2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9433996

ENHANCED-QUALITY GAN (EQ-GAN) ON LUNG CT SCANS: TOWARD TRUTH AND POTENTIAL HALLUCINATIONS

Martin Jammes-Floreani[1], *Andrew F. Laine*[2], *Elsa D. Angelini*[1,2,3]

[1]NIHR Imperial Biomedical Research Centre, ITMAT Data Science Group,
Imperial College London, London, UK
[2]Department of Biomedical Engineering, Columbia University, New York, USA
[3]Department of Metabolism-Digestion-Reproduction, Imperial College London, UK

ABSTRACT

Lung Computed Tomography (CT) scans are extensively used to screen lung diseases. Strategies such as large slice spacing and low-dose CT scans are often preferred to reduce radiation exposure and therefore the risk for patients' health. The counterpart is a significant degradation of image quality and/or resolution. In this work we investigate a generative adversarial network (GAN) for lung CT image enhanced-quality (EQ). Our EQ-GAN is trained on a high-quality lung CT cohort to recover the visual quality of scans degraded by blur and noise. The capability of our trained GAN to generate EQ CT scans is further illustrated on two test cohorts. Results confirm gains in visual quality metrics, remarkable visual enhancement of vessels, airways and lung parenchyma, as well as other enhancement patterns that require further investigation. We also compared automatic lung lobe segmentation on original versus EQ scans. Average Dice scores vary between lobes, can be as low as 0.3 and EQ scans enable segmentation of some lobes missed in the original scans. This paves the way to using EQ as pre-processing for lung lobe segmentation, further research to evaluate the impact of EQ to add robustness to airway and vessel segmentation, and to investigate anatomical details revealed in EQ scans.

Index Terms— Lung CT, Generative Adversarial Network (GAN), image enhancement, super resolution.

1. INTRODUCTION

Several recent approaches have exploited GANs to generate super-resolution medical images. The main class of models that addresses image enhancement is Super Resolution GAN (SRGAN) [1]. Initially developed to upsample images while preserving realistic appearance, its architecture can be adapted for CT denoising and deblurring via training on pairs of images: original high quality image (the target) and corresponding image degraded with blur and noise.

As CT volumes are too large to fit on common GPUs, most approaches either work on 2D slices, downsampled 3D volumes or 3D patches [2, 3]. In lung CT scans, another challenge arises from the variability of slice thickness and spacing and inconsistency in number of axial slices because of variable morphology and selected field of view. For EQ, an additional difficulty is the lack of ground-truth pairs on high/low quality CT images, due to ethical consideration. In this work, we chose to work on fixed-size 2D axial slices (most consistent scan resolution across cohorts) artificially degraded for training and tested on low quality CT scans.

1.1. Data

We used 3 publicly available cohorts of lung CT scans as detailed in Table 1 to train, validate and test our GAN. The ILD cohort (N=102 scans annotated for interstitial lung disease) was selected as our train/validation cohort because it displays a large variety of lung disease patterns and has high visual quality in 2D axial slices. Tests were performed on two cohorts from The Cancer Imaging Archive (TCIA): NSCLC-Lung3 (N=89 scans of low visual quality annotated for lung cancer) and RIDER (N=59 scans of average quality annotated for lung cancer).

1.2. CT scan preparation

We set the field of view size to 512×512 pixels per axial slice, clip intensity values between [-1000, 1000] Hounsfield units (HU) and then rescale between 0 and 1. During the training part, we artificially degraded ILD CT scans to simulate low quality CT scans with random blur and noise as follows:

- To **add blur:** 2D convolution kernels are applied with the following coefficients: $G(i,j) = \frac{1}{2\pi\sigma^2} e^{-\frac{i^2+j^2}{2\sigma^2}}$, where $\sigma \sim \mathcal{U}(0,\, \sigma_{max})$ and σ_{max} determines the maximum level of blur to add within the degraded images.
- To **add noise:** the following values are added to the pixels and new values are clipped to $[0, 1]$: ϵ where $\epsilon \sim \mathcal{N}(0,\, \mu^2)$ with $\mu \sim \mathcal{U}(0,\, \mu_{max})$, and μ_{max} sets the maximum level of noise added to the images.

