lahiri, A., Ayush, K., Biswas, P. K., and Mitra, P. (2017). Generative adversarial learning for reducing manual annotation in semantic segmentation on large scale miscroscopy images: Automated vessel segmentation in retinal fundus image as test case. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 42–48.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690.

Odena, A., Olah, C., and Shlens, J. (2017). Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pages 2642–2651.

Salehinejad, H., Valaee, S., Dowdell, T., Colak, E., and Barfett, J. (2018). Generalization of Deep Neural Networks for Chest Pathology Classification in X-Rays Using Generative Adversarial Networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2017). Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 6.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. (2018). Deep image prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Van Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In *International Conference on Machine Learning*, pages 1747–1756.

Villani, C. (2009). *Optimal transport : old and new*. Springer.

Wang, Z., Simoncelli, E., Bovik, A., et al. (2003). Multi-scale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems, and Computers*, volume 2, pages 1398–1402.

Wolterink, J. M., Leiner, T., Viergever, M. A., and Išgum, I. (2017). Generative adversarial networks for noise reduction in low-dose ct. *IEEE transactions on medical imaging*, 36(12):2536–2545.

Yadav, A., Shah, S., Xu, Z., Jacobs, D., and Goldstein, T. (2018). Stabilizing adversarial nets with prediction methods. In *Proceedings of the International Conference on Learning Representations*.

Yi, X. and Babyn, P. (2018). Sharpness-aware low-dose ct denoising using conditional generative adversarial network. *Journal of digital imaging*, pages 1–15.

