Tunable CT Lung Nodule Synthesis Conditioned on Background Image and Semantic Features

Ziyue Xu(✉), Xiaosong Wang, Hoo-Chang Shin, Holger Roth, Dong Yang, Fausto Milletari, Ling Zhang, and Daguang Xu

Nvidia Corp., Bethesda, USA
ziyuex@nvidia.com

Abstract. Synthetic CT image with artificially generated lung nodules has been shown to be useful as an augmentation method for certain tasks such as lung segmentation and nodule classification. Most conventional methods are designed as "inpainting" tasks by removing a region from background image and synthesizing the foreground nodule. To ensure natural blending with the background, existing method proposed loss function and separate shape/appearance generation. However, spatial discontinuity is still unavoidable for certain cases. Meanwhile, there is often little control over semantic features regarding the nodule characteristics, which may limit their capability of fine-grained augmentation in balancing the original data. In this work, we address these two challenges by developing a 3D multi-conditional generative adversarial network (GAN) that is conditioned on both background image and semantic features for lung nodule synthesis on CT image. Instead of removing part of the input image, we use a fusion block to blend object and background, ensuring more realistic appearance. Multiple discriminator scenarios are considered, and three outputs of image, segmentation, and feature are used to guide the synthesis process towards semantic feature control. We trained our method on public dataset, and showed promising results as a solution for tunable lung nodule synthesis.

1 Introduction

Among the three major factors enabling the success of deep learning - data, algorithm, and computation power, data covering sufficient population distribution is often most critical and most difficult to achieve. This is especially true for medical image domain, in which labeled data availability is limited by its unique characteristics: (1) medical images often involves high cost to produce, and sensitivity in sharing; (2) pathological cases can have large variability in appearances, and are often unbalanced/long tail in distribution; (3) accurate labeling of the data requires high professional expertise, and can nevertheless have large inter- and intra- observer variability even among experts.

Therefore, current work in medical domain mostly relies on using labeled large public datasets [1], automated and/or semi-automated methods [5], and

N. Burgos et al. (Eds.): SASHIMI 2019, LNCS 11827, pp. 62–70, 2019.
https://doi.org/10.1007/978-3-030-32778-1_7

existing clinical report mining [12]. Recently, the development of generative adversarial networks (GAN) [2] has enabled a promising way in data augmentation: generate realistic synthetic data for training purpose. Preliminary works along this direction has demonstrated the potential of such approach in lung segmentation [6], brain tumor segmentation [11] and lung nodule classification [13].

Although shown to be promising, current GAN-based methods generate synthetic images based on limited information such as segmentation [8] and surrounding images [6]. Few recent works investigated finer control over the synthesis process, for example, controlling the malignancy property of the generated lung nodules [13]. However, to our best knowledge, there is no prior work that has the capability of controlling the semantic features of the synthesized nodules.

Meanwhile, most of previous methods model the synthesis process as an "inpainting" problem, in which a portion of the background image is removed before inpainting the synthesized nodule. One shortcoming of such model is that the fusion between synthetic region and background image may not be natural. To address this challenge, previous work used multi-mask reconstruction loss [6], or decoupled mask-appearance generation [8]. However, since the original information is lost in the background image input, it is difficult to recover the spatial continuity, even with the proposed methods.

In this work, we develop a 3D multi-conditional GAN model learning the shape and appearance distributions of lung nodules related to semantic features in 3D space. We aim to generate not only realistic but also tunable nodules according to its semantic features. Hence, our GAN is conditioned on both surrounding background information and a controllable feature set. In order to ensure a natural fusion with background image, we use two outputs of image and its corresponding nodule mask to reinforce the blending of the two, rather than erasing the region from base image. Multiple generator and discriminator losses are used to guide the network towards controlling the semantic feature inputs. We apply our strategy to public lung nodule dataset of LIDC [1], where each nodule is linked with a series of semantic annotations describing its appearances.

This work's main contributions are: (1) we synthesize 3D lung nodules and control its properties by using a 3D multi-conditional GAN with both surrounding images and semantic features; (2) instead of inpainting, we address the object/background fusion by multi-output and fusion block within network design; (3) both feature learning and fusion learning are performed by designing their corresponding outputs and losses during network training.

