CS²: A Controllable and Simultaneous Synthesizer of Images and Annotations with Minimal Human Intervention

Xiaodan Xing[1], Jiahao Huang[1,5], Yang Nan[1], Yinzhe Wu[1,2], Chengjia Wang[3], Zhifan Gao[4], Simon Walsh[1], and Guang Yang[1,5(✉)]

[1] National Heart and Lung Institute, Imperial College London, London, UK
`g.yang@imperial.ac.uk`
[2] Department of Biomedical and Engineering, Imperial College London, London, UK
[3] Edinburgh Centre for Robotics, Heriot-Watt University, Edinburgh, UK
[4] School of Biomedical Engineering, Sun Yat-sen University, Guangdong, China
[5] Cardiovascular Research Centre, Royal Brompton Hospital, London, UK

Abstract. The destitution of image data and corresponding expert annotations limit the training capacities of AI diagnostic models and potentially inhibit their performance. To address such a problem of data and label scarcity, generative models have been developed to augment the training datasets. Previously proposed generative models usually require manually adjusted annotations (e.g., segmentation masks) or need pre-labeling. However, studies have found that these pre-labeling based methods can induce hallucinating artifacts, which might mislead the downstream clinical tasks, while manual adjustment could be onerous and subjective. To avoid manual adjustment and pre-labeling, we propose a novel controllable and simultaneous synthesizer (dubbed CS²) in this study to generate both realistic images and corresponding annotations at the same time. Our CS² model is trained and validated using high resolution CT (HRCT) data collected from COVID-19 patients to realize an efficient infections segmentation with minimal human intervention. Our contributions include 1) a conditional image synthesis network that receives both style information from reference CT images and structural information from unsupervised segmentation masks, and 2) a corresponding segmentation mask synthesis network to automatically segment these synthesized images simultaneously. Our experimental studies on HRCT scans collected from COVID-19 patients demonstrate that our CS² model can lead to realistic synthesized datasets and promising segmentation results of COVID infections compared to the state-of-the-art nnUNet trained and fine-tuned in a fully supervised manner.

Keywords: Generative model · Semi-supervised segmentation · Data augmentation


---

1   Introduction

Medical images with ample annotations are difficult to obtain. This paucity of medical images and labels compromises the performance of AI-based computer-aided diagnosis, including medical image segmentation and classification. Medical image synthesis [3,4] provides an effective and practical strategy for data augmentation that can mitigate the problem of data and labeling scarcity. Generative Adversarial Networks (GAN) [6] is a potent paradigm for realistic medical image synthesis, which can be potentially used for effective data augmentation. In medical image synthesis, three types of GANs are widely used:

The first type of model is called vector-to-image (V2I) GANs, which synthesize augmented images from a vector of noises (Fig. 1 (a)). A recent study [21] used a V2I styleGAN model to produce annotated augmented images from a vector of noises without a pre-labeled dataset. By labeling a handful of those synthesized images and re-inferencing their styleGAN, augmented images with annotations could be obtained. However, the generation of anatomy and the location of the lesions could not be controlled in these V2I models, and the explainability of these generative models was limited, which might sabotage following clinical interpretation and decision making.

The second type is called mask-to-image (M2I) GANs (Fig. 1 (b)). The annotations of these algorithms could be manually adjusted, i.e., combining existing segmentation masks [19,20], or transferred from a labeled dataset in another domain [9]. Unfortunately, fully paired mask-to-image synthesis reduces the variance of synthetic data. Moreover, these models require large human input to feed them with manual pre-labeling, conflicting with our primary aim to alleviate manual labeling workload. Besides, cross-modality models require the labeled dataset in another domain, which might not always be available due to exorbitant costs of additional scanning and patients' physiological limitations.

The third type is vector-to-mask-to-image (V2M2I) GANs. To obtain the segmentation mask, a particular synthesis network in addition to the networks in M2I was designed and trained [1,16]. However, the added segmentation mask synthesis network demanded a large number of labeled samples to make it well trained, which could be up to hundreds for example in [1] thus labor-intensive.