Appendices

A Further examples

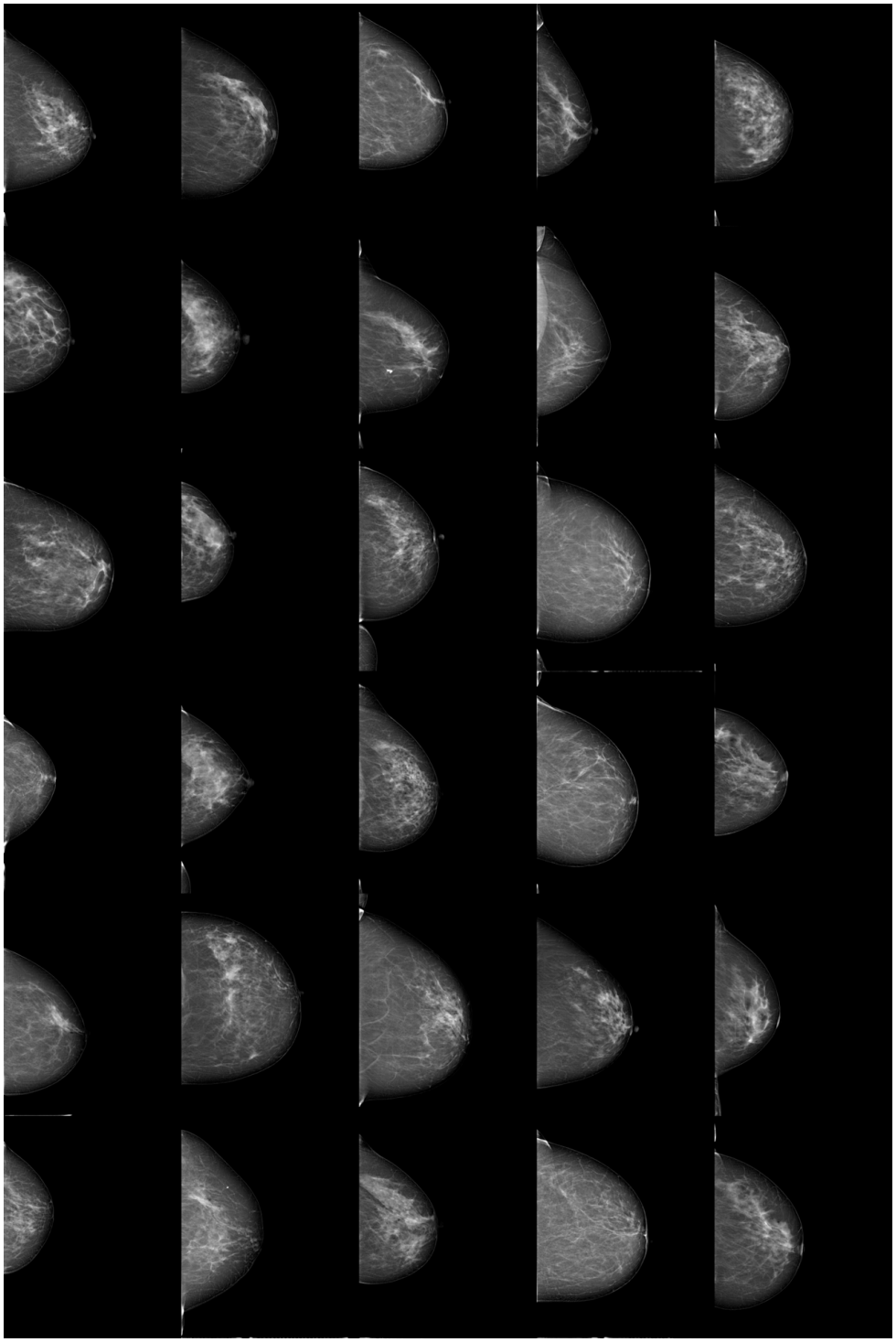


Figure 9: Random samples of generated CC views.

