Collaborative Adversarial Networks for Joint Synthesis and Segmentation of X-ray Breast Mass Images

1st Tianyu Shen
State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
shentianyu2016@ia.ac.cn

2nd Chao Gou*
School of Intelligent Systems Engineering
Sun Yat-sen University
Guangzhou, China
gouchao@mail.sysu.edu.cn

3rd Jiangong Wang
State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences
Beijing, China
School of Artificial Intelligence
University of Chinese Academy of Sciences
Beijing, China
wangjiangong2018@ia.ac.cn

4th Fei-Yue Wang
State Key Laboratory of Management and Control for Complex Systems
Institute of Automation, Chinese Academy of Sciences
Beijing, China
Qingdao Academy of Intelligent Industries,
Qingdao, China
feiyue.wang@ia.ac.cn

2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9326848

***Abstract*—In this paper, we propose Collaborative Adversarial Networks (CAN) to enable simultaneous forward synthesis and backward segmentation of X-ray breast mass image. The proposed CAN consists of a generator (*G*), an inverter (*I*) and a discriminator (*D*). *G* aims to reconstruct mass images from corresponding annotated masks, while *I* is trained for mapping images back to accurate segmentation masks. All the obtained mask-image pairs are fed to *D* trained in an adversarial learning scheme. Through the collaborative adversarial training using a joint loss function, *G* synthesizes realistic mass images consistent with provided masks and *I* effectively segments the tumor regions from the images. Qualitative and quantitative evaluations on publicly available INbreast database demonstrate the effectiveness of our model. Furthermore, different from conventional GANs-based methods that can only perform either image synthesis or segmentation, the proposed model can be generalized to other bidirectional image-to-image translation of multimodal medical data.**

***Index Terms*—generative adversarial network, medical image synthesis, mass segmentation, X-ray breast mass**

I. Introduction

Breast cancer is one of the most common cancers among women worldwide. Digital X-ray mammography is the most effective tool for early diagnosis of breast cancer, which is essential for the survival of patients. Currently, deep learning (DL) models have led to a significant breakthrough in automatic detection and segmentation of mammograms, which provides desirable assistance in the accurate diagnosis and clinical treatment. However, for DL models, obtaining precisely annotated training dataset remains a challenging work in medical domain. This is mainly because of the protection of patient privacy, the scarcity of corresponding disease and the expense of expert annotation. Therefore, an effective synthesis of realistic X-ray breast mass appearances and a precise segmentation of tumor regions in mammograms have immediate practical significance.

Nevertheless, synthesizing realistic X-ray breast mass images is still challenging due to the variety of mass in terms of texture and shape as well as the presence of intricate and diverse breast tissue surrounding the masses. Recent development of mass image synthesis can be divided into transformation-based and generative model-based approaches. Transformation-based approaches generate new samples by applying operations such as mathematical affine transformation [1] or feature transformation [2] to the existing samples. However, these small mathematical modifications can only provide a little information alteration and the obtained texture distribution can be inconsistent with real mass. On

Research supported by National Natural Science Foundation of China under Grant No.61806198, No.61533019 and the Key Research and Development Program of Guangzhou under Grant 202007050002.
*Corresponding author: Chao Gou.

the other hand, generative model-based approaches leverage on massive training data to acquire the potential distribution of realistic samples and produce new samples from the same distribution [3]. The advantage is that the synthetic samples are similar in appearance to the real ones but possess a variety of properties different from the existing ones. At present, generative adversarial networks (GANs) [4] along with various stabilization methods produce state-of-the-art results in natural and biomedical image synthesis tasks [5].

The main difficulties in X-ray breast mass segmentation are caused by the intense noise and low contrast of mammogram, as well as strong irregularities and ambiguous boundaries of tumor regions. Traditional methods such as region growing, active contour and Markov random field (MRF) for automatic mass segmentation have been gradually replaced by DL models like convolutional neural networks (CNNs) [1], which overcome the shortcomings of artificial designed features. Some mostly used CNN models, such as fully convolutional network (FCN) [6], SegNet [7] and U-Net [8], have produced state-of-the-art performance in biomedical segmentation tasks.