To constrain the synthesis of images and annotations at the same time without using a large scale of pre-labeling, we propose a novel controllable and simultaneous synthesizer (so-called CS^2) in this study. The novelty of our work is three-fold: 1) we develop a novel unsupervised mask-to-image synthesis pipeline that generates images controllably without human labeling; 2) instead of directly using the numeric and disarranged unsupervised segmentation masks, which are cluttered with over-segmented super-pixels, we assign the mean Hounsfield unit (HU) value for each cluster in the unsupervised segmentation masks to obtain an ordered and well-organized labeling; and 3) we propose a new synthesis network structure featured by multiple adaptive instance normalization (AdaIN) blocks that handles unaligned structural and tissue information. The code is publicly available at https://github.com/ayanglab/CS2.
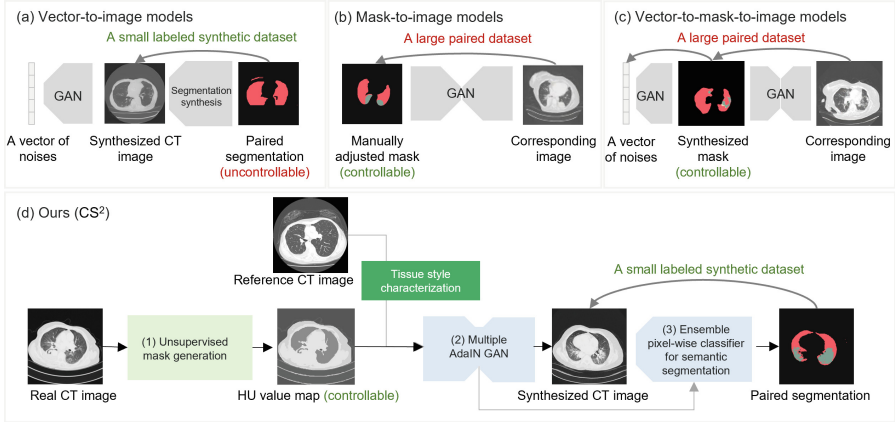
Fig. 1. The rationale of our developed CS^2 compared to previously proposed data augmentation methods, including vector-to-image (a), mask-to-image (b), and vector-to-mask-to-image (c). Our CS^2 model can produce a large labeled synthetic dataset controllably without using large-scale pre-labeling or human-labeled datasets.

2   Methods

Our model consists of three major components, including 1) an **unsupervised mask generation** module, which generates unsupervised segmentation masks from input images; 2) a **multiple AdaIN GAN** that receives an unsupervised mask and a reference CT image as inputs, and outputs a synthetic CT image; then we use a small labeled subset of synthetic images and leverage the feature maps from the mask-to-image model to train 3) an **ensemble pixel-wise classifier for semantic segmentation**. Once trained, our model can produce synthetic images and corresponding annotations simultaneously. We present the overall network structure in Fig. 1 (d) and introduce the details of each component as follows.

2.1   Unsupervised Mask Generation

To achieve a controllable synthesis, an adjustable input that contains the structure of synthetic images must be provided. As we mentioned in the Introduction section, the supervised generation of structural masks requires a large scale of pre-labeling. Thus, in our algorithm, we adopt an unsupervised structural mask generation [13] as shown in Fig. 2. Unsupervised segmentation masks are obtained by a super-pixel guided CNN.

