Medical (CT) Image Generation with Style

Arjun Krishna[a], **Klaus Mueller**[a]
[a]Stony Brook University, Computer Science Department

Abstract. We propose the use of a conditional generative adversarial network (cGAN) to generate anatomically accurate full-sized CT images. Our approach is motivated by the recently discovered concept of *style transfer* and proposes to mix style and content of two separate CT images for generating a new image. We argue that by using these losses in a style transfer based architecture along with a cGAN, we can increase the size of clinically accurate, annotated datasets by multiple folds. Our framework can generate full-sized images with novel anatomy at spatial high resolution for all organs and only requires limited annotated input data of a few patients. The expanded datasets our framework generates can then be utilized within the many deep learning architectures designed for various processing tasks in medical imaging.

1 Introduction

Deep learning has shown great promise for a myriad of applications in CT imaging such as improving image quality in low dose acquisition, cross modality translations etc. However training of deep networks requires an abundance of clinical training data. This remains a challenge due to scarcity/privacy issues and the high interpatient anatomical variability. Also, often these datasets are not comprehensively annotated, owing to the costliness and scarcity of expert annotation in the medical domain. Hence we present an approach to increase the training data multiple folds with as few as ten training samples. Our framework also ensures that the full-sized generated CT images are anatomically correct and contain enough anatomical variation from training data.

Our method is inspired by the work of Zhao et al.[1] which simulates retinal fundus images and neuronal images with a cGAN[2] by conditioning on segmented filamentary ground truths. The main difference between their setup and ours is that the retinal image generation process through style transfer is fairly straightforward due to the near uniform texture of a retinal image. The content loss along with the adversarial loss ensures that filamentary structure is respected while the style loss transfers the texture of the style image onto the generated image. Conversely, CT images consist of different structures with different textures. The overall style loss of the entire image will not work since the style changes dramatically from one part of the image to the next. Our paper describes our solution to this challenge. To the best of knowledge, our proposed approach is the first attempt to incorporate style transfer techniques involving a cGAN for the synthesis of new full-sized CT images with correct and varied anatomical features of all organs.

2 Methods

In our approach, the generator G_θ is viewed as a mapping function from the segmentation map of a CT image y to a plausible CT image x inheriting the style of a particular style image x_s, which is another CT image from a different patient, on a per-organ basis. For example, the heart of the generated CT image will match in appearance the heart of the style image; likewise for lungs, spinal cord, bones, etc. Let us denote $\mathrm{x} \in \mathrm{R}^{WxH}$ as the gray-scale CT image to be generated, while $\mathrm{y} \in \{0, 50, 100, 150, 200, 230\}^{WxH}$ is the segmentation map of a real image x_o, given as conditional input to the generator. Let $\mathrm{G}_\theta : (\mathrm{y} \in \mathrm{R}^{WxH}, \mathrm{z} \in \mathrm{R}^{Z}) \rightarrow \mathrm{x} \in \mathrm{R}^{WxH}$ denote the image generation function that takes a segmentation map image y and a noise code z as input to produce a CT image x of a particular style. As shown in Fig. 2, the numerical values 0, 50, 100... are gray-scale values denoting the segmented regions of the corresponding organs. The style image x_s will be used in calculating the style loss over the generated image whereas the original image x_o (corresponding to the segmentation map) will be used for determining the content loss. Hence, x_o determines the new anatomy and x_s makes sure it has the proper texture appearance. Our network architecture is shown in Fig. 1.

Our contributions are: (1) We define a mapping function G learned from a very small training set $\{(\mathrm{x}_i, \mathrm{y}_i)\}_{i=1}^{n}$ for a style image; (2) We can obtain distinct plausible gray-scale CT image instances by varying the noise code z for a particular style and segmentation; (3) Our method facilitates the generation of even more anatomical variations in the generated images by perturbing the boundaries of the segmentation map by some input, whereby the correctness of these variations is ensured by the style transfer of the corresponding parts of the style CT image.