2 Method

To address the challenges of (1) incorporating semantic features, and (2) object/background fusion, inspired by works for 2D natural image synthesis [7,10], we design our network as a 3D multi-conditional GAN with style specification by additional regression branch. The generator takes in two conditions

of background image and semantic feature, and produces three outputs of synthetic image, nodule mask, and predicted feature. The object/background fusion is performed with fusion blocks at each resolution level. The inter-relationships among background, semantic feature, and target nodule are controlled via multiple losses from generator and discriminator. Figures 1 and 2 depicts an overview of our method. Below, we outline the GAN architecture, loss function design, and training strategy for learning appearance together with the semantic features.

2.1 GAN Architecture

Figure 1 illustrates the structure of the proposed generator. Background image is encoded via a series of convolutional layers with three resolution levels, each downsampling doubles the feature channel. The semantic features are transformed via a fully connected layer and reshaped to bottleneck image size. The blending of object (nodule) and background image is performed via fusion block.

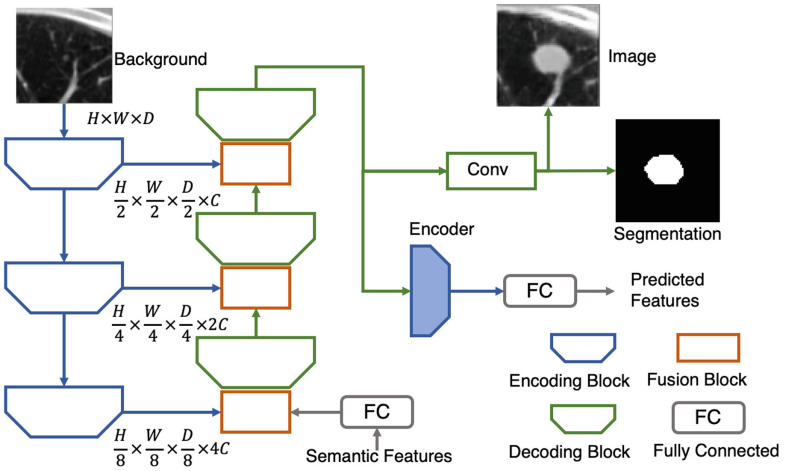

Fig. 1. Proposed generator of the 3D multi-conditional GAN for tunable nodule synthesis. Generator utilizes both background image and semantic feature code to synthesize image, nodule segmentation, and also a regression branch for feature code prediction.

As shown in Fig. 2, following [10], the fusion block is designed so that half of the object code is used to control the "soft" merging of the two feature sets in order to produce the synthetic image and its corresponding segmentation mask. Such fusion is enforced by the prediction of segmentation mask as an auxiliary output during training. As compared with "inpainiting", this strategy performs better in natural blending of the object/background. Also, the mask output is potentially helpful for data augmentation in tasks such as detection and segmentation.

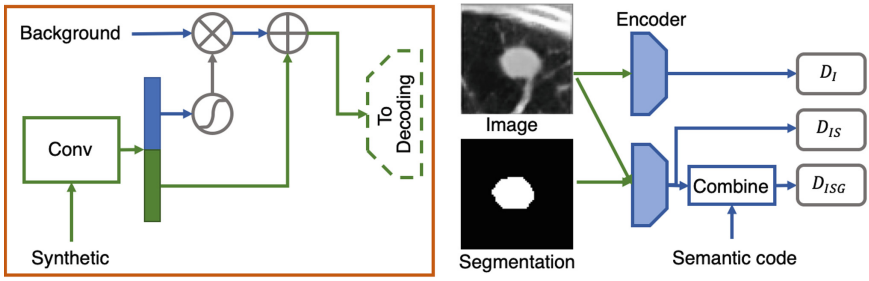

Fig. 2. Fusion block and discriminator of the 3D multi-conditional GAN. Left: fusion block at each resolution layer helps to fuse the information from background with that from previous layer. Right: with image, segmentation, and feature code, discriminator distinguishes three types of real/fake scenarios (Sect. 2.2).

To address the challenge of semantic feature specification, in addition to discriminator pairing, we added a regression branch (Fig. 1) beyond synthetic image and mask generation. Specifically, a encoding block is added to the output layer of the generator followed by fully connected layer to predict the vector of semantic features from the synthesized feature map. Furthermore, to control the size of the generated nodule, a loss is computed from the size of mask prediction in comparison with that of the ground truth segmentation of training data.

