CT Images GAN-based Augmentation with AdaIN for Lung Nodules Detection

Maksim Kryuchkov[1], Natalia Khanzhina[1], Ilya Osmakov[2], Pavel Ulyanov[2]
[1]Machine Learning Lab, ITMO University, 49 Kronverksky Pr., St. Petersburg 197101, Russia
[2]Burtsev Lab, Skolkovo Innovation Center, 42 Bolshoy Blvd., Moscow 121205, Russia

ABSTRACT

In this work we made an attempt to improve 3D detection of pulmonary nodules on CT images using Conditional GANs extended with Adaptive Instance Normalization and combined Wasserstein Loss for data augmentation (DA). Nodule generating GAN model used for DA was built upon an open-source CT-GAN network which provides high and reproducible results in the nodule generation task. For the evaluation purpose we used DeepSEED model, which is a 3D end-to-end one-stage detector. We tested our approach on the LUNA16 dataset, the subset of LIDC-IDRI. The proposed model outperformed the baseline detection model trained on the original dataset by 3% of average sensitivity. The augmentation helped achieve a remarkable classification rate: 91% of sensitivity and 86% of specificity.

Keywords: Lung nodules generation, lung nodules detection, Adaptive Instance Normalization, Conditional GAN, Wasserstein GAN.

1. INTRODUCTION

Statistically, lung cancer is considered the most widespread type of cancer due to the environmental factors and the prevalence of smoking in the population. Nowadays it is one of the most severe cancers in terms of mortality rate, which is caused by frequent late diagnosis of the disease.

This type of cancer most often manifests itself as malignant nodules of various sizes. The standard method of diagnostics involves computed tomography (CT) lung scanning for nodules. This work is laborious and requires highly experienced specialists which however can make errors of various kinds.

To treat the disease, it is critical to diagnose it at its early stages. The fact that the recognition of small neoplasms is much more complicated than the recognition of large ones leads to an even greater risk of a specialist error. These considerations determine the great relevance of the nodule automated recognition task. Unlike other methods of radiation diagnostics, computed tomography has a high resolution, thus small nodules (less than 3 mm in diameter) detection is feasible. However, this makes diagnostics made by humans even more difficult and also turns automatic recognition and detection of lung nodules into a challenging problem to solve for deep learning and computer vision methods.

Deep neural networks require large amounts of data to train efficiently, while data collection is very time consuming and human resources demanding. For example, labelling the LIDC-IDRI [1] dataset, which is used as the benchmark for lung nodules recognition task and includes only scans collected from 1012 patients in total, involved four specialists. This results in the need of data augmentation, which can also be achieved using generative adversarial networks (GANs). GANs are widely applied for data generation. Generated data can later be used as an extension for an existing dataset, which the detector is supposed to be trained on.

Despite that segmentation task for medical screening problems is much more widely spread than detection task, in the specific case of lung nodule recognition domain detection task is common. The main reason is that nodule tissues are much smaller than CT scans. Mean nodule diameter calculated across data provided by LIDC-IDRI [1] is 12.8 mm. In-plane pixel size of LIDC-IDRI scans ranges from 0.461 mm to 0.977 mm, while average slice thickness ranges from 1.5 mm to 2.5 mm [1].

Recent advances for GANs training boosted its performance in different tasks significantly. One of the advanced practices is called Adaptive Instance Normalization (AdaIN) [2]. AdaIN is a technology for normalizing the outputs of the latent network layers, which was proposed in [2] to improve style transferring. This is an affine transformation of the

input x parametrized by some y. The authors proposed to use y as the style, but later this normalization was found to be useful in image generation in general. Thus, AdaIN can be used not only to solve the problem of the style transfer, but also to effectively train normalization parameters.

The contribution of this work is the usage of AdaIN for nodules generation on 3D CT images of lungs based on CT-GAN network, that, to our knowledge, was not applied by other researchers for such kind of tasks. We modified the CT-GAN architecture by including AdaIN layers and modified the loss function combining it with Wasserstein Loss component. Our approach is aimed at solving both the specific problem of class imbalance and increasing the diversity of nodules. We evaluated our modification using DeepSEED [3] network for nodules detection to show the sustainable increase of the performance in terms of sensitivity on the subset of LIDC-IDRI, namely LUNA16 [4].