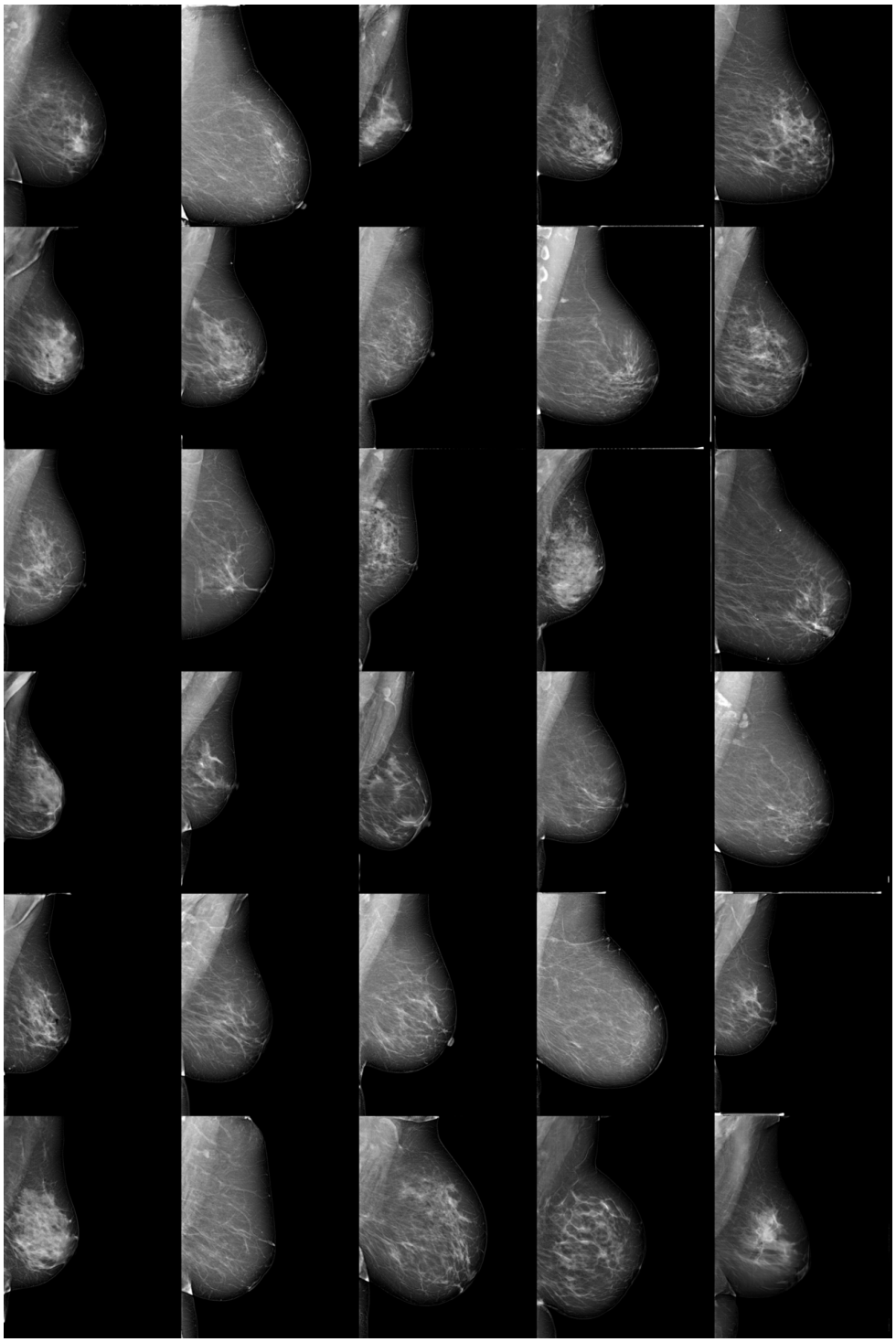
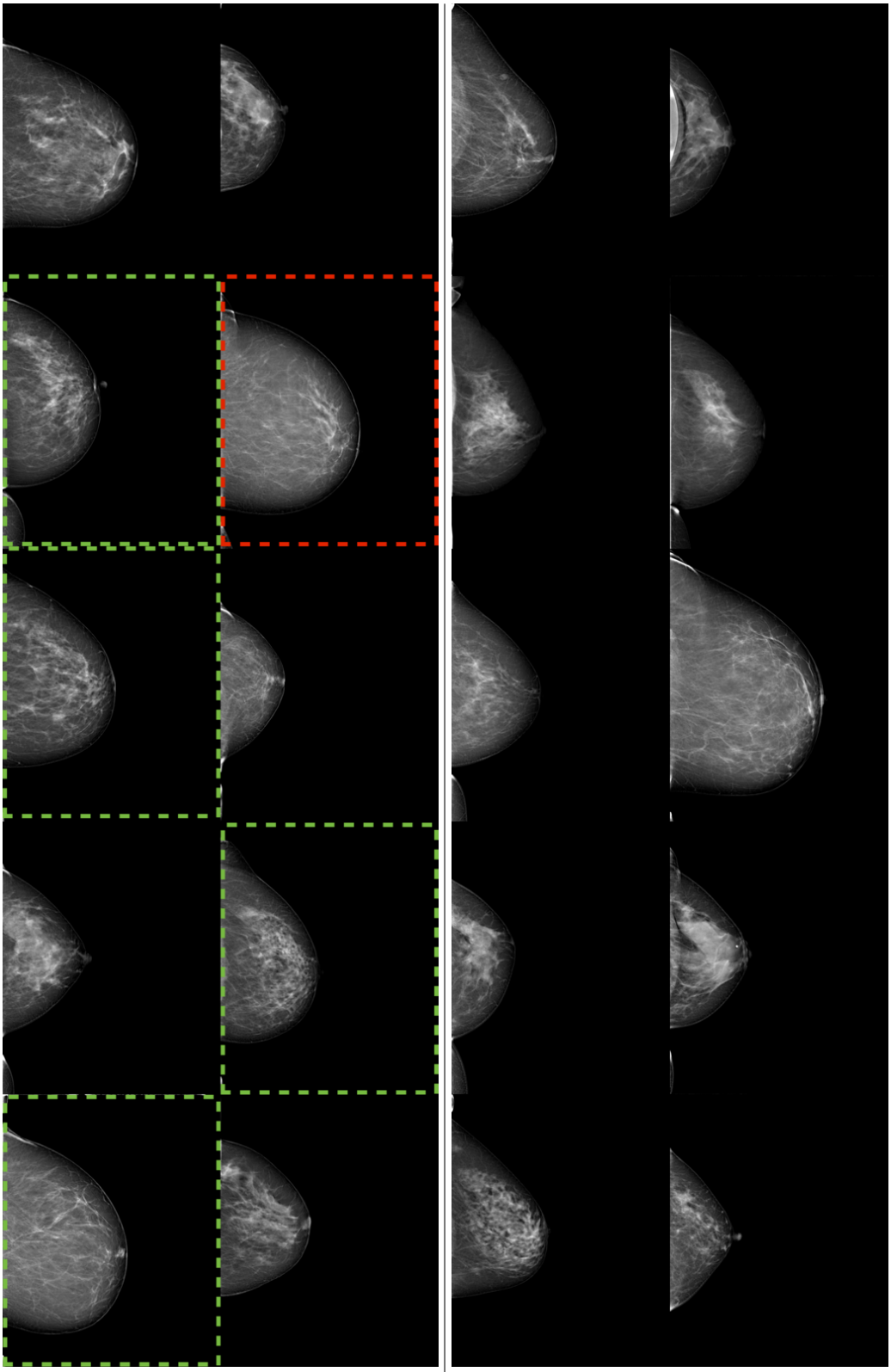