Figure 3 shows a result example for the proposed GAN from three views. The second column is the weighting mask from the last fusion block. It can be observed that the nodule and background are naturally separated and fused with the proposed fusion block and network.

2.2 Loss Functions and Training Strategy

The proposed GAN synthesize nodule with segmentation according to semantic feature vector. In order to guide the training process, several losses are proposed to supervise different aspects of the network.

The discriminator is illustrated in Fig. 2. The input to the discriminator is a tuple of image-segmentation-semantic feature code. Two encoders are utilized to encode: (1) image for discriminator D_I, and (2) image-segmentation pairs for discriminator D_{IS}. The second encoder's output is further combined with feature code f and further encoded via convolution, batch normalization, and leaky ReLU activation layers for discriminator D_{ISG}. Discriminators are trained with least squares loss functions [9]. Given image x, matched semantic feature code f, and matched segmentation mask m, tuples to be discriminated against it include cases containing mismatched feature code $\bar{f}$, mismatched segmentation mask $\bar{m}$, synthetic image G_x, and synthetic mask G_m. Let p_d and p_G denote the distributions of real and synthetic data, we have $x, f, m, \bar{f}, \bar{m} \sim p_d$ and

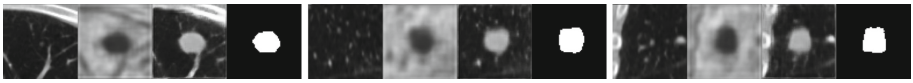

Fig. 3. Example of results produced by proposed synthesis GAN from three views: left to right - background image, background weight image during fusion, synthesized nodule image, and output segmentation mask.

$G_x, G_m \sim p_G$. With different combinations, we have

$$\begin{aligned} L_{D_I} &= \mathbb{E}[(D_I(x)-1)^2] + \mathbb{E}[D_I(G_x)^2] \\ L_{D_{IS}} &= \mathbb{E}[(D_{IS}(x,m)-1)^2] + \mathbb{E}[D_{IS}(x,\bar{m})^2] + \mathbb{E}[D_{IS}(G_x,G_m)^2] \\ L_{D_{ISG}} &= \mathbb{E}[(D_{ISG}(x,m,f)-1)^2] + \mathbb{E}[D_{ISG}(x,\bar{m},f)^2] \\ &\quad + \mathbb{E}[D_{ISG}(x,m,\bar{f})^2] + \mathbb{E}[D_{ISG}(G_x,G_m,f)^2] \end{aligned}$$

For training the generator, in addition to discriminator loss, we further reinforced background reconstruction, semantic feature prediction, and size control with their corresponding losses. Let $G_{\bar{M}}$ be a morphological eroded version of segmentation mask G_m's inverse (i.e. background region), $\odot$ denote element-wise multiplication. The background reconstruction loss $L_{G_{BG}}$ is formulated as the $L1$ loss over background between synthetic image G_x and base image x. The semantic feature prediction loss L_{G_F} and size loss L_{G_S} are formulated as the $L2$ loss between predictions G_f, G_s and ground truth f, s, where $G_s = \sum(G_m > 0)$ and $s = \sum(m > 0)$

$$\begin{aligned} L_{G_{BG}} &= \mathbb{E}[\|G_x \odot G_{\bar{M}} - x \odot G_{\bar{M}}\|_1] \\ L_{G_F} &= \mathbb{E}[\|G_f - f\|_2] \\ L_{G_S} &= \mathbb{E}[\|G_s - s\|_2] \end{aligned}$$

With all the proposed losses, the generator loss is

$$\begin{aligned} L_G &= \mathbb{E}[(D_I(G_x)-1)^2] + \mathbb{E}[(D_{IS}(G_x,G_m)-1)^2] \\ &\quad + \mathbb{E}[(D_{ISG}(G_x,G_m,g)-1)^2] + \lambda_1 L_{G_{BG}} + \lambda_2 L_{G_F} + \lambda_3 L_{G_S} \end{aligned}$$