The rest of the paper is organized as follows. The related works are briefly summarized in Section 2. In Section 3, we describe implementation details of the proposed improvements to the baseline network. The experiments are described in Section 4.1, followed by their results in Sections 4.2, 4.3. Section 5 summarizes the results and outlines the further work.

2. RELATED WORK

2.1 Nodules Augmentation

A number of papers have proposed augmentation of pulmonary datasets using GANs. In particular, Conditional GANs were used in [5] for 2D nodules segmentation, in [6] for 3D segmentation. Authors [7,8] use CGAN for 3D nodules detection. Works [7,9] considered nodules position taken surroundings tissues as the model input condition. WGAN [10] was used in [11,12], and Multi-Conditional GAN with Wasserstein Loss (WL) added was used in [9], that was proved to generate more diverse images.

Based on the success of Conditional GANs for the task, in our work we chose CT-GAN from [7] as the baseline, which provides not only plausible 3D images, but also an open framework for nodule injection. We decided to modify the loss function of GAN combining it with WL to address the problem of diversity.

CT-GAN

The model is built upon Conditional GAN (CGAN) and is trained to generate a crop of a CT image of size 32^3 containing a nodule. To train CT-GAN is fed with CT scan crops of the same size with a zero-mask placeholder in the middle where the original nodule is cut out. The remainder of the crop serves as the context i.e. condition. The generator network consists of several encoder layers followed by a similar decoder. The architecture is enriched with skip connections and batch normalization applied after each convolutional block.

AdaIN

Adaptive Instance Normalization (AdaIN) is a new technique, which could be considered as an alternative to Batch Normalization [13] and Instance Normalization [14] often used in neural networks. In AdaIN mean and standard deviation are computed across spatial dimensions which is similar to Instance Normalization. Each convolution block output is fed to an auxiliary network, which outputs the affine parameters used in AdaIN. Thus, normalization utilizes learnable affine parameters, which can lead to significant improvement of a neural network performance. We attempt to use this technique to improve the baseline CT-GAN network.

2.2 Nodules detection

End-to-end 3D convolutional neural networks have demonstrated its efficiency for the detection of pulmonary nodules [15-18]. Some of the authors use Faster R-CNN [19] based models [16-18], performing detection in two stages. We chose the DeepSEED [3] model, that works in one stage showing state-of-the-art comparable results and has an open source code.

DeepSEED

This is a convolutional network with the Squeeze-and-Excitation (SE) Encoder Decoder architecture that was used for nodules detection in [3]. The network uses the 3D ResNet18 architecture as an encoder with built-in SE modules [20] after each ResNet convolutional block (residual block) [21]. The output of the decoder is fed to the Region Proposal

Network (RPN) [19], which is trained to predict bounding boxes as well as probabilities of anchors to be nodules. The model could be easily trained and was used in our experiments for the proposed model evaluation.

3. THE APPROACH

3.1 GAN Implementation Details

CT-GAN was adopted as a baseline GAN architecture. Several problems emerged while baseline GAN performance evaluation including the mode collapse and the discriminator over generator domination. Thus, common architectural improvements were proposed, which involved generator and discriminator weight updates frequency ratioing and using Wasserstein Loss (WL). WL was combined with a baseline MSE loss between generated and original scan crops according formula 1.

$$L(o,g,D_{output},labels)=MSE(o,g)+WL(D_{output},labels), \quad (1)$$

where *o, g* stand for original and generated images corresponding to the same context, D_{output} is the output of the discriminator, $labels \in [-1,1]$.

Applying these methods to the GAN training procedure resulted in substantial improvement of visual quality and diversity of generated data along with the generator slightly dominating over the discriminator.

AdaIN was added to the network instead of Batch Normalization. It is applied after each convolutional block in both the encoder and decoder except the first layer. An auxiliary dense network computes affine parameter passing convolutional block output through three dense layers, two of which are shared across all the AdaIN auxiliary networks in the GAN. The modified convolution and deconvolution blocks of the original network are presented in Fig. 1. The normalization parameters are applied by the formula 2.

$$AdaIN(x,y) = \sigma(y)\left(\frac{x-\mu(x)}{\sigma(x)}\right) + \mu(y), \quad (2)$$

where *x* is the previous layer output, *y* is the affine transformation.