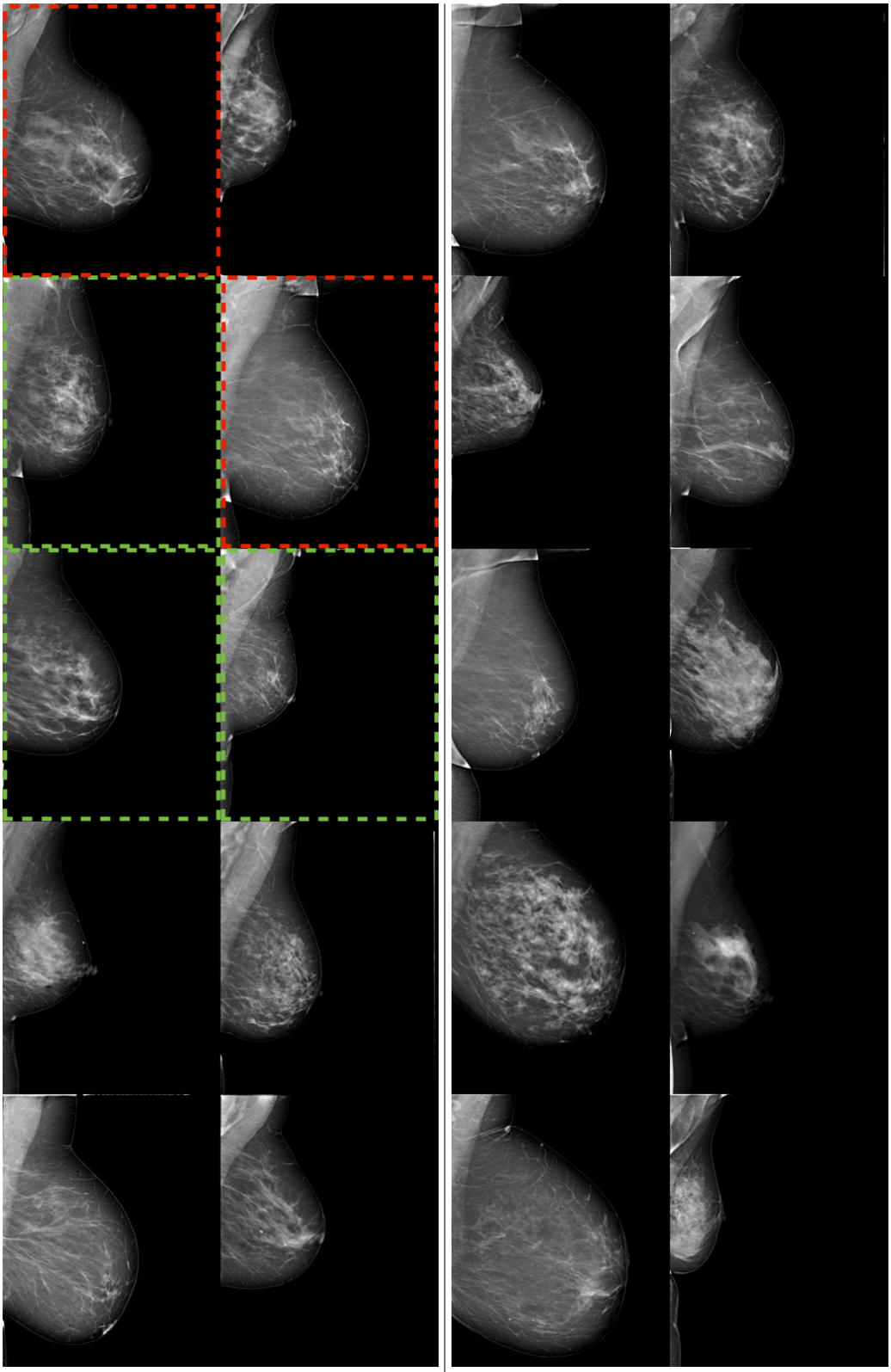


