Mask Embedding for Realistic High-Resolution Medical Image Synthesis

Yinhao Ren[1], Zhe Zhu[2], Yingzhou Li[3], Dehan Kong[4], Rui Hou[5], Lars J. Grimm[2], Jeffery R. Marks[6], and Joseph Y. Lo[1,2,5(✉)]

[1] Department of Biomedical Engineering, Duke University, Durham, USA
{yinhao.ren,joseph.lo}@duke.edu
[2] Department of Radiology, Duke University School of Medicine, Durham, USA
{zhe.zhu,lars.grimm}@duke.edu
[3] Department of Mathematics, Duke University, Durham, USA
yinzhou.li@duke.edu
[4] Department of Automation, Beijing Institute of Technology, Beijing, China
dehan.kong@duke.edu
[5] Department of Electrical Engineering, Duke University, Durham, USA
rui.hou@duke.edu
[6] Department of Surgery, Duke University School of Medicine, Durham, USA
jeffery.marks@duke.edu

Abstract. Generative Adversarial Networks (GANs) have found applications in natural image synthesis and begin to show promises generating synthetic medical images. In many cases, the ability to perform controlled image synthesis using masked priors such as shape and size of organs is desired. However, mask-guided image synthesis is challenging due to the pixel level mask constraint. While the few existing mask-guided image generation approaches suffer from the lack of fine-grained texture details, we tackle the issue of mask-guided stochastic image synthesis via mask embedding. Our novel architecture first encodes the input mask as an embedding vector and then inject these embedding into the random latent vector input. The intuition is to classify semantic masks into partitions before feature up-sampling for improved sample space mapping stability. We validate our approach on a large dataset containing 39,778 patients with 443,556 negative screening Full Field Digital Mammography (FFDM) images. Experimental results show that our approach can generate realistic high-resolution (256×512) images with pixel-level mask constraints, and outperform other state-of-the-art approaches.

Keywords: Generative Adversarial Networks · Image synthesis · Mask embedding · Mammogram

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-32226-7_47) contains supplementary material, which is available to authorized users.

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11769, pp. 422–430, 2019.
https://doi.org/10.1007/978-3-030-32226-7_47

1 Introduction

The rapid development of generative models especially on training methodology [1,9] and architecture [10], has led to the significant improvement of resolution and quality of the output images. In the image editing scenario, semantic control of the generated images such as object category and shape is highly desired. Many studies have explored conditional Generative Adversarial Networks (cGANs) [4,12] using class labels (one-hot vectors) [4] and mask labels [8]. Producing highly stochastic outputs as well as capturing the full entropy of the conditional distributions are of a great challenge for current cGANs.

Most cGANs derive from the basic generator–discriminator architecture. By adding conditional information to both the generator and the discriminator, cGANs can control some characteristics of the generated images. The most straightforward way to incorporate class label information is to directly concatenate the label vector with the latent vector in the generator and then to concatenate the conditional information with the latent features in the discriminator [12]. On the other hand, incorporating a pixel-level mask label requires special design of the networks to preserve the fine-grained texture details while satisfying the mask constraint [8].

In this paper we propose a novel approach to improve the generation of realistic high-resolution medical images with semantic control. We use a U-Net [16] style generator that takes both a **latent vector** (Gaussian noise vector) and **semantic mask**. Our generator first perform embedding of the mask input and then concatenate the mask embedding vector to the latent noise vector as the input of the feature projection path. The mask is an image providing constraints and could be an edge map, a segmentation map, a gradient field, etc.

We summarize our contributions as follows:

1. We propose to use mask embedding in semantic cGANs that takes both a mask and a random latent vector as the input. With this novel structure we can generate highly stochastic images with fine-grained details.
2. We apply the proposed approach to a medical image synthesis task, and generated realistic high-resolution images. Specifically we synthesize mammograms with a binary mask that indicates the breast shape. To our best knowledge this is the first work that can generate realistic high-resolution mammograms with semantic control.

2 Motivation

The ability to synthesize FFDM images with cancer lesions in a controlled fashion is greatly desired by the medical imaging machine learning community. Specifically, a mask-guided stochastic generator for medical image data augmentation could potentially yield gains in detection and classification algorithms given the low occupancy of pathology related pixels in most medical imaging modalities. To actually realize this gain, the generator needs to (1) have efficient mechanisms for sample space mapping to majority of the available training images

(e.g. Latent Vector); (2) learn joint distribution of semantic inputs and feature realizations (e.g. Mask Embedding). This work represent the first stage for our multi-stage study to synthesize mammogram with lesions. We investigate the feasibility of synthesizing clinically normal FFDM images with mask constraint.