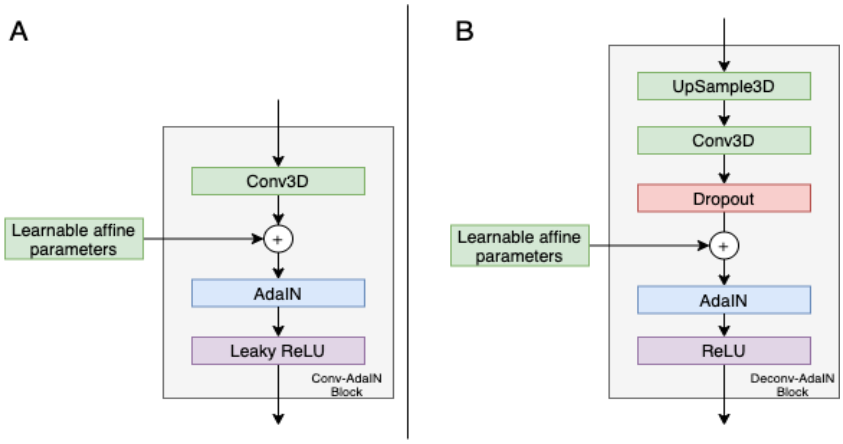

Figure 1. The modified blocks of CT-GAN: (A) Convolutional block, BatchNorm layer is replaced with AdaIN, the affine parameters come from the auxiliary Dense network, activation function is placed at the end of the block; (B) Deconvolutional block, BatchNorm layer is replaced with AdaIN, the affine parameters come from the auxiliary Dense network, activation function is placed at the end of the block.

Proposed Training Procedure

The end-to-end procedure of augmenting data and training the detection model is as follows:

1. Selecting labeled data from the original dataset;
2. Cropping nodules surrounded with context on selected images to form the dataset suitable for GAN training;

3. Applying standard augmentation methods to the GAN training set including random flips, rotations and shifts;
4. Applying histogram equalization and normalization to the GAN training set images;
5. Training GAN;
6. Feeding preprocessed crops obtained from the original dataset to the trained GAN;
7. Extending the original dataset with generated data obtained from GAN;
8. Detection model training on augmented dataset.

The detection model used was fully adopted from DeepSEED [3].

3.2 GAN Augmentation

Data fed to trained GAN was collected from the original dataset. Crops included original nodule surrounding contexts with nodules cut in the middle. Though with such an approach augmented data seems spatially isomorphic to the original data, there are significant advantages:

1. The spatial features of the nodules are not violated, i.e. a nodule will never be injected somewhere it is unlikely to appear from the biological point of view;
2. Since Wasserstein Loss was adopted, visually, generated nodules essentially differ from the original ones compared to the baseline GAN model;
3. Detecting places suitable for injection is not a trivial task and requires either solving an additional CT scan segmentation task and evaluating the solution, or manual scan labelling.

4. EXPERIMENTS AND RESULTS

4.1 Experimental setup

4.1.1 Dataset

LIDC-IDRI [1], is an open dataset in DICOM format labeled by 4 radiologists. In total, the database contains 1018 scans and more than 7,000 nodules. In this work, a subset of LIDC-IDRI, namely LUNA16 (LUng Nodule Analysis 2016) [4], was used; it contains 888 scans and 1187 nodules, each labeled with a bounding box. Unlike LIDC-IDRI, LUNA16 excludes scans thicker than 2.5 mm and includes only those nodules that were marked by at least three of four radiologists. Another feature of LUNA16 compared to LIDC is that the average size of nodules in LUNA16 is 8.3 mm with a standard deviation of 4.8 mm, when for LIDC-IDRI these indicators are 12.8 mm and 10.6 mm, respectively. As a result, the nodule detection objective on LUNA16 is considered less challenging than on LIDC-IDRI. Furthermore, train data collected from LUNA16 seems better for the GAN augmentation task due to its higher plausibility compared to LIDC-IDRI data.