The levels of blur and noise controlled with σ and μ are chosen randomly in the data loader during training to simulate various levels of noise and blur.

Table 1*: Datasets used. ILD=interstitial lung disease, NSCLC=non-small-cell lung cancer. Date=publication date. Resolution: axial pixel size $\times$ slice thickness (can be different from slice spacing).*

Name	Origin	Date	# scans	Resolution	Disease	Usage
ILD [4]	Medgift [5]	2012	102	0.410.94 mm $\times$ 1-2 mm	ILD	Training/Validation
NSCLC-Lung3 [6]	TCIA [7]	2014	89	0.601.37 mm $\times$ 1.5-8.0 mm	NSCLC	Test
RIDER [8, 9]	TCIA [7]	2008	59	0.500.92 mm $\times$ 1.25 mm	NSCLC	Test

2. METHOD

2.1. Model architecture

After several pre-training experiments, we ended up using a model inspired from the ESRGAN architecture [10] adapted from initial SRGANs as follows. This was the one providing the best visual quality metrics. Batch normalization (BN) layers were removed from the generator, and residual blocks (RB) were replaced by the proposed residuals in Residual Dense Blocks (RRDBs). All upsampling layers were also removed as resolution of input and output images are the same. The parameters were adapted to handle axial images of size 512×512 and simplified to handle gray-level rather than color images, leading to gain in computation time.

Following the work presented in SRGAN [1] and TomoGAN [2], we used a hybrid loss function for the generator via a linear combination of adversarial loss, MSE loss (content loss) and MSE loss on the activated features of the third block of a VGG19 network pre-trained on Imagenet [11] (perceptual loss). Weights between the loss terms were part of the optimized parameters.

2.2. Training parameters

The best model for final experiments has been trained using Adam optimizer with a learning rate of 0.0002 over $20,000$ epochs, taking approximately 8 days on a TITAN RTX GPU.

3. RESULTS

All GAN-based images were generated in the range [0, 1] and rescaled to pixel intensities in the range [-1000, 1000] (HU) (with similar clipping applied to the input images). We trained our GAN on 80 randomly selected ILD scans and used the rest (22 scans) for validation. For training/validation image degradation, we set $\sigma_{max} = 3$ and $\mu_{max} = 0.3$ to make the most degraded ILD images look worse than the worst scans of the test cohorts (based on visual evaluation). Optimal weights between the 3 loss terms are: 1, 5×10^{-3} and 1×10^{-2} for perceptual, adversarial and content loss.

For validation, we evaluated the predictions on degraded images and compared the results with the ground-truth corresponding original CT scans (target predictions) using visual quality metrics. Visual quality metrics are: Mean Square Error on VGG19 features (MSE_{VGG}), Structural Similarity (SSIM) and Peak Signal to Noise Ratio (PSNR).

For testing, we visually inspected the EQ scans, and compared lung lobe segmentation results on original versus EQ scans.

3.1. Limitations of SR-GAN for upsampling

Generating lung CT scans from pure noise, as used in classic GAN approaches, remains quite challenging. To test the ability for GAN models to generate realistic lung CT images, we first relied on training a custom SR-GAN to interpolate from pairs of (downsampled, upsampled) training images, as illustrated in Fig. 1. Low-resolution axial images were resized to 28×28 pixels while high-resolution target images were of size 224×224 pixels. The generated images, obtained after several tuning and long training times, highlight the ability for a GAN model to output images on which the shape of the lungs is preserved and some relevant details are revealed even if not humanly-perceptible in the original low-resolution images. However, vessels and airways are not properly recovered in this setup. This suggested that SR-GAN model can learn details present in chest CT scans (vertebra and rib bones, round large vessels) but can't recover small lung vessels and airway structures if downsampled.