3 Related Work

Medical image synthesis has been a well-motivated research topic for a long time. The ability to generate infinite number of realistic looking medical phantom greatly enables studies such as virtual clinical trials and data augmentation for computer aided diagnosis algorithms.

Recently many studies have been using GANs to synthesize medical images. Those methods can be grouped into unconditioned synthesis [6,13] and conditioned synthesis [3,14,15]. There is also a detailed survey of medical image synthesis using GANs [18]. Note that mammogram synthesis using GANs has been proposed in [11], but their approach focuses on image realism and resolution in the unconditional setting, hence the shape or other morphological characteristics of their synthesized mammograms cannot be controlled.

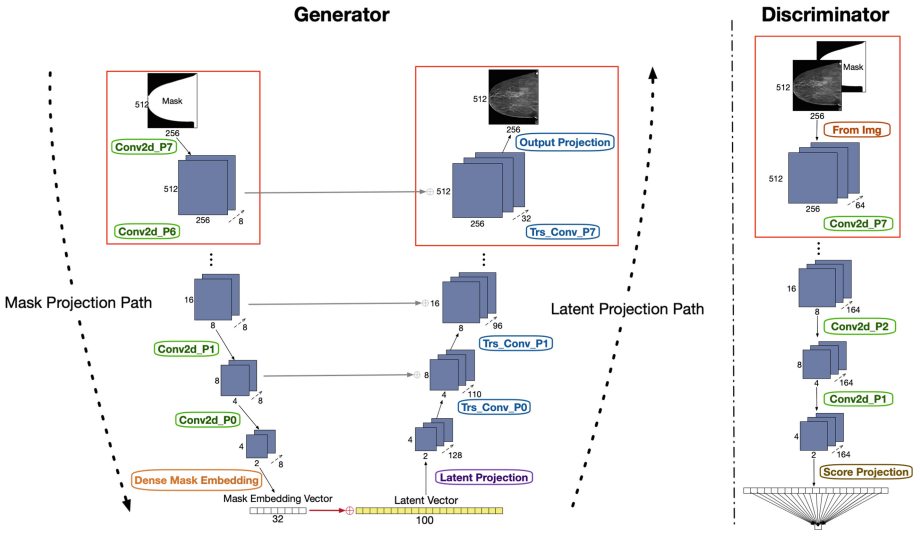

Fig. 1. Proposed architecture with two inputs to the Generator: (1) **Mask** and (2) **Latent Vector**. When doubling the dimension of the networks at beginning of each training phase, layers at the positions indicated by red boxes are newly initialized and faded in as described in the progressive training strategy for GAN [9]; For our without embedding baseline model, the **Dense Mask Embedding Layer** in the generator is removed. The latent vector input is also adjusted to a 132-dimensional vector to maintain the same amount of parameters in the latent projection path for a fair performance comparison. (Color figure online)

4 Network Architecture

The challenge to train an image translation model with latent vector input is to produce latent features that are compatible with the input mask constraint. In other words, the initial latent projection layer could produce features that fall outside the manifold constrained by the semantic mask input in the latent space, resulting in the following deconvolution layers to compensate for this inefficiency, and eventually leading to model capacity reduction. Our solution is to pose an additional constraint on the initially projected features by injecting the mask embedding vector into the input latent vector. This process allows a more efficient initial feature projection that produces latent features that are compatible with the mask input, thus preserving the output image quality.

Our model consists of a generator and a discriminator shown in Fig. 1. The key concept is to perform mask embedding in the generator before the latent feature projection layer to increase the overall feature projection efficiency. The generator follows a U-Net style design that can be divided into the **mask projection** path and the **latent projection** path. The discriminator takes the output of the generator as well as the corresponding mask and produce a probability score.

The input of the generator's mask projection path is a 256×512 mask (a binary image in our case). This mask projection path has 7 convolution layers each with a stride of 2 and depth of 8 features. The output of the mask projection path is a 32-dimensional vector (mask embedding) and is injected into the latent vector as the input of the latent projection layer. The latent vector is a 100-dimensional vector thus the input of the latent projection path is a 132-dimensional vector. Each mask projection feature block (except for the last one) is then concatenated onto the corresponding latent projection feature block to complete the U-Net structure. The initial latent feature block is produced by a dense layer followed by 7 deconvolution layers with stride of 2 and size of 4. The number of kernels of each deconvolution layer starts from 128 and decreases by a factor of 0.8 (rounded to the nearest integer) in each following layer. The output of the projection layer is the synthesized image.