4.1.2 Experiments

We used the 5-folds cross-validation in our experiments. During each experiment a test set was fixed while two detection models were trained on the original train set and augmented train set respectively. The test set contained crops made from 80 scans, each scan producing 3 positive and 3 negative crops, which can be considered as a Test Time Augmentation [22]. The augmented train set contained 400 scans generated via GAN, which means that about 30% data was augmented. Scans from the train set produced crops via random crop augmentation. We trained our networks on Tesla P100 PCIe 16GB, provided by Google Colab, the training of one model took about 30 hours.

4.1.3 Detection Evaluation

To evaluate the detection model, we used a slightly changed FROC analysis. Usually, in FROC analysis sensitivity is calculated over a fixed number of average false positive detections per scan, while we calculated sensitivity over average false positives per a scan crop of size 128^3. In this analysis sensitivity is defined as a percentage of crops on which the *Intersection over Union* of predicted and labeled bounding boxes is greater than 0.5. Original FROC analysis used in DeepSEED requires the network to process crops of size 208^3, which we could not afford due to computer performance limitations. We did not only reduce test crop sizes from 208^3 to 128^3, but also refused from per-scan metric calculations

in order to perform representative per-crop evaluation on a diverse test set, which was achieved by test-time crop augmentation.

4.2 GAN Results

Images generated by the baseline CT-GAN model had similar synthesized nodules which also can be found very blurry as shown in Figure 4. Such GAN performance does not seem to satisfy the major concept of using GANs for data augmentation, specifically extending original dataset with diverse generated data.

AdaIN layer added substantial contrast between nodule and healthy tissue (shown in Figure 3), which was an achievement compared to baseline images. Now the detector model could learn how nodule tissue is different from the surroundings due to high contrast. However, the baseline extended with single AdaIN produced no plausible results. When Wasserstein Loss was adopted, the transition between nodule and healthy tissues became much smoother and, as the result, images looked much more plausible as shown in Figure 2. Still synthesized nodules were enough different from each other which was also achieved by WL because after WL was added the generator substantially increased its ability to learn and managed to minimize loss.

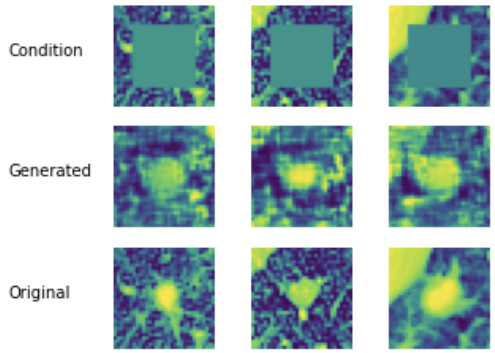

Figure 2. Proposed GAN model (with AdaIN and WL) output image slice examples

Besides described methods an attempt to solve mode collapse problem by increasing latent space dimensions was made. Crop sizes were changed from 32^3 (as originally proposed in [7]) to 48^3, extending the nodule surrounding context and thus increasing the GAN latent space. The zero-masked volume of size 20^3 was left as it is. However, a model trained on modified data showed even worse results, thus extending crops was not included in the proposed model.