There is little work on a unified model for joint mask-to-image synthesis and image-to-mask segmentation despite their close correlation. To achieve this goal, we present a novel Collaborative Adversarial Networks (CAN) inspired by two recent researches Pix2Pix network [9] and bidirectional-GAN [10]. The motivation is two-fold. Firstly, we aim to synthesize X-ray breast mass images with pixel-wise masks, different from conventional class-specific synthesis that the generated samples are confined to a specified category. To address this issue, we consider the mask-to-image reconstruction inherited from Pix2Pix network [9]. It is worth noticing that mask-to-image is more versatile compared with label-to-image synthesis in medical domain, since it preserves the semantic information and allows the generalization of the framework to different category fields. Secondly, we aim to achieve accurate segmentation of lesion region at the same time. Therefore, we consider the bidirectional idea from BiGAN [10] to generate the masks from images. Our work is different from [9] and [10] because we not only alter the whole network architecture, but also propose an algorithm of parameter learning. In addition, different from BiGAN [10] that decodes the images into latent representation z through feature learning, our model is designed to acquire the masks from images.

To summarize, our main contributions include: (1) Propose a novel Collaborative Adversarial Network (CAN) for joint forward synthesis and backward segmentation of X-ray breast mass image, which is capable of being applied beyond mass images, to any bidirectional image-to-image related fields. (2) Enable synthesizing large amount of mask-annotated X-ray breast mass images with variations in appearance. (3) Enable effective segmentation of the tumor regions from the mass images and achieve better performance compared with existing deep segmentation models. (4) Introduce a joint loss function for the collaborative adversarial training of proposed model.

II. Collaborative Adversarial Networks

As shown in Fig.1, the CAN is composed of a generator (G), an inverter (I) and a discriminator (D), and achieves joint synthesis and segmentation tasks through a collaborative adversarial training. In this section, we first briefly review the Pix2Pix and BiGAN model, then introduce the architecture and learning algorithm of proposed CAN in details.

A. Preliminaries

Pix2Pix Pix2Pix network [9] generalizes the cGANs [11] for various image-to-image translation tasks by altering the condition c to a modality of image x and adding a traditional L1 loss for ensuring the consistency between expected output y and input x. As a result, the generator plays a role in fooling the discriminator, as well as producing output images close to the corresponding ground truth in terms of L1 loss metric. The objective of Pix2Pix can be expressed:

$$\min_G \max_D V(D;G) = \lambda \mathcal{L}_{L1}(G) + \mathbb{E}_{x,y\sim p_{\text{data}}(x,y)}[\log D(x,y)] + \mathbb{E}_{x\sim p_{\text{data}}(x),z\sim p_{\text{data}}(z)}[\log(1-D(x,G(x,z)))] \tag{1}$$

where z represents a random input noise vector.

BiGAN Bidirectional generative adversarial networks [10] provide a bidirectional mapping between the random noise vector z and generated image y by adding an encoder (E) into the original $G-D$ architecture. The G is defined the same as the original GAN, while E models a mapping $E(z;y):\Omega_Y \to \Omega_Z$ and induces a distribution $z\sim p_E(z|y)$. The D is modified to take two kind of data pairs, $(z;G(z))$ obtained from G and $(E(y);y)$ from E, as input to make the predictions $D(z;G(z))$ or $D(E(y);y)$ of real or fake. G and E are inverse of each other and are trained in an adversarial scheme trying to fool the discriminator. The training objective of BiGAN is defined as a new min-max objective:

$$\min_{G;E} \max_D V(D;G;E) = \mathbb{E}_{y\sim p_y}\left[\mathbb{E}_{E(y)\sim p_E(z|y)}[\log D(E(y);y)]\right] + \mathbb{E}_{z\sim p_z}\left[\mathbb{E}_{G(z)\sim p_G(y|z)}[\log(1-D(z;G(z)))]\right] \tag{2}$$

B. CAN Architecture and Learning

In our research, the problems we expect to address are two-fold: (1) How to ensure the authenticity of mass images as well as consistency between masks and images; (2) How to obtain the segmentation mask from mass images simultaneously while achieving better performance compared with a conventional encode-decode segmentation model.

We present the architecture as shown in Fig.1 and a collaborative adversarial learning for solving the above-mentioned problems. Firstly, in order to solve problem (1), we set the segmentation masks as input conditional information, the black pixels of which correspond to the normal breast tissue and white pixels to the lesion region. Thus, the resulted generation is a function over the input mask pixels, which is learned by G. And we design the