2.1 Generator and Discriminator

Our method employs an encoder-decoder strategy for the generator.[3,4] Since we are using a cGAN architecture, the segmentation map (y) along with the noise vector z, a 200-dimensional random code, are given as inputs. The noise code z is fully connected to the first layer, which is then reshaped. We use kernel size 4 and stride 2 without any pooling layer for all layers of G and D. The basic network architecture proposed in[5] is followed to build the layers

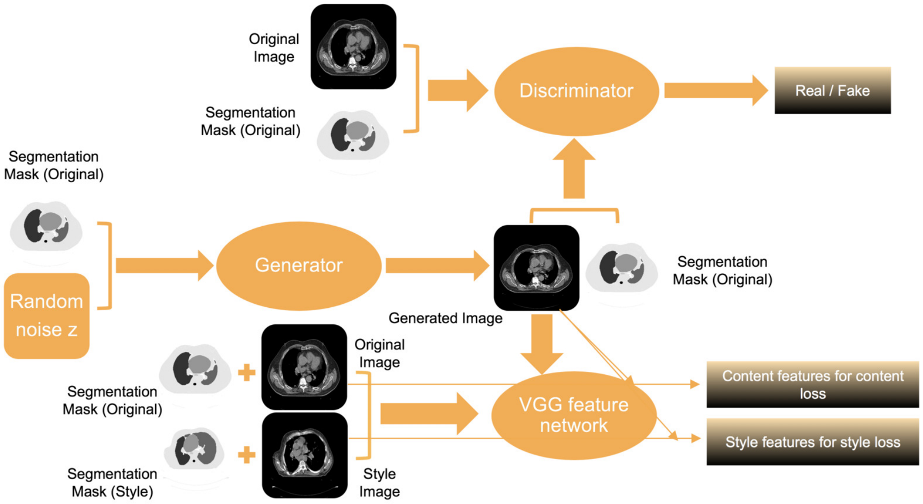

Original Image
Segmentation Mask (Original)
Discriminator
Real / Fake
Segmentation Mask (Original)
Generator
Random noise z
Segmentation Mask (Original)
Generated Image
Original Image
Segmentation Mask (Original)
VGG feature network
Content features for content loss
Style features for style loss
Segmentation Mask (Style)
Style Image

Fig 1 Framework for training style transfer of every segment of style image over input segmentation map.

of the generator with multiple Convolution-BatchNormLeakyRelu components. The activation function of the output layer of G is tanh to squash the value between - 1 and 1. On par with our generator, the same ConvolutionBatchNorm-LeakyRelu building blocks are used in building our discriminator. The activation function of the output layer of D is a sigmoid function. An Adam optimizer with mini-batch of size 1[6] and Stochastic gradient descent are employed to update the parameters (θ of G and γ of D) during the training process.

2.2 *Loss Functions*

Since the mapping function G has to be learned with respect to a style image x$_s$, perceptual losses are needed to be optimized for learning that function, i.e. we need to add style and content loss as part of the generator G. The motivation for the aforementioned perceptual losses for our experiments is to transfer anatomically correct texture and content (location, size) for each synthesized part of the generated CT image. The content loss, on the other hand, ensures anatomical correctness while the style of another patient provides texture variation for the generated CT image.

We use a VGG-19[7] convolutional neural network (CNN) to extract features from its multiple layers.[8] For style, we extract features from the RELU activations of the first layer and every other RELU layer that succeeds the pooling layer in the VGG. We extract these features for both the style image (x$_s$) and the generated image (x) to calculate the style loss. Similarly for the content loss we extract feature maps from the RELU activations of the tenth layer from both the original image (x$_o$) and the generated image (x).

Let $\phi^l(\cdot)$ be the function implemented by the part of the VGG-19 network from the input up to layer l, and let O^l, S^l, and R^l denote the feature maps extracted from the VGG at layer l, for the original image x$_o$, the style image and the stylized image x, respectively.