5 Progressive Training

We used the progressive training strategy for GAN [9]. The training was divided into 7 phases. In each phase we doubled the network dimensions and gradually faded in the newly added layer. The model was grown from the resolution of 4×8 to 256×512. We stopped at this resolution due to hardware limitation. We adjusted the batch size and learning rate for each phases so that the standard WGAN-GP [5] converging mechanism can be achieved. We trained our model on three 1080 Ti for approximately a week to reach the maximum resolution. For each phase we train the network until the discriminator loss converges and no further observable improvement is made on the synthesized images. More details can be found in our open sourced implementation.

6 Experiments

6.1 Dataset

We used the mammography dataset collected from our institution. The dataset contains 39,778 negative screening subjects. Each exam has at least 4 images (Craniocaudal view and Mediolateral oblique view for each side of the breast), resulting in 443,556 images in total. The pixel values are truncated and normalized to $[0, 1]$ using the window level settings provided by the DICOM header. Each image was padded and resized to 256×512. For each mammography image a binary skin mask is generated using Otsu thresholding, where 1 denotes breast region and 0 denotes background. For comparison against the pix2pix model [8], we extracted the edge map of each mammography images using Sobel filters in both horizontal and vertical direction and then overlay the texture map with the corresponding skin mask.

6.2 Results

Several example results using randomly sampled skin masks and latent vectors are shown in Fig. 2. We compared our proposed model with the pix2pix model and our baseline model without embedding mechanism. Our proposed model generates mammograms with much more realistic texture details.

6.3 Comparison to Pix2Pix Method

We compare the results of our approach with the well-known pix2pix image translation model that takes only semantic mask as input. Results are shown in Fig. 3(c). Due to the image transformation nature of this model, our first approach using a smooth skin mask as the only input failed to generate any meaningful texture details. In order to evaluate the response of pix2pix model to perturbation of mask input, we constructed the texture map as mentioned in Sect. 6.1. Even trained on masks with the prior information of high frequency tissue structures, the standard pix2pix model still under performs our proposed model in terms of fine-grained texture details and variety of parenchymal patterns generated. The pix2pix result lacks stochasticity in which a very limited mapping between mask input space and sample space is possible, thus limiting the output variation. Moreover, the same problem limits training stability since the model is forced to map similar input binary patterns with drastically different realization of tissue structures without having the proper mechanism.

6.4 Comparison to Baseline Method

We explore the effect of mask embedding mechanism by removing the mask embedding layer from our proposed model and training this baseline model from scratch. The design of our baseline model is equivalent to Tub-GAN [19]. The