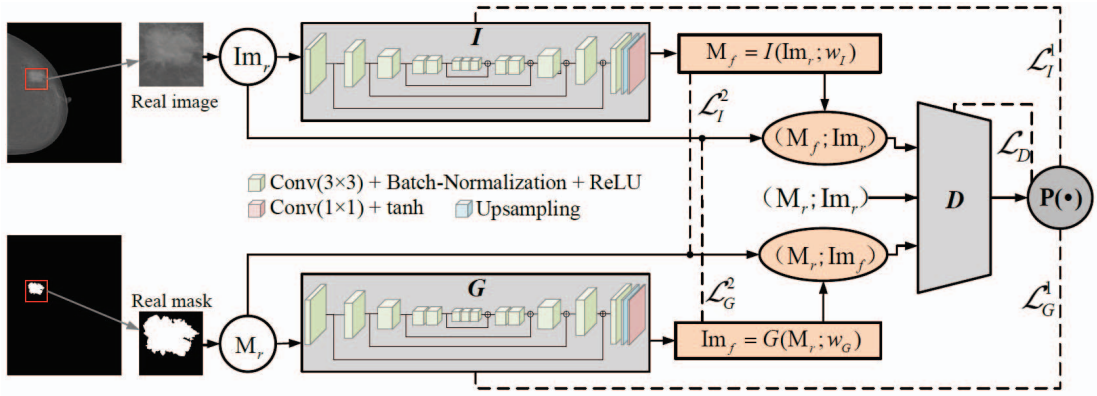

Fig. 1. Proposed CAN architecture.

Algorithm 1: The learning of CAN

Input: $(\mathrm{M}_r; \mathrm{Im}_r)$, batchsize, epoch, learning rate α, weight factors λ_1, λ_2

Initialization: $\{\mathbf{w}_G\}$; $\{\mathbf{w}_I\}$; $\{\mathbf{w}_D\}$;

while *not converge* **do**

- mass image synthesis: $\mathrm{Im}_f = G(\mathrm{M}_r; \mathbf{w}_G)$;
- mass segmentation: $\mathrm{M}_f = I(\mathrm{Im}_r; \mathbf{w}_I)$;
- discriminative probability for mask-image pairs: $P_r = D(\mathrm{M}_r; \mathrm{Im}_r; \mathbf{w}_D)$; $P_f^1 = D(\mathrm{M}_r; \mathrm{Im}_f; \mathbf{w}_D)$; $P_f^2 = D(\mathrm{M}_f; \mathrm{Im}_r; \mathbf{w}_D)$;;
- G learning: $\mathcal{L}_G = \log\left(1 - P_f^1\right) + \lambda_1 \mathbb{E}\left[\|\mathrm{Im}_r - G(\mathrm{M}_r)\|_1\right]$; $\mathbf{w}_G = \mathbf{w}_G - \alpha \frac{\partial \mathcal{L}_G}{\partial \mathbf{w}_G}$;
- I learning: $\mathcal{L}_I = \log\left(1 - P_f^2\right) + \lambda_2 \mathbb{E}\left[\|\mathbf{M}_r - I(\mathrm{Im}_r)\|_2\right]$; $\mathbf{w}_I = \mathbf{w}_I - \alpha \frac{\partial \mathcal{L}_I}{\partial \mathbf{w}_I}$;
- D learning: $\mathcal{L}_D = \log(P_r) + \log\left(1 - P_f^1\right) + \log\left(1 - P_f^2\right)$; $\mathbf{w}_D = \mathbf{w}_D - \alpha \frac{\partial \mathcal{L}_D}{\partial \mathbf{w}_D}$

discriminator D to guarantee not only the authenticity of mass images but also the consistency of the output image with the associated mask. Secondly for problem (2), we incorporate an inverter I, which learns the inverse mapping from mass image samples to condition factors, in the CAN framework. I is trained jointly with G and D using a collaborative adversarial loss function to enable effective learning, as well as improve the generation and segmentation performance through better mode coverage and robustness against mode collapse.

Network Description For G and I network of the CAN, we choose a U-Net-based [8] architecture, whose capability has been proved on biomedical segmentation tasks. As shown in Fig.1, the architecture is composed of encoding and decoding network with skip connections between mirrored layers. The skip connections enable the encoder and the decoder to share information. For the problem we consider, a mass image and its corresponding mask differ in surface appearance and texture, but share the same underlying structure and shape. In the conventional encoder-decoder net, all the information passing through the layers directly results in the loss of low-level features.

D is designed as a Patch-based network inspired from [9] with 16×16 patch. It only needs to determine if each 16×16 patch in an input mask-image pair is real or fake. D is run across the whole mask-image pair by computing a convolution integral. Then all the responses are averaged to produce the ultimate prediction. This operation assumes each patch to be independent in discrimination process and emphasizes the perception on the authenticity of local texture/style instead of the content of whole image.