Figure 10: Random samples of generated MLO views.



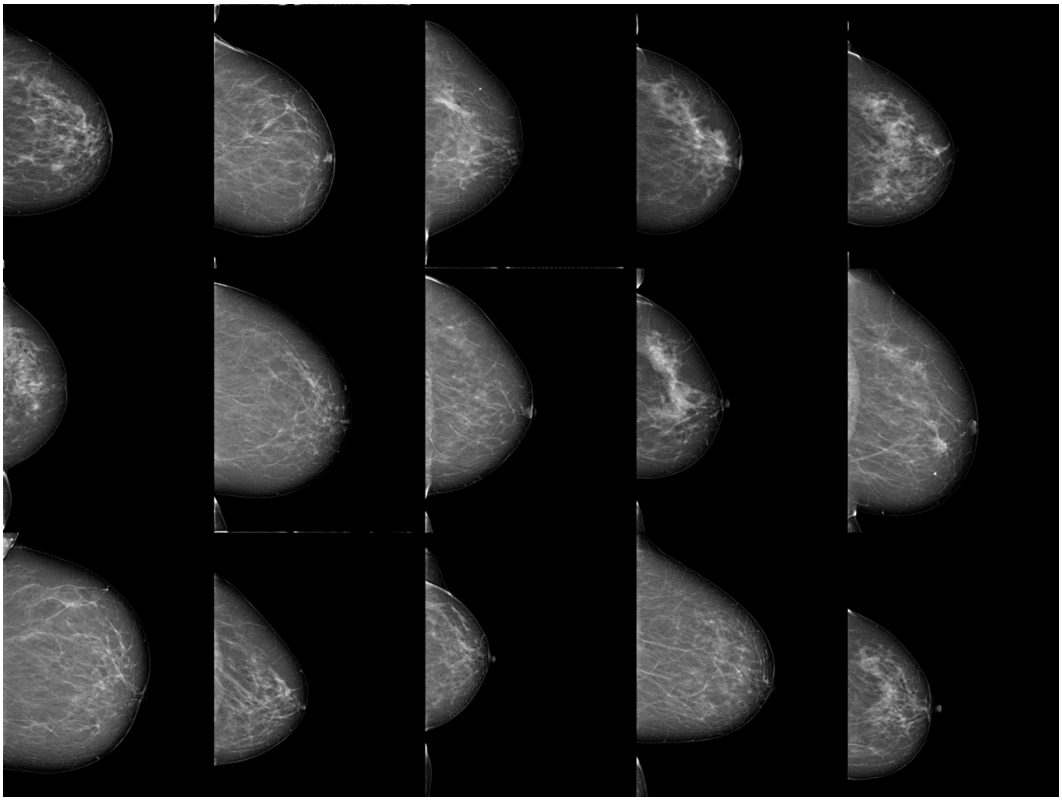
(a) Generated (b) Original

Figure 11: Randomly sampled original and generated CC views. The green dashed line denotes particularly convincing samples and the red dashed line denotes images with obvious artifacts.

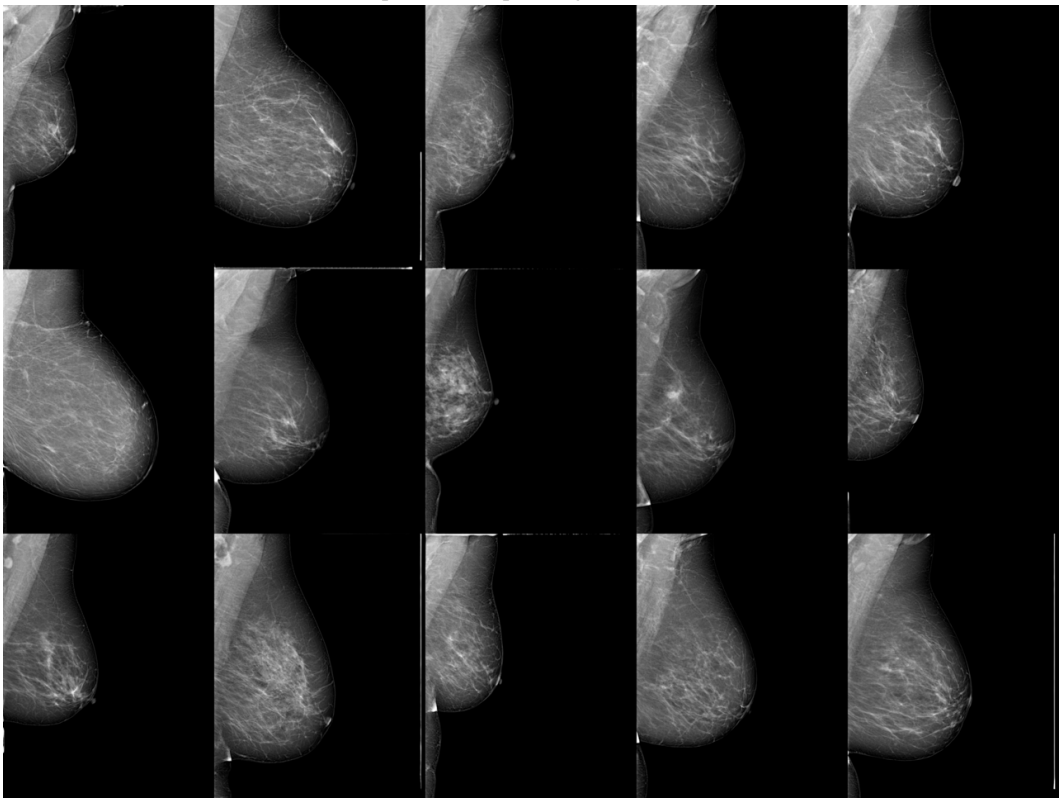


(a) Generated (b) Original

Figure 12: Randomly sampled original and generated MLO views. The green dashed line denotes particularly convincing samples and the red dashed line denotes images with obvious artifacts.