Instead of using the disarrayed unsupervised masks directly, we compute the average HU value inside each class and assign this value to each cluster to create the structural guidance maps. In so doing, we can achieve an ordered and well-organized labeling for our unsupervised segmentation mask generation. We refer to this step as the mean HU assignment. According to the mean HU assignment,
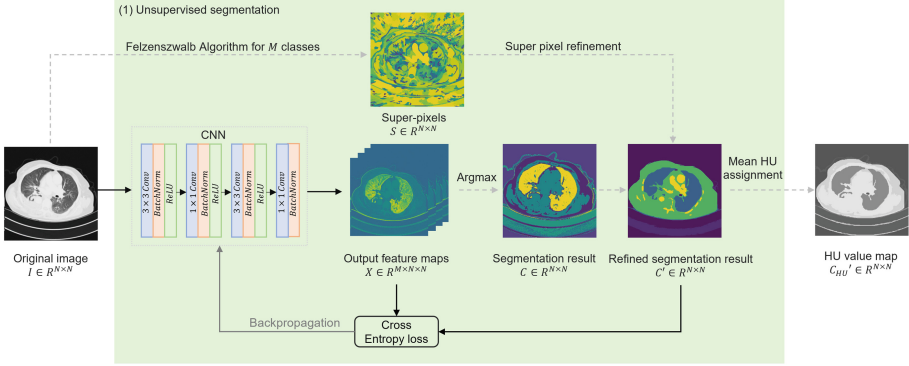
Fig. 2. The workflow for unsupervised mask generation. Dashed lines indicate no gradient back-propagation during the training procedure.

we can easily edit the unsupervised segmentation mask. For example, by adding a patch with an approximate HU value of the COVID infections, we can artificially add a lesion in the synthesized healthy lungs. It is of note that this mean HU assignment step is also crucial for the content matching loss defined in Sect. 2.2.

The unsupervised segmentation algorithm is optimized by the cross-entropy loss between CNN output X and the super-pixel refined mask C'. M is the number of classes initialized for our algorithm and is a pre-defined super parameter. For each cluster m in the super-pixel mask S, we extract the corresponding region $C^{(m)}$ in segmentation result C. Then we count the mostly appeared class in $C^{(m)}$, and assign this value $C_{\max}$ as the new label to all elements in $C^{(m)}$ The elements in $C^{(m)}$ is mathematically defined as

$$C_{i,j}^{(m)} = \begin{cases} 0 & S_{i,j} \neq m \\ 1 & S_{i,j} = m \end{cases}.$$

Here, $C_{\max}, m, C_{i,j}, S_{i,j}$ are class numbers in the range of $(0, M]$. The refined mask $C' = \sum_{m=1}^{M} C^{(m)}$. This refinement can decrease the number of classes in the segmentation results. Once the number of clusters in the refined segmentation result reaches four or the number of iterations reaches 60, we stop the training of the CNN and use the refined segmentation result C' as our unsupervised masks.

2.2 Multiple AdaIN GAN

For the image synthesis branch, we use a multiple AdaIN GAN shown in Fig. 3. The input of our generator is one HU value map and one non-corresponding reference CT image. An additional reference CT image helps the synthesis network to adopt the style of the whole dataset better, and the unpaired inputs increase the variability of the synthetic images. The encoder extracts feature from both
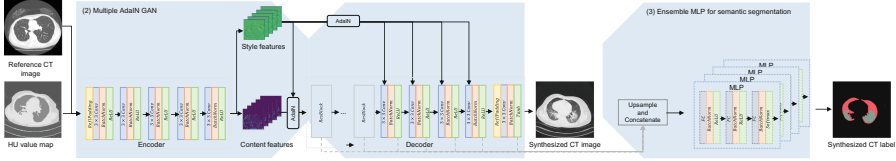
Fig. 3. The network structure of our UM2I model and the ensemble MLP pixel-wise classifier for the segmentation synthesis. Dashed lines indicate no gradient back-propagation during training in this operation.

HU value maps and CT images simultaneously, and an Adaptive Instance Normalization (AdaIN) [8] block normalize the content feature (from the HU value map) with the style feature (e.g., from the CT image). In our multiple AdaIN GAN, AdaIN blocks are added in all convolutional blocks except for the residual blocks (ResBlocks) [7]. This is because AdaIN changes the distribution of feature maps; thus, it corrupts the identity mapping between input feature maps and output feature maps in residual blocks.

The final loss function is composed of three parts: 1) a discriminator loss, which is computed by a patch-based discriminator [5]; 2) a style matching loss, which is the VGG loss [12] between the reference CT and our synthesized CT images; and 3) a content matching loss that is the Mean Squared Error (MSE) loss between the content and our synthesized CT.