Parameter Learning As shown in Fig.1, G takes a mask as input condition M_r (r represents $real$) and outputs a generated mass image Im_f (f represents fake): $G(\mathrm{M}_r; \mathrm{Im}_f) : \Omega_{\mathrm{M}} \to \Omega_{\mathrm{Im}}$. On the contrary, I outputs a segmentation mask from an input mass image: $I(\mathrm{Im}_r; \mathrm{M}_f) : \Omega_{\mathrm{Im}} \to \Omega_{\mathrm{M}}$ and induces a new distribution $\mathrm{Im}_r \sim p_I(\mathrm{Im}_r | \mathrm{M}_f)$. Therefore, we construct the mask-image pairs including three forms, which are: $(\mathrm{M}_r; \mathrm{Im}_f)$ obtained from G, $(\mathrm{M}_f; \mathrm{Im}_r)$ from I, and $(\mathrm{M}_r; \mathrm{Im}_r)$ from the matched real mask and image. The obtained mask-image pairs are fed to the D and a probability P of real or fake is predicted.

For D, only the third form $(\mathrm{M}_r; \mathrm{Im}_r)$ is real while the other two forms are fake. D is trained by maximizing the ability of distinguishing real and fake pairs (L_D). For G, the first form $(\mathrm{M}_r; \mathrm{Im}_f)$ is desired to be realistic and the generated image is expected as close as possible to the corresponding ground truth. Thus, G is learned through maximizing the authenticity, expressed as the adversarial loss term (L_G^1), and minimizing the distance between Im_f and corresponding Im_r, expressed as L1 loss term (L_G^2). For I, similar to G, the second form $(\mathrm{M}_f; \mathrm{Im}_r)$ is desired to be realistic (L_I^1) and the distance between M_f and corresponding M_r is expected to be close defined by L2 loss term (L_I^2).

The final joint loss function for the collaborative adversarial

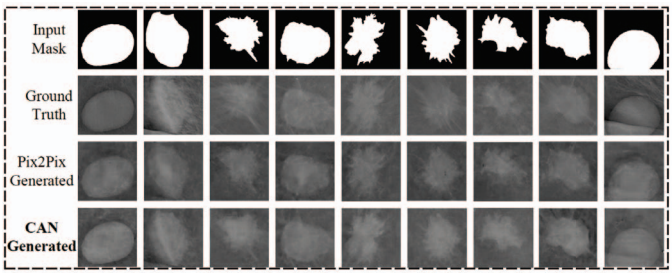

Fig. 2. Visual comparison for real and synthesized images of Pix2Pix and proposed CAN on the INbreast dataset: improvement of the normal synthesized mass images.

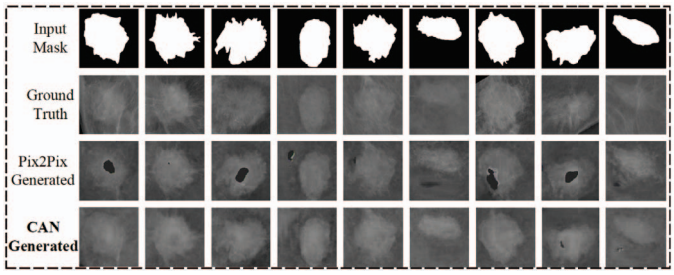

Fig. 3. Visual comparison of Pix2Pix and proposed CAN: improvement of the synthesized mass images with tissue defects.

training is defined as:

$$\begin{aligned}\mathcal{L}_{CAN}(G;I;D) &= \mathbb{E}_{M,\mathrm{Im}\sim p_{data}(\mathrm{M},\mathrm{Im})}[\log D(\mathrm{M},\mathrm{Im})] \\ &+ \mathbb{E}_{\mathrm{M}\sim p_{data}(\mathrm{M})}[\log(1-D(\mathrm{M};G(\mathrm{M})))] + \lambda_1\mathcal{L}_{L1}(G) \\ &+ \mathbb{E}_{\mathrm{Im}\sim \mathrm{p}_{data}(\mathrm{Im})}[\log(1-D(I(\mathrm{Im});\mathrm{Im}))] + \lambda_2\mathcal{L}_{L2}(I)\end{aligned} \tag{3}$$

Overall, the parameter learning of CAN is summarized in Alg.1. When the final convergence is reached, G can map a given mask to the corresponding mass image with the realistic distribution: $\mathrm{Im} = G(\mathrm{M};\mathbf{w}_G)$. And I can acquire the segmentation mask from an input mass image: $\mathrm{M} = I(\mathrm{Im};\mathbf{w}_I)$.