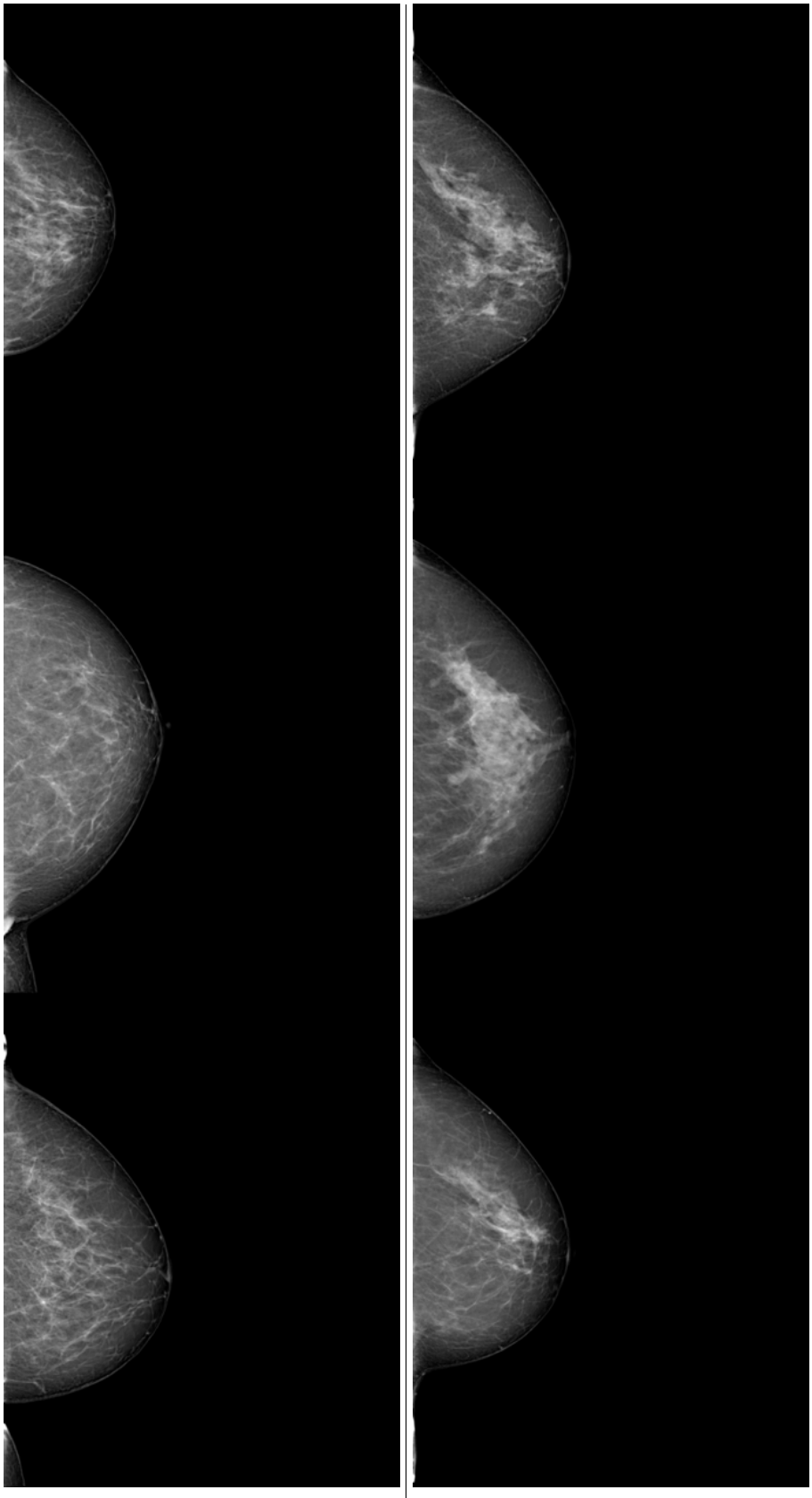


(a) Hand-picked examples of generated CC views.