In the previously published style transfer network using AdaIN [8], the content matching loss was computed by the VGG loss between content features and the feature maps extracted from the synthesized images. However, this loss function required extra GPU memory and limited the training batch size. We observed in our experiments that a large batch size led to a stable optimization procedure, which was also in accordance with a previous study [2]. Therefore, we replace the content matching loss using the MSE between the HU value map and our synthesized CT images. Due to our mean HU assignment step, this MSE loss function performs better compared to using the original content matching loss.

2.3 Ensemble MLP Classifier for Semantic Segmentation

Once the generative model is trained, we can generate synthetic images. The multiple AdaIN GAN provides effective image representations for synthetic images. We label several synthetic images (30 in our experiments) and then up-sample their features from the decoder into the same size as the original CT image. The features in the decoder are so powerful that even labeling a handful of synthetic images can lead to an accurate segmentation synthesis. We then train an ensemble multi-layer perceptron (MLP) classifier for the pixel-wise classification, as shown in Fig. 3. The ensemble MLP classifier and multiple AdaIN GAN are trained separately. During inference, synthesized CT images and corresponding annotations are produced. In our experiments, we trained 10 MLP classifiers and obtained the final segmentation mask by majority voting. The number of pixel-wise classifiers was selected based on empirical studies according to [21].

3   Experiments

3.1   Dataset

We used two datasets to validate the performance of our algorithm. The target of our experiment is to segment both lung tissues and ground glass opacities (GGOs) from CT images in those datasets. The first dataset is our in-house multi-center chest CT dataset, which contains 7,140 3D CT volumes from 2,071 COVID patients. An experienced radiologist manually labeled 1,192 volumes for this dataset. We randomly selected 192 volumes for independent testing. Detailed descriptions of our private dataset can be found in our supplementary file. The second is an open-access chest CT dataset, COVID-19-CT-Seg, containing 20 3D CT volumes of COVID patients [15]. All CT images from this open-access dataset have left lung, right lung, and infections labeled by two radiologists and verified by an experienced radiologist. It should be noted that we only used the open-access dataset for independent testing.

Our experiments were performed on 4-channel 2.5D CT images, i.e., we selected 4 2D slices from the 3D CT volume and stacked them channel-wise. The 2.5D concept [18] is widely used to represent 3D medical imaging volumes and to reduce computational costs. In this study, we used 2.5D CT representation to handle the varied slice thickness among multi-centered CT data. The selection strategy was a uniform selection from the middle (25%–75%) slices.

3.2   Experimental Settings

We compared our CS^2 with three state-of-the-art models, as shown in Fig. 1. For vector-to-image (V2I) synthesis, we used StyleGAN [14]. For mask-to-image synthesis (M2I), we used the pix2pix [11] model, and the masks are generated by randomly adding circular GGO patches to lung segmentation masks. For the vector-to-mask-to-image synthesis (V2M2I), we used styleGAN to generate synthesis masks and then used the pix2pix model to generate patches. For each generative model, we created synthetic datasets with 4,000 2.5D CT images. We will refer to 2.5D CT images as images or samples in the following section.

We further trained a UNet on these synthetic images together for lung and GGO segmentation. To prove that our synthetic data are representative, we also compared the segmentation result with the transfer-learning based method, i.e., we used a pretrained nnUNet model [10] (task numbered 115 [17]) and fine-tuned the last layer according to 10 volumes from our labeled in-house training dataset.