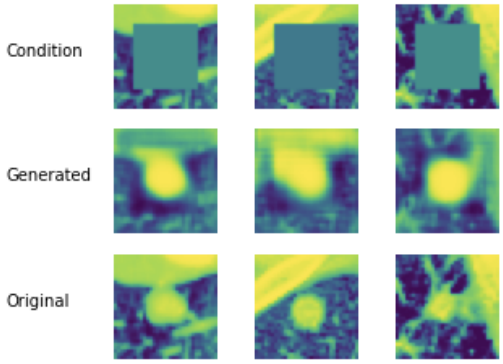

Figure 3. CT-GAN extended with AdaIN output image slice examples

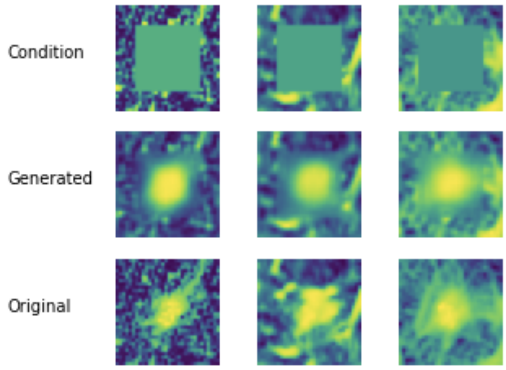

Figure 4. CT-GAN output image slice examples

Due to GAN being fed crops with nodules of diameter between 10 and 16 mm, plenty of augmented scans were significantly different from the original scans because of much bigger injected nodules as shown in Fig 6. This issue could be considered as an advantage because the generated nodule was completely different from the original one, which implies that the dataset was enriched with diverse augmented data, different from the original data, but still plausible. On the other hand, this might force the detector model to adapt to bigger nodules and ignore smaller ones.

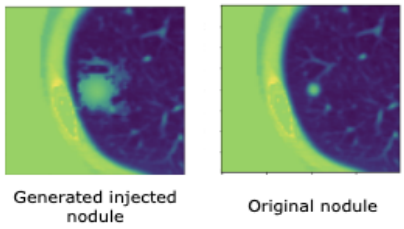

Figure 5. Nodule injected via GAN compared to the original nodule. Tissues surrounding nodules are the same

4.3 Detection Results

The detection model trained on the augmented dataset outperformed baseline by 3% as shown in Table 1. The model proposed has shown reasonable results compared to the similar work [9] where Conditional GAN was adopted for CT scan augmentation and 3D Faster-RCNN was used as a detection model. In [9] the baseline was outperformed by 3.2% which corresponds to the results achieved in our work.

Table 1. FROC metrics obtained from augmented and baseline model evaluation. Metrics mean and standard deviation are computed across 5-fold cross-validation experiments.

Average FP / crop	**0.25**	**0.5**	**1**	**2**	**4**	**8**	**Average**
Sensitivity (augmented)	**0.330±0.049**	**0.433±0.056**	**0.555±0.060**	**0.684±0.058**	**0.794±0.048**	**0.854±0.037**	**0.608±0.043**
Sensitivity (baseline)	0.302±0.042	0.414±0.040	0.542±0.041	0.647±0.030	0.743±0.020	0.822±0.036	0.578±0.028

Detection models can also help solve classification problems. Classification problem is defined as follows: positive instances are those containing at least one nodule, negative instances do not contain nodules. Making use of the detection model built upon 3D RPN, network output can be considered positive or negative depending on whether it contains any bounding box with probability score greater than some fixed threshold or not. Varying the threshold changes the number of average FP predictions per instance as well as false positive rate in the defined classification problem.

To evaluate the classification model ROC metrics have been obtained as shown in Figure 6. The best classifier (marked as a point on the ROC curve) reached sensitivity at 91% and specificity at 86%. ROC AUC metric reached 0.95.

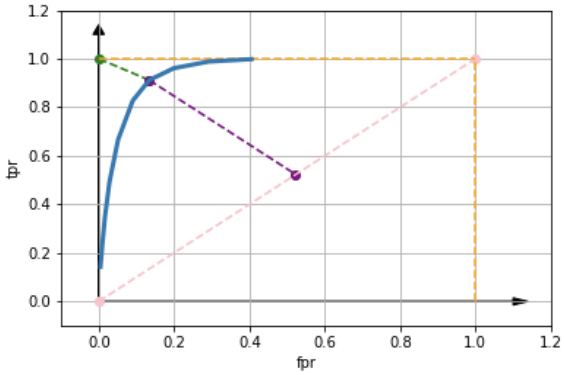

Figure 6. ROC curve illustrates the detection model transformed into classifier performance. Metrics were calculated across crops of size 128^3 obtained from the test set. Y-axis stands for TP rate i.e.sensitivity, X-axis stands for FP rate.

5. CONCLUSION

In this paper we proposed using Adaptive Instance Normalization and Wasserstein Loss as CT-GAN improvement for lung nodules generation on CT scans in 3D. The generated images were used for augmentation of LUNA16 dataset, which is the subset of LIDC-IDRI, and increased the size of the dataset by 30%. We evaluated the boost of our improvement based on nodules detection results. For the detection we used DeepSEED architecture that includes Squeeze-and-Excitation block and outperforms state-of-the-art models for this task. The experiments showed that our CT-GAN with AdaIN trained with Wasserstein Loss helped increase average sensitivity by 3%, giving 60.8% in total. This result is comparable with the work [9] and can be considered as the first attempt of applying AdaINs for lung nodules generation in CT images. The classifier based on the augmented images achieved 91% sensitivity and 86% of specificity.