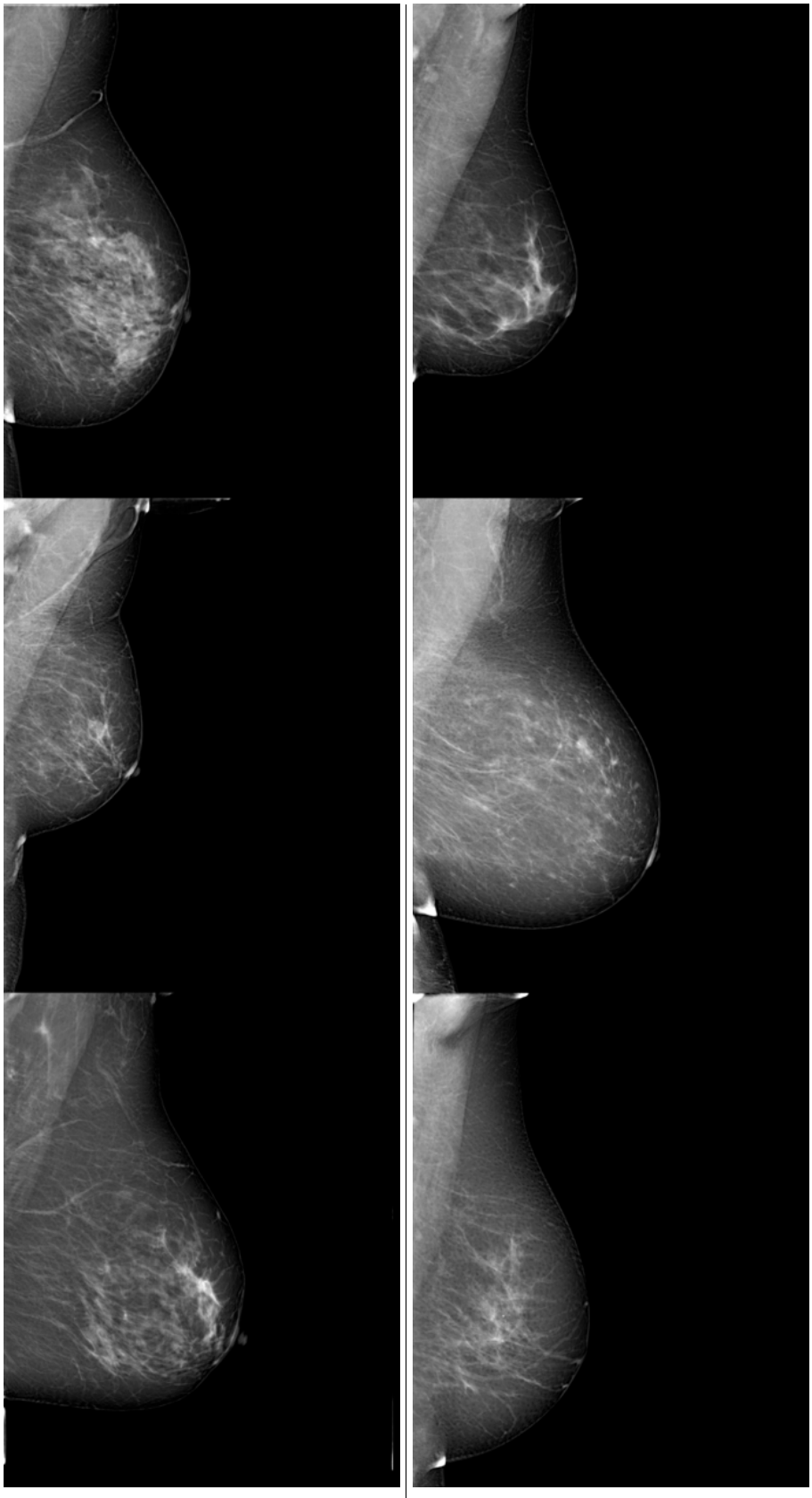
(b) Generated images from MLO view.

Figure 13: Handpicked examples of both CC and MLO views.



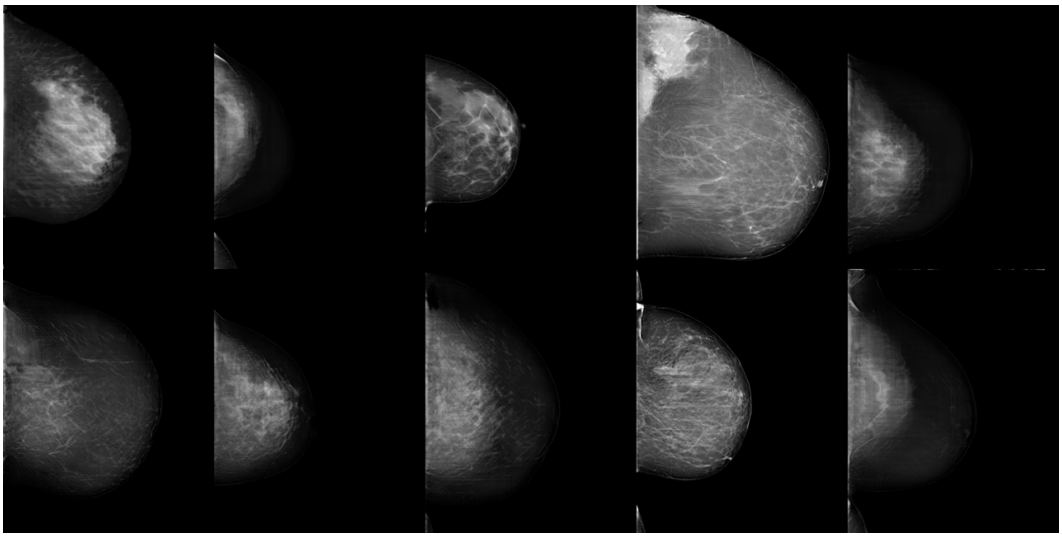
(a) Generated (b) Original

Figure 14: Handpicked generated CC views alongside random original CC views.