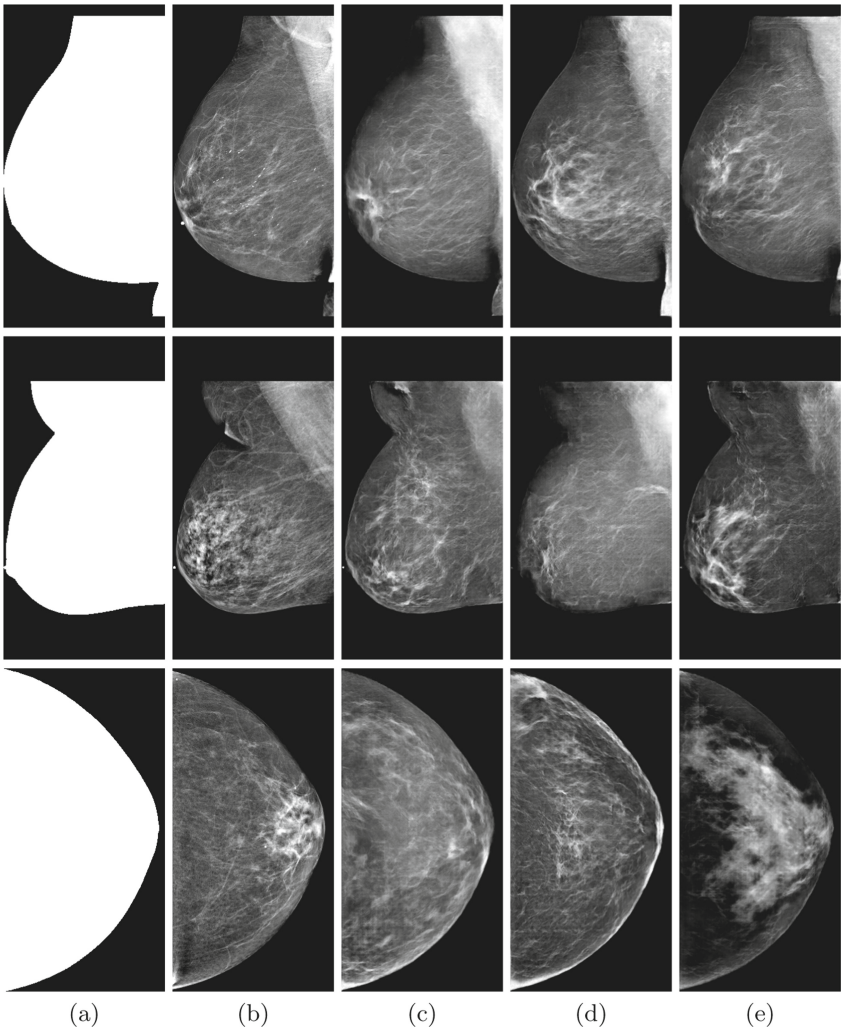

Fig. 2. (a) Input mask. (b) Original mammogram. (c), (d), (e) Generated mammograms using our mask embedding approach with different random latent vectors.

latent input vector is adjusted to be 100+32 so that the total number of parameters in the latent projection layer stays the same to our proposed model. The exact same training schedule for our proposed model is repeated. The results are shown in Fig. 3(d). The generated images have more high resolution details compared to pix2pix mode, but lack parenchyma complexity and usually contain obvious artifacts formed during up-sampling. This is an indication of model losing capacity due to the constraint posed by the mask input. A larger collection of comparison images can be found in supplementary material.

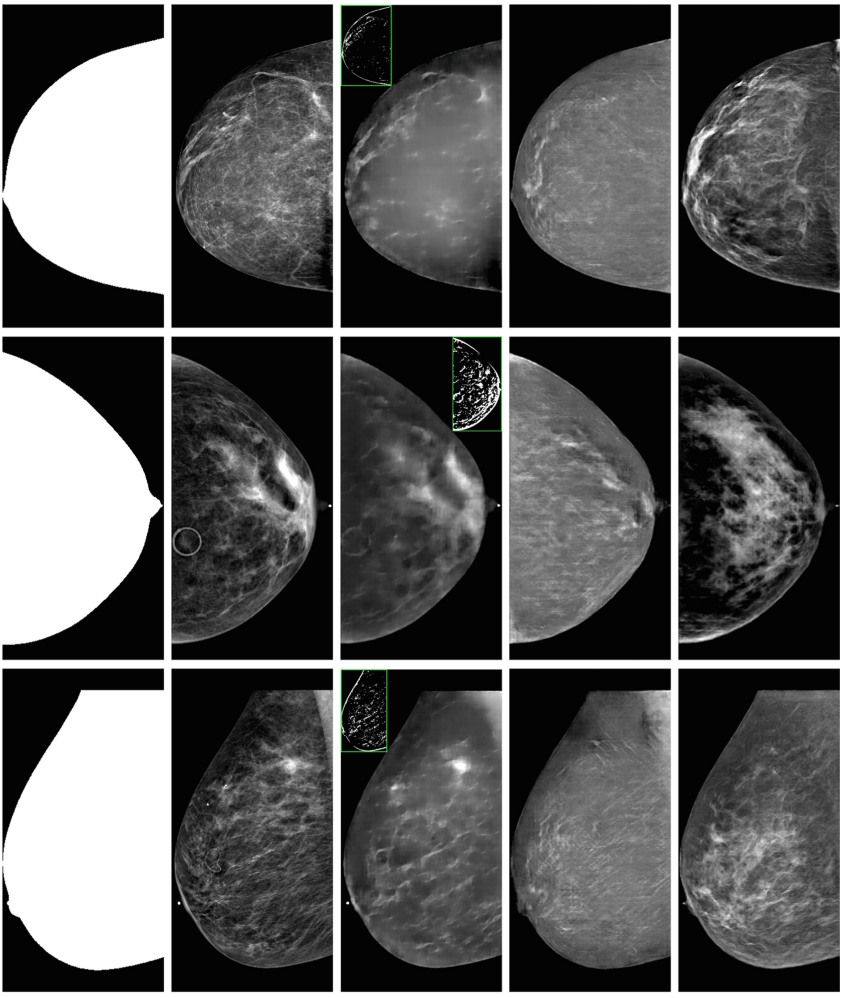

Fig. 3. (a) Input mask. (b) Original mammogram. (c) Synthesized mammogram using Pix2Pix. (d) Synthesized mammogram using our approach without mask embedding. (e) Synthesized mammogram using our approach with mask embedding.