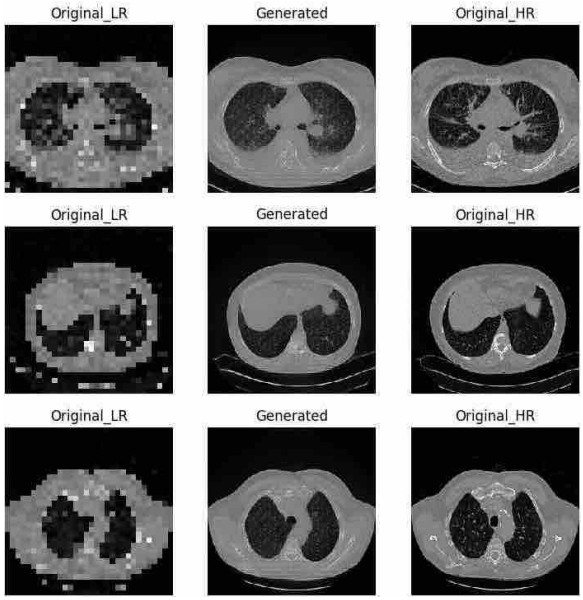

Fig. 1*: SR-GAN up-sampling capability as proof of concept: Up-sampling 2D images from validation dataset (ILD) from 28×28 pixels to 224×224 pixels, after 6,250 training epochs.*