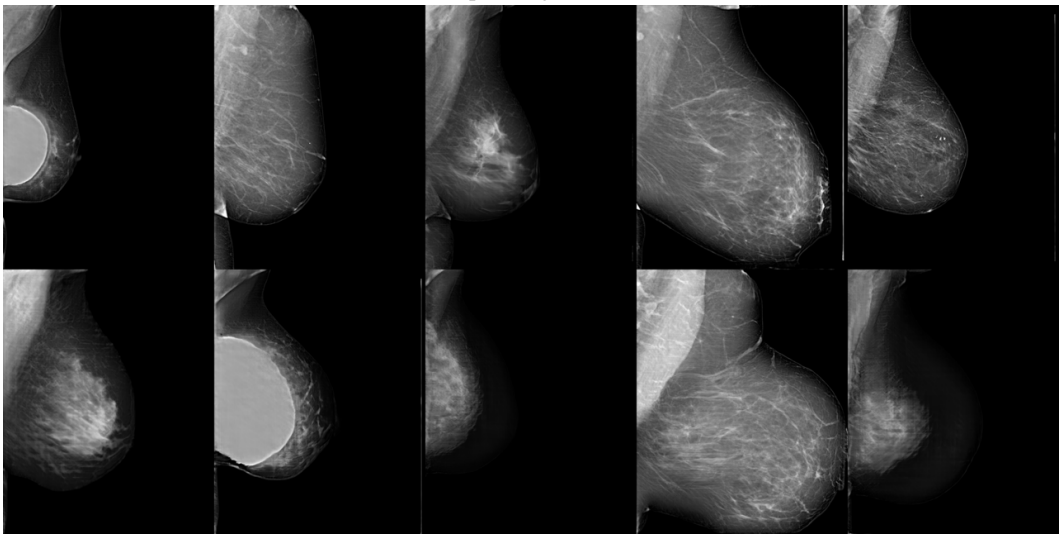


(a) Generated (b) Original

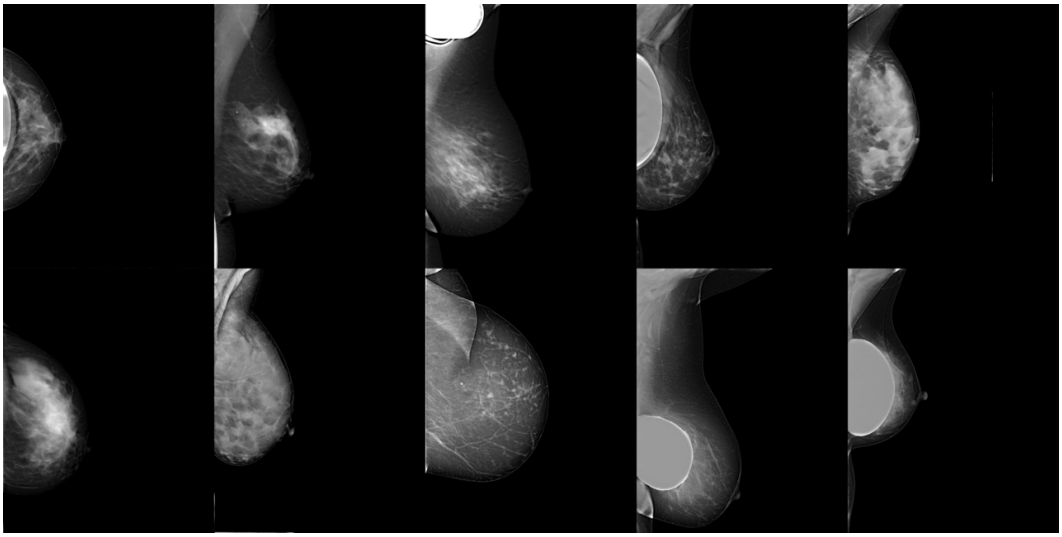
Figure 15: Handpicked generated CC views alongside random original CC views.



(a) Worst examples of generated CC views.



(b) Worst examples of generated MLO views.



(c) Original images with problematic appearances.

Figure 16: Worst examples we could find from both CC and MLO views.