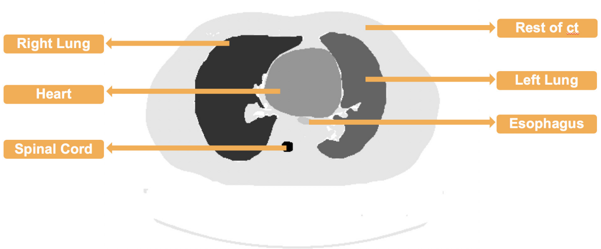
$$O^l = \phi^l(x_o), S^l = \phi^l(x_s), R^l = \phi^l(x) \tag{1}$$

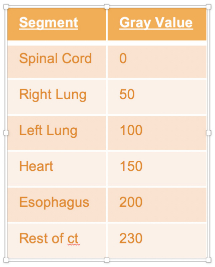
Let the dimensionality of these feature maps be N$_l$ $\times M_l$ and let R$^l_{ij}$ and O$^l_{ij}$ be the jth position of filter i in layer l of the network. Then the content loss at layer l as per Gatys et al.,[9] can be defined as:

$$L^l_c(x, x_o) = \frac{1}{2N_l M_l} \|\phi^l(x) - \phi^l(x_o)\|^2_F = \frac{1}{2N_l M_l} \sum_{i,j} |R^l_{ij} - O^l_{ij}|^2 \tag{2}$$

As per Gatys, et al.,[9] the style of an image if measured from layer l, consists of the correlations between the different feature responses which can be encoded into the Graham Matrices. We can define this for style image as:

$$G(S)^l_{ij} = S^l_{i*}.S^l_{j*} = \sum_{k=1}^{M_l} S^l_{ik} S^l_{jk} \tag{3}$$





Segment	Gray Value
Spinal Cord	0
Right Lung	50
Left Lung	100
Heart	150
Esophagus	200
Rest of ct	230

Fig 2 Left: An example of a segmentation map of a CT image. Right: Fixed Pixel values used to denote the segment.

As such the style loss at layer l would be:

$$L_s^l(x, x_s) = \frac{1}{4N_l^2 M_l^2} \|G(R)^l - G(S)^l\|_F^2 = \frac{1}{4N_l^2 M_l^2} \sum_{i,j} |G(R)_{ij}^l - G(S)_{ij}^l|^2 \tag{4}$$

Since medical images are generally spatially smooth in texture we add total variation loss to L$_G$ as well.

$$L_{tv}(x) = \sum_{w,h} (\|x_{w,h+1} - x_{w,h}\|_2^2 + \|x_{w,h+1} - x_{w,h}\|_2^2) \tag{5}$$

As mentioned earlier, each segment of the CT image has its own style. Hence we use the segmentation masks of each image, namely style image and generated image, to extract the style features of each segmented region separately and thereby calculate the style loss of each segmented region. Hence the total style loss is:

$$L_s(x, x_s) = \sum_{sg \in SG} \sum_{l \in SL} \frac{1}{4N_l^2 M_l^2} \|RR_{sg}^l - SS_{sg}^l\|_F^2 \tag{6}$$

where SG is the set of all six segments in the images and SL is the set of all style layers in VGG net As per above the content loss is given by:

$$L_c(x, x_o) = \sum_{sg \in SG} \sum_{l \in CL} \frac{1}{2N_l M_l} \|R_{sg}^l - O_{sg}^l)\|_2^2 \tag{7}$$

where CL is the set of all content layers. We will calculate these perceptual losses by considering features of the aforementioned layers and combine them with the adversarial loss of the generator when generating an image. The loss for G becomes:

$$L_G(G_\theta) = -\sum_i \log D_\gamma(G_\theta(y_i, z_i), y_i) + w_s L_s + w_c L_c + w_{tv} L_{tv} \tag{8}$$

For the discriminator nothing extra needs to be added; hence the loss includes the adversarial component only. The loss for D is:

$$L_D(D_\gamma) = -\sum_i \log D_\gamma(x_i, y_i) + \log(1 - D_\gamma(G_\theta(y_i, z_i), y_i)) \tag{9}$$

It is clear that by training our cGAN to optimize the above loss functions, the style transfer of every organ of the CT image will contribute to the texture of the corresponding part of the generated CT image.