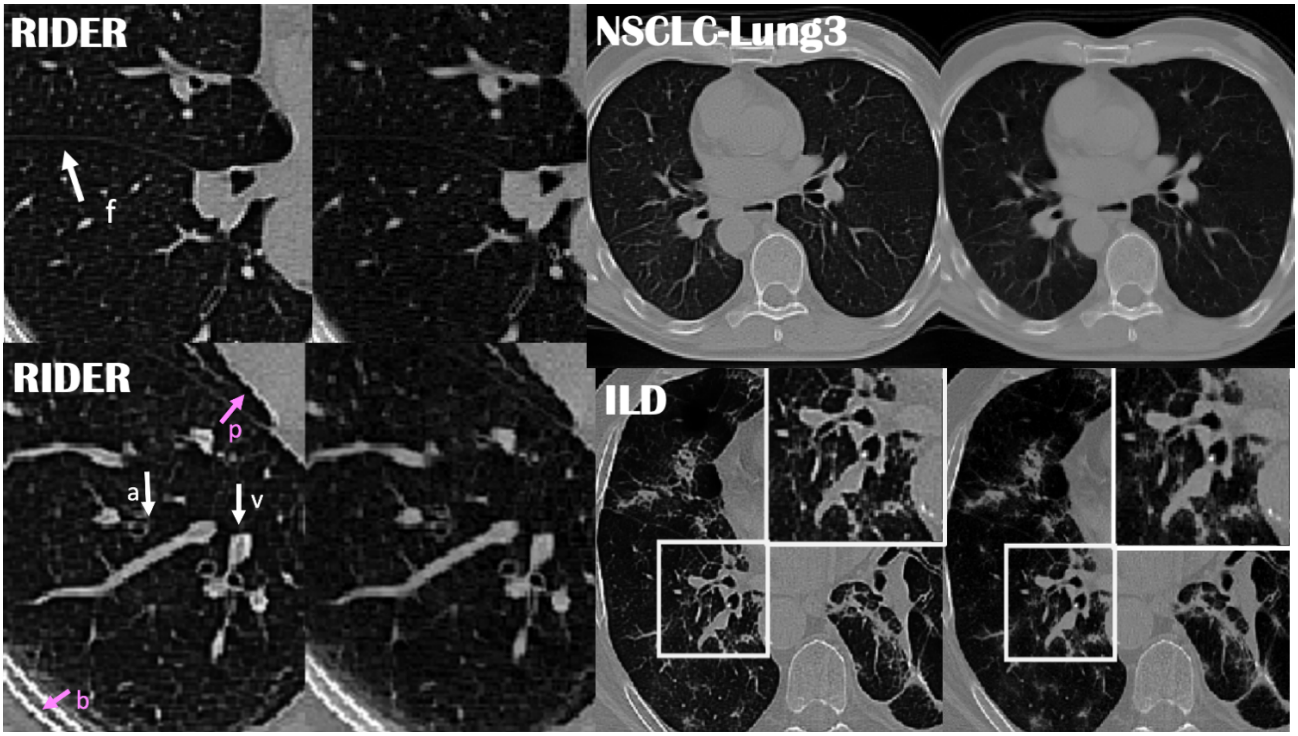

Fig. 2: *EQ-GAN reconstructions on validation (ILD) and test (NSCLC-Lung3 and RIDER) cohorts. Left=EQ scan, Right=original scan. (f=fissure, a=airway, v=vessel, p=pleura, b=bone).*

3.2. Visual quality at fixed resolution

We illustrate visual results on the validation ILD cohort and test NSCLC-Lung3 and RIDER cohorts in Fig 2. These visualisations show clear anatomically-plausible enhancements such as debluring effect (NSCLC-Lung3), enhanced vessels and airway walls (all), and contrast on the fissures (RIDER). Additional enhancement patterns, such as bright enhancement of the pleural border (RIDER, ILD), dark outline of the bones (RIDER, ILD) and overall more small-scale texture patterns in the lung parenchyma (all) are cohort-specific and need further investigation to explain "hallucination" effects (e.g. arising from phenomena similar to ringing artefacts generating "ghost" edges).

3.3. Quantitative evaluation on generated images

As common metrics such as pixel intensity RMSE are not able to fully catch differences in visual quality of important details, we chose to quantitatively evaluate our results using MSE on high-level features computed on the third block of a pre-trained VGG19 network [11]. Using the original images as ground-truth, we obtained the following MSE_{VGG} metrics on the validation ILD cohort: $MSE_{VGG} = 49 \pm 66$ for the EQ images, versus $MSE_{VGG} = 69 \pm 44$ for the degraded images. Other classic metrics were not very discriminative: SSIM = 0.606 ± 0.118 (degraded) versus 0.614 ± 0.117 (EQ); PSNR = 1.65 ± 9.22 (degraded) versus 1.81 ± 9.34 (EQ).

3.4. Quantitative evaluation for lung lobe segmentation

We ran on the test NSCLC-Lung3 cohort of 89 subjects the publicly-available lung lobe segmentation tool from [12], pre-trained on lung CT scans with severe pathologies, such as fibrosis and trauma, and using the proposed default settings. Lung lobe segmentation was performed for 5 lobes indexed from 1 to 5 in the following order: Left (superior-inferior) and Right (superior-middle-inferior). We measured the average Dice coefficients on all lobes and per lobe between the original and EQ 3D volumes. Results are provided in Fig. 3. While whole-lung Dice are all above 0.99 the average Dice over lobes has a median of 0.97 and minimum of 0.78 (case with large upper-body Field of view (FOV)). The 0.3 Dice score for lobe #1 is an upper-body FOV case where part of left superior lobe was detected on the EQ scan while entirely missed on the original scan. There are 2 other upper-body FOV cases where lobes #2 and #4 were missed entirely in original scans while detected in EQ scans. Per lobe, percentages of scans with lobe Dice scores above 0.98 are: 70%, 52%, 55%, 7%, 55% versus above 0.97 are: 84%, 68%, 81%, 34%, 77%. Thus, enhancement on a single CT can affect Dice agreement of "repeated" segmentation of a given lobe by 0.01 on 70% of scans if using a small cohort of 89 scans from TCIA (hence likely close to clinical "quality" image scanning setting).