6.5 Evaluation

For natural image generation there have been several objective metrics to measure the performance of the generative models such as Inception Score [17], Mode Score [2] and Fréchet Inception Distance [7], in medical image generation however there is no such metric available. Thus we design a reader study and let the expert radiologists assess the realism and quality of the synthesized results.

We randomly picked 50 real breast masks and generated mammograms using the three different approaches: pix2pix, our approach without mask embedding and our approach using mask embedding. All images were presented to readers in random order. Two expert radiologists were asked to rate each mammogram using 5 scores (5: definitely realistic, 4: realistic, 3: neutral, 2: fake, 1: definitely fake). The averaged score for real mammograms, synthesized results using pix2pix, synthesized results using our approach without and **with mask embedding** are 3.78, 1.08, 1.34, **2.38** respectively. Although subjective, these numerical results confirm that our approach with mask embedding provides a considerable improvement in realism.

7 Conclusion

We have proposed to use binary mask constraint to guide image synthesis while preserving output variety and fine-grained texture details. The challenge was to compensate for the generator capacity reduction caused by the pixel-level mask constraint. Our solution is to use mask embedding to further guide the initial projection of latent features to increase the probability of latent features falling within the manifold constrained by the mask. Our approach enables the semantic control of the synthesized mammograms while ensuring the fine-grained texture details are looking realistic. This technique can potentially be applied to other high resolution medical image modalities as well as natural images.

Acknowledgments. This work was supported in part by NIH/NCI U01-CA214183 and U2C-CA233254, and an equipment donation by NVIDIA Corporation.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, vol. 70, pp. 214–223 (2017)
2. Che, T., Li, Y., Jacob, A.P., Bengio, Y., Li, W.: Mode regularized generative adversarial networks. CoRR abs/1612.02136 (2016)
3. Costa, P., et al.: End-to-end adversarial retinal image synthesis. IEEE Trans. Med. Imaging **37**(3), 781–791 (2018)
4. Goodfellow, I., et al.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680 (2014)
5. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of Wasserstein GANs. In: Guyon, I., et al. (eds.) Advances in Neural Information Processing Systems, pp. 5767–5777 (2017)
6. Han, C., et al.: GAN-based synthetic brain MR image generation. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 734–738, April 2018. https://doi.org/10.1109/ISBI.2018.8363678
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a Nash equilibrium. CoRR abs/1706.08500 (2017)

8. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
9. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. CoRR abs/1710.10196 (2017)
10. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR abs/1812.04948 (2018)
11. Korkinof, D., Rijken, T., O'Neill, M., Yearsley, J., Harvey, H., Glocker, B.: High-resolution mammogram synthesis using progressive generative adversarial networks. CoRR abs/1807.03401 (2018)
12. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv e-prints arXiv:1411.1784, November 2014
13. Moradi, M., Madani, A., Karargyris, A., Syeda-Mahmood, T.F.: Chest X-ray generation and data augmentation for cardiovascular abnormality classification, p. 57, March 2018
14. Nie, D., et al.: Medical image synthesis with deep convolutional adversarial networks. IEEE Trans. Biomed. Eng. **65**(12), 2720–2730 (2018)
15. Nie, D., et al.: Medical image synthesis with context-aware generative adversarial networks. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 417–425. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_48
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. CoRR abs/1505.04597 (2015)
17. Salimans, T., et al.: Improved techniques for training GANs. In: Lee, D.D., Sugiyama, M., Luxburg, U.V., Guyon, I., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 29, pp. 2234–2242 (2016)
18. Yi, X., Walia, E., Babyn, P.: Generative adversarial network in medical imaging: a review. CoRR abs/1809.07294 (2018)
19. Zhao, H., Li, H., Maurer-Stroh, S., Cheng, L.: Synthesizing retinal and neuronal images with generative adversarial nets. Med. Image Anal. **49**, 14–26 (2018)