III. Experiments

The INbreast dataset [12] is a public mammogram dataset created by the Breast Research Group, INESC Porto, Portugal. 107 X-ray breast images with their masks are cropped into 256×256 pixels size containing the mass ROIs. Original dataset is augmented by rotating and fiipping to meet the requirement of a large amount of data. A total of 850 samples is obtained. Thereinto, 700 samples are used for training the model and 150 samples are used for testing. The CAN model is implemented using the Keras framework. The parameter settings are as follows: the batchsize is 1, the training epoch is 300, an Adam optimizer is adopted with a learning rate of 0.0002. The weight λ_1 in objective is 10 and λ_2 is 100 through experiment and observation. Since there is little work focusing on simultaneous mass image synthesis and segmentation, separate comparisons with existing methods are conducted in this work.

A. Mask-to-Image Synthesis

For mask-to-image synthesis, we employ qualitative evaluations and compare the results with Pix2Pix [9] under the same experimental settings. It can be observed that the generated images have shape features resembling to the input masks while their texture features are similar to the ground truth images. We achieve comparable results as Pix2Pix as shown in Fig.2. The reason is that both methods adopt similar architecture and training objective for the generator. However, it is observed that CAN presents some slight improvements at the edge of the synthesized tumor region because of the collaborative optimization of D. In addition, our model reduces the tissue defects of synthesized images observably as shown in Fig.3, although there are still slight deficiencies in a few ones.

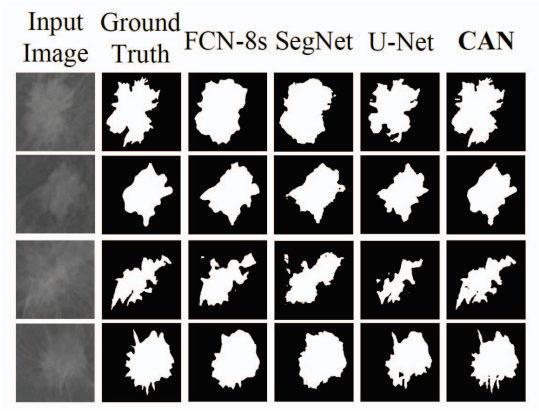

Fig. 4. Qualitative comparison of image-to-mask segmentation.

B. Image-to-Mask Segmentation

The segmentation performance is evaluated quantitatively as shown in Table I and qualitatively as shown in Fig.4. Our method outperforms other segmentation networks and is capable of obtaining better segmentation boundaries and masks due to the adversarial training for I.

More specifically, the CAN gives comparable performance with other models in benign mass segmentation, but gives more desirable performance in malignant tumors, especially for the malignant samples with lobulated or spiculated borders. It is more significant for the accurate segmentation of malignant tumor regions in the clinical applications, because the treatment for malignant breast tumor usually requires an operation to remove the lesions.

TABLE I
Quantitative evaluation for image-to-mask segmentation.

Models	Measurement metrics (%)			
	Acc.	Dice	Jac.	MCC
FCN-8s [6]	93.4	91.0	83.8	85.8
SegNet [7]	93.2	90.4	82.9	85.1
U-Net [8]	93.1	90.5	83.0	85.2
Proposed CAN	**94.1**	**91.7**	**85.1**	**87.2**

IV. Conclusion

To summarize, we propose a Collaborative Adversarial Network with joint training for simultaneous end-to-end synthesis and segmentation of X-ray breast mass image. For mask-to-image synthesis, we achieve desirable results with fewer defects in synthesized images using a more generalized framework, as compared to Pix2Pix network. For image-to-mask segmentation, our model outperforms other segmentation models due to the addition of adversarial loss. Furthermore, the proposed model can be generalized beyond mass images to other bidirectional image-to-image translation fields.

References

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[2] L. Maaten, M. Chen, S. Tyree, and K. Weinberger, "Learning with marginalized corrupted features," in *International Conference on Machine Learning*, 2013, pp. 410–418.

[3] D. Nie, R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen, "Medical image synthesis with context-aware generative adversarial networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 417–425.

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

[5] E. Wu, K. Wu, D. Cox, and W. Lotter, "Conditional infilling gans for data augmentation in mammogram classification," in *Image Analysis for Moving Organ, Breast, and Thoracic Images*. Springer, 2018, pp. 98–106.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[7] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[8] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[9] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[10] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *arXiv preprint arXiv:1605.09782*, 2016.

[11] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[12] I. C. Moreira, I. Amaral, I. Domingues, A. Cardoso, M. J. Cardoso, and J. S. Cardoso, "Inbreast: toward a full-field digital mammographic database," *Academic radiology*, vol. 19, no. 2, pp. 236–248, 2012.