3 Experiment and Result

We evaluate the proposed method using the publicly available LIDC dataset [1]. This dataset contains 1018 chest CT scans of patients with lung nodules. There are 9 semantic features for each nodule: subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture, and malignancy. Additionally, we can calculate the volume V of each nodule's manual segmentation, and estimate its diameter using sphere model $d = \sqrt[3]{6V/\pi}$. For this work, we select a subset of all nodules with approximate diameter between 3 mm and 30 mm

following clinical standard of micro-nodule (<3 mm) and mass (>30 mm) [3]. In total there are 5942 semantic records from 826 patients. Note that multiple records can be related to the same nodule, as a single nodule can be annotated by several experts. Therefore, the annotation inherently contains certain amount of variability/noise. A $60 \times 60 \times 60$ mm^3 volume-of-interest (VOI) centered at each nodule is first cropped from the original image, then resampled to a fixed size of $64 \times 64 \times 64$.

To generate background image, we first: (1) segment the lung region of each CT volume using [4] from the whole CT volume; (2) make binary union of all manual nodule segmentations; and (3) exclude the nodule mask from lung mask. Hence there will be no nodule presence within the resulting mask after step (3), so that "painting nodule over existing nodule" can be avoided. Next, distance transform is computed from this mask, and centers for 3D background VOI patches are selected at a random location 5 to 25 mm from the mask boundary. The VOIs of the same size as nodule cases are cropped and resized to a fixed size of $64 \times 64 \times 64$.

The aim of our proposed method is to (1) generate realistic nodules and natural blending with the specified background, and (2) control the nodule appearance with semantic features.

Figure 4 shows the performance of image synthesis with multi-conditional GAN. As shown in the image, based on random nodule-free background B, the proposed method generates realistic images D, which reflects the semantic features as the reference training samples A (clear/fuzzy boundary, solid/ground-glass, etc.). As comparison, we implemented a 3D version of baseline [10], although it also have feature vector matching during discriminator phase, it failed to achieve same level of semantic feature control without the help of regression branch.

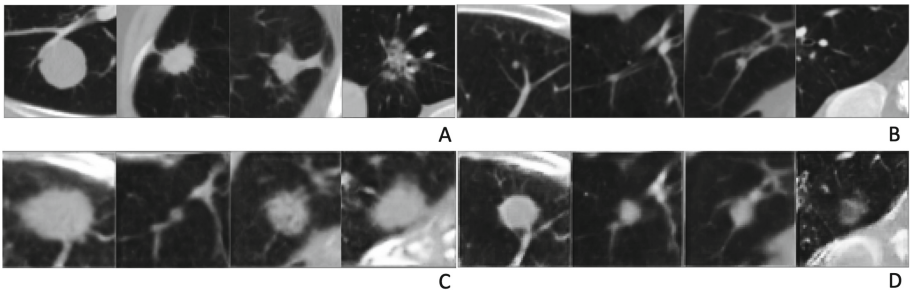

Fig. 4. Result of nodule synthesis, A: 4 different training image, B: random nodule-free background image, C: synthetic image generated by 3D version of baseline method [10], and D: synthetic images generated by the proposed method. Note that A, C, and D shared the same semantic features.

Figure 5 shows the synthesis result using the same background image with various semantic features and sizes. Two sets of examples are given under three

views, and one additional result on changing size only is presented with one view (last row). As shown in the image, the proposed method has the capability of generating various nodules from a background image under different configurations of semantic features and sizes. From the result, we can observe some distortion of the background image, especially for the ground-glass, heterogeneous case of the last column due to its challenging nature. Last row shows the change with small to large size parameters. We observe that although the size changed as expected, they are not very accurate with regard to the real "expected" size (as input parameter). Therefore, potential improvements and future work include the investigation into annotation uncertainty/correlation among semantic features, better network structure design for higher quality image and more accurate control, and application to other tasks as data augmentation.