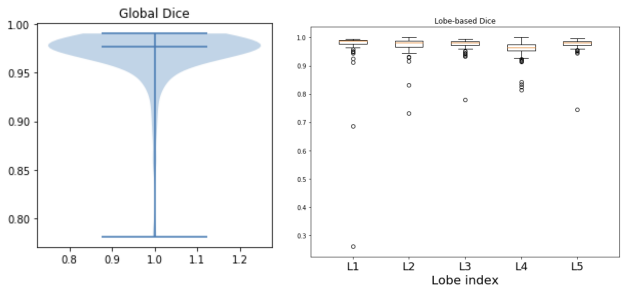

Fig. 3*: Boxplots of Dice coefficients for lung lobe segmentation on NSCLC-Lung3 test cohort (N=89) comparing original and EQ volumes. (left) Violin plot on average lung lobes Dice coefficients. (right) Lobe-based Dice coefficients. Lobe indices L_i (in order)= Left (superior-inferior) and Right (superior-middle-inferior).*

4. CONCLUSION

Improving the quality of lung CT scans using GAN deep learning algorithms is a promising but complex task due to the challenges related to their training, the lack of availability of real pairs of high/low resolution scans, and the difficulty to objectively evaluate improvement in visual perception. This work highlights the capabilities of a recent EQ-GAN architecture specifically trained to enhance axial slices on cohorts of lung CT scans with various origins and baseline image quality. Qualitative and quantitative evaluations as reported confirm promising potentials to enhance expected anatomical structures as well as require further investigations of potentially "ghost-hallucination" contrasts. Future work will explore transfer learning toward 3D by following for example the method from [3]. However, variability in slice thickness between CT scans and non-constant FOV used in public cohorts need to be handled properly. Our proposed EQ tool can also be considered as standard pre-processing, similar to Gaussian smoothing. This would constitute a drastic change of paradigm, aiming for enhanced versus degraded (smoothed) versions of CT scans to be used for quantification.

Compliance with Ethical Standards: Data from open access sources (details in the text).

Acknowledgements: Authors report no conflict of interest. This work was partially supported by NIH 2R01-HL121270-05.

5. REFERENCES

[1] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE CVPR*, 2016, pp. 4681–4690.

[2] Z. Liu, T. Bicer, R. Kettimuthu, D. Gursoy, F. De Carlo, and I. Foster, "TomoGAN: low-dose synchrotron X-ray tomography with generative adversarial networks: discussion," *Journal of the Optical Society of America A*, vol. 37, no. 3, pp. 422–434, 2020.

[3] H. Shan, Y. Zhang, Q. Yang, U. Kruger, M. K. Kalra, L. Sun, W. Cong, and G. Wang, "3-D convolutional encoder-decoder network for low-dose CT via transfer learning from a 2-D trained network," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 15221534, 2018.

[4] A. Depeursinge, A. Vargas, A. Platon, A. Geissbuhler, P. A. Poletti, and H. Muller, "Building a reference multimedia database for interstitial lung diseases," *Computerized Medical Imaging and Graphics*, vol. 36, no. 3, pp. 227–238, 2012.

[5] "MEDGIFT content-based Image Retrieval," http://medgift.hevs.ch/, Accessed: 2020.

[6] H. J. Aerts, E. R. Velazquez, R. T. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, et al., "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Communications*, vol. 5, no. 1, pp. 1–9, 2014.

[7] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al., "The cancer imaging archive (TCIA): maintaining and operating a public information repository," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[8] B. Zhao, L. P. James, C. S. Moskowitz, P. Guo, M. S. Ginsberg, R. A. Lefkowitz, Y. Qin, G. J. Riely, M. G. Kris, and L. H. Schwartz, "Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non–small cell lung cancer," *Radiology*, vol. 252, no. 1, pp. 263–272, 2009.

[9] S. G. Armato III, C. R. Meyer, M. F. McNitt-Gray, G. McLennan, A. Reeves, B. Y. Croft, L. P. Clarke, and R. R. Group, "The reference image database to evaluate response to therapy in lung cancer (RIDER) project: A resource for the development of change-analysis software," *Clinical Pharmacology & Therapeutics*, vol. 84, no. 4, pp. 448–456, 2008.

[10] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *The European Conference on Computer Vision Workshops (ECCVW)*, 2018.

[11] "Keras applications," https://keras.io/api/applications/, Accessed: 2020-07-30.

[12] J. Hofmanninger, F. Prayer, J. Pan, S. Rohrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is a data diversity problem, not a methodology problem," *arXiv:2001.11767*, 2020.