3.3   Results

Our Model Can Successfully Synthesize Both the Overall Shape of Lungs and Textures of the GGOs. Examples of synthetic images and annotations are shown in Fig. 4. The V2I model, although generates a realistic overall shape of the lung and human body, falsely captures the texture of lung tissues, as shown in Fig. 4 (a1). Since the segmentation mask of the M2I model is manually
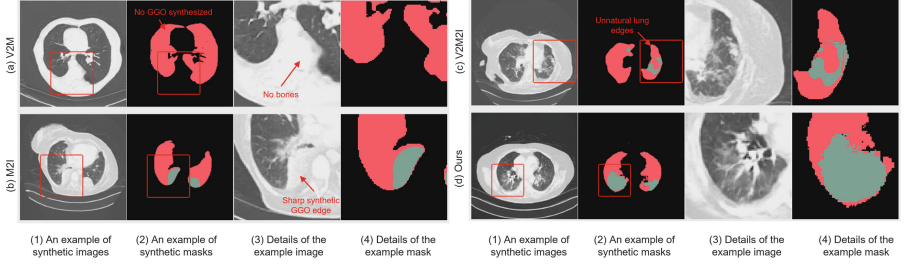
Fig. 4. Example synthetic images from four generative models. The masks are synthetic segmentation masks that corresponded with these images. Red pixels indicate lung tissues and green pixels are GGOs. **We also include synthetic samples of different modalities in our supplementary file.** (Color figure online)

generated, the M2I model failed to capture the gradual change between normal tissues and GGOs, as shown in Fig. 4 (b3). As for the V2M2I synthesized masks shown in Fig. 4 (c2), the lung edges are unnatural. Our model successfully generates realistic synthetic images and annotations. Both M2I and V2M2I models are trained with 500 2.5D CT images.

The Downstream Segmentation Model Trained on CS^2 Synthesized Images Can Reach an Accuracy as High as Fully Supervised Pretrained nnUNet. We trained our ensemble MLP classifier with a different number of labeled synthetic 2.5D CT images and generated 4,000 synthetic masks for each group. Then we trained our segmentation network with these synthetic data. The result is shown in Fig. 5 (a). In our in-house dataset, the UNets trained on CS^2 synthetic datasets with 20, 25, and 30 manual labels even outperform nnUNet which was trained on 199 manually labeled 3D CT volumes and fine-tuned by 10 volumes in a fully supervised manner. It is of note that in the public dataset, the UNet trained on our synthetic datasets did not perform
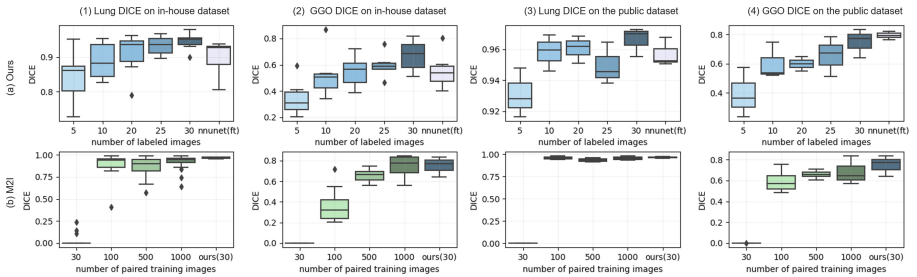
Fig. 5. Dice scores of the lung and GGO on both in-house and public datasets from models trained on CS^2 synthesized (a) and M2I model synthesized (b) datasets.
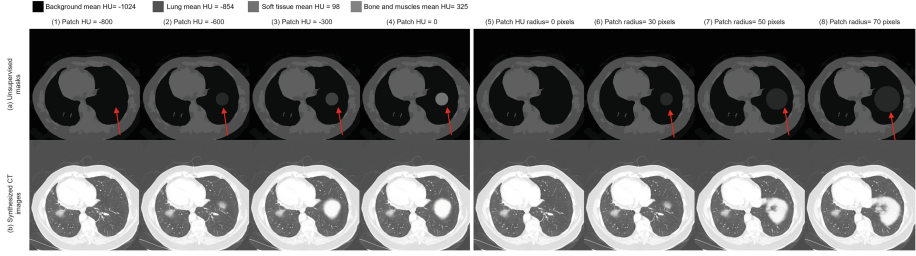
Fig. 6. An example of our synthetic images (b) structurally edited with the Unsupervised masks (a) by adding circular patches of different HU values (1–4) and radii (5–8). The patches in (a1) to (a4) have a radius of 30 pixels, and the patches in (a1) to (a4) have a mean HU value of −600.

as well as the fine-tuned nnUNet. However, our segmentation results are still comparable in such a low-labeled data scenario.