2.3 Dataset and Implementation

We downloaded a set of normal-dose lung CT images from the Cancer Imaging Archive[10] of 12 patients, each of size of 512 × 512. We manually selected 1-2 images from a similar anatomical position of each patient to obtain 20 images. The proposed neural network was implemented using the Tensorflow deep learning library in a Python3 environment on the Google Cloud Platform. All experiments were performed using an NVIDIA Tesla P100 graphics card with 30 GB RAM. The training was done for 100 epochs.

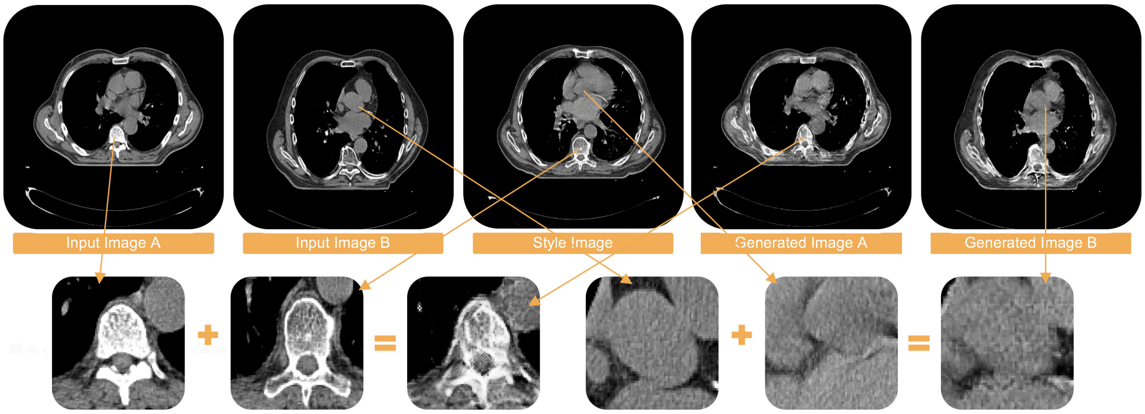


Fig 3 First Row: The last two images are generated corresponding to the segmentation maps of the first two images. Second Row: Segments are zoomed to show the effect of style transfer on input images (content image + style image = result image)

3 Results and Conclusion

Figure 3 shows some first results we obtained with our method. We observe that the generated CT images preserve the segmented input anatomy (the organs), while exhibiting different yet realistic-looking texture appearances throughout the image. For the generated image A, for example, we observe that the spinal cord is now completely surrounded by bone as in the style image, while the shape and location of the spine is similar to input image A. The spine texture is also similar to that of the style image. On the other hand, for the generated image B, we observe that the texture is more dense, similar to the style image (i.e., the dark areas between the different structures are reduced), while the regions themselves are more like input image B. All this suggests that our generation model can capture such intrinsic features without explicit human interventions for conveying such prior knowledge.

We presented a first implementation of a promising new cGAN based approach to synthesize large (512x512) CT images given a segmentation map. The synthesized images look realistic, possess acceptable anatomical detail and texture variations. Moreover, the model is capable of learning from small training sets of as few as 10-20 examples. Future work will build on our method and study how to create perturbations to existing segmentation maps or make new ones, so as to generate more anatomically diverse CT images. By applying a pre-trained segmentation network we could use readily available non-annotated CT images for style transfer, adding a lot more data with different texture patterns. We also intend to design a more sophisticated scheme to mix texture and content in varied ratios in different parts of the CT. Finally, we also plan to conduct user studies with clinicians to verify and further refine our method.

References

1 H. Zhao, H. Li, and L. Cheng, "Synthesizing filamentary structured images with gans," *CoRR* **abs/1706.02185** (2017).

2 M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR* **abs/1411.1784** (2014).

3 X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *European Conference on Computer Vision*, 318–335, Springer (2016).

4 X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in neural information processing systems*, 2802–2810 (2016).

5 A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434* (2015).

6 D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).

7 K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556* (2014).

8 J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 694–711, Springer (2016).

9 L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv:1508.06576* (2015).

10 K. Clark, B. Vendt, K. Smith, *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging, vol. 26, no. 6, pp. 10451057* (2013).