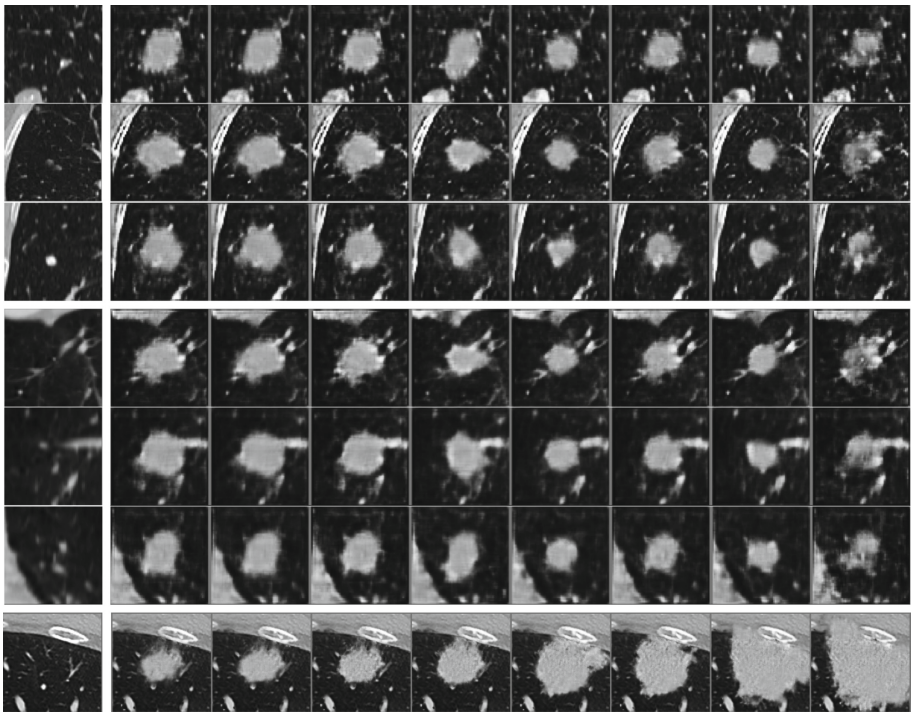

Fig. 5. Two sets of results of nodule synthesis based on the same background image under different semantic features and sizes, three views are provided. First column is the background image, and the following columns are synthetic cases, each column using a semantic feature/size combination. Last row showed an experiment of changing size parameter only.

4 Conclusion

We use a multi-conditional GAN, coupled with fusion structure, multiple outputs, and loss functions, to effectively generate realistic nodules with control over appearance by semantic features and size. Without erasing any portion of condition image, the proposed method achieves realistic nodule generation and smooth background fusion. The tunable size and semantic features ensures further diversified and targeted data augmentation. Current results showed promising diversity, however, more vigorous study is needed to verify their actual "controllability" over the image generation. As such, our approach can provide a potentially effective means for nodule image sample generation.

References

1. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
2. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27, pp. 2672–2680 (2014)
3. Hansell, D.M., Bankier, A.A., MacMahon, H., McLoud, T.C., Mller, N.L., Remy, J.: Fleischner society: glossary of terms for thoracic imaging. Radiology **246**(3), 697–722 (2008)
4. Harrison, A.P., Xu, Z., George, K., Lu, L., Summers, R.M., Mollura, D.J.: Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 621–629. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_71
5. Jin, D., Xu, Z., Harrison, A.P., George, K., Mollura, D.J.: 3D convolutional neural networks with graph refinement for airway segmentation using incomplete data labels. In: Wang, Q., Shi, Y., Suk, H.-I., Suzuki, K. (eds.) MLMI 2017. LNCS, vol. 10541, pp. 141–149. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_17
6. Jin, D., Xu, Z., Tang, Y., Harrison, A.P., Mollura, D.J.: CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 732–740. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_81
7. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR abs/1812.04948 (2018)
8. Liu, S., et al.: Decompose to manipulate: manipulable object synthesis in 3D medical images with structured image decomposition. CoRR abs/1812.01737 (2018)
9. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: Least squares generative adversarial networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2813–2821, October 2017
10. Park, H., Yoo, Y., Kwak, N.: MC-GAN: multi-conditional generative adversarial network for image synthesis. In: The British MachineVision Conference (BMVC) (2018)

11. Shin, H.-C., et al.: Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: Gooya, A., Goksel, O., Oguz, I., Burgos, N. (eds.) SASHIMI 2018. LNCS, vol. 11037, pp. 1–11. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00536-8_1
12. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017
13. Yang, J., et al.: Class-aware adversarial lung nodule synthesis in CT images. CoRR abs/1812.11204 (2018)