Our CS^2 Demonstrate Promising Results for Large-Scale Annotated Data Synthesis When Only a Limited Number of Annotations is Required. However, when the number of annotations increased, the performance of the M2I model and V2M2I model improved, as shown in Fig. 5 (b). We only plotted the M2I model performance here because the performance of the V2M2I model was limited by its M2I component. To achieve similar synthesis performance, the M2I model required at least 500 paired training images.

The Synthetic Images of Our Model are Structurally Editable. As shown in Fig. 6, we added circular patches in the lung masks with different mean HU values and different shapes. The mean HU values of the added patches are more decisive than their radius. For example in (b2) to (b4), even though we did not increase the radius of the added patches, our model detected a change from GGOs patches toward background tissue patches, thus automatically adjusting the size of the generated infectious areas. In contrast, as (b4) to (b8) show, once the mean HU value is fixed, the corresponding object is synthesized (e.g., no extraneous tissue or lesion is generated). The example HU values were given by an experienced radiologist, and we present the statistical distributions of HU values in unsupervised masks in our supplementary file.

4   Conclusion

In this study, a generative model called CS^2 has been proposed for data augmentation. By labeling only a handful of synthetic images (e.g., 30 2.5D CT images in our experimental studies), our CS^2 model has been able to generate realistic images and annotations at the same time. In addition, our algorithm has no requirement of any large pre-labeling or manually adjusted segmentation

masks. Our experiments have proven our hypothesis that the proposed CS^2 has produced realistic synthetic images with annotations, which have enhanced the accuracy of lung tissues and infections segmentation of COVID-19 CT images. One limitation of our work is that we obtained our segmentation masks from a pixel-wise classifier, so the segmentation masks are fuzzy and have a few (8%) scattered wrongly labeled pixels. Post-processing algorithms on segmentation masks such as connected component analysis can remove these artifacts easily.

References

1. Bailo, O., Ham, D., Min Shin, Y.: Red blood cell image generation for data augmentation using conditional generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
2. Brock, A., Donahue, J., Simonyan, K.: Large scale GAN training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
3. Tsaftaris, S.A., Gooya, A., Frangi, A.F., Prince, J.L. (eds.): SASHIMI 2017. LNCS, vol. 10557. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-68127-6
4. Dalmaz, O., Yurt, M., Çukur, T.: ResViT: residual vision transformers for multi-modal medical image synthesis. arXiv preprint arXiv:2106.16031 (2021)
5. Demir, U., Unal, G.: Patch-based image inpainting with generative adversarial networks. arXiv preprint arXiv:1803.07422 (2018)
6. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27 (2014)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
8. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1501–1510 (2017)
9. Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R.G., Landman, B.A.: Adversarial synthesis learning enables segmentation without target modality ground truth. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1217–1220. IEEE (2018)
10. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No New-Net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 234–244. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_21
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)

12. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43

13. Kanezaki, A.: Unsupervised image segmentation by backpropagation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1543–1547. IEEE (2018)

14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)

15. Ma, J., et al.: Towards efficient Covid-19 CT annotation: a benchmark for lung and infection segmentation. arxiv 2020, arXiv preprint arXiv:2004.12537 (2020)

16. Pandey, S., Singh, P.R., Tian, J.: An image augmentation approach using two-stage generative adversarial network for nuclei image segmentation. Biomed. Signal Process. Control **57**, 101782 (2020)

17. Roth, H., et al.: Rapid artificial intelligence solutions in a pandemic-the Covid-19-20 lung CT lesion segmentation challenge (2021). researchsquare.com

18. Roth, H.R., et al.: A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014. LNCS, vol. 8673, pp. 520–527. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10404-1_65

19. 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

20. Sun, Y., Yuan, P., Sun, Y.: MM-GAN: 3D MRI data augmentation for medical image segmentation via generative adversarial networks. In: 2020 IEEE International Conference on Knowledge Graph (ICKG), pp. 227–234. IEEE (2020)

21. Zhang, Y., et al.: DatasetGAN: efficient labeled data factory with minimal human effort. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10145–10155 (2021)