As a direction for the future work two drawbacks of the current approach should be addressed, resolving which can significantly improve the results. The first one is that the CT-GAN-generated images have checkerboard artifacts. There are two main ways to combat this effect: upsampling stride tuning and the usage of some type of interpolation before convolution layer instead of upsampling or deconvolution [23]. The elimination of this effect can help to generate more plausible and diverse nodules. The second drawback is that a trivial way of inserting the generated nodule into the scan was implemented, specifically the generated nodule takes the place of the original one. Automatic recognition of the suitable place on the scan for injection of the generated nodule is not a simple task, but its solution in future can further diversify the augmented data, and thus contribute to nodule detection improvement.

ACKNOWLEDGMENTS

This work was financially supported by the Government of the Russian Federation, Grant 08-08, and Russian Science Foundation, Grant 19-19-00696.

REFERENCES

[1] S.G. Armato III, G. McLennan, et al. "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans." *Medical physics,* vol. 38. no. 2, pp. 915-931, 2011.

[2] X. Huang, and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization." *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[3] Y. Li, Y. Fan, DeepSEED: 3D squeeze-and-excitation encoder-decoder convnets for pulmonary nodule detection. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020.*

[4] A.A.A. Setio, A. Traverso, et al. "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA1616 challenge." *Medical image analysis,* vol. 42, pp. 1-13, 2017.

[5] C. Han, et al. "Learning more with less: Conditional PGGANbased data augmentation for brain metastases detection using highly-rough annotation on MR images." *In Proc. ACM International Conference on Information and Knowledge Management (CIKM*), 2019

[6] H. Shi, J. Lu and Q. Zhou, "A Novel Data Augmentation Method Using Style-Based GAN for Robust Pulmonary Nodule Segmentation," *2020 Chinese Control and Decision Conference (CCDC)*, Hefei, China, 2020, pp. 2486-2491, doi: 10.1109/CCDC49329.2020.9164303.

[7] Y. Mirsky et al. "CT-GAN: Malicious tampering of 3D medical imagery using deep learning." *28th {USENIX} Security Symposium ({USENIX} Security 19)*, pp. 461-478, 2019.

[8] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura. "CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation." *In Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp.732–740, 2018.

[9] C. Han et al. "Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection." *2019 International Conference on 3D Vision (3DV)*. IEEE, 2019.

[10] M. Arjovsky, S. Chintala and L. Bottou. "Wasserstein gan." *arXiv preprint arXiv:1701.07875,* 2017.

[11] C. Gao, et al. "Augmenting LIDC dataset using 3D generative adversarial networks to improve lung nodule detection." *Medical Imaging 2019: Computer-Aided Diagnosis*. Vol. 10950. International Society for Optics and Photonics, 2019.

[12] S. Liu, et al. "No Surprises: Training Robust Lung Nodule Detection for Low-Dose CT Scans by Augmenting with Adversarial Attacks." *arXiv preprint arXiv:2003.03824*, 2020.

[13] S. Ioffe, and C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." *arXiv preprint arXiv:1502.03167*, 2015.

[14] D. Ulyanov, A. Vedaldi, and V. Lempitsky. "Instance normalization: The missing ingredient for fast stylization." *arXiv preprint arXiv:1607.08022*, 2016.

[15] N. Khosravan, U. Bagci. "S4ND: single-shot single-scale lung nodule detection." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2018.

[16] W. Zhu et al. "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification." *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018.

[17] F. Liao, et al. "Evaluate the malignancy of pulmonary nodules using the 3-d deep leaky noisy-or network." *IEEE transactions on neural networks and learning systems* 30.11, pp. 3484-3495, 2019.

[18] J. Ding, et al. "Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, Cham, 2017.

[19] S. Ren, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." *Advances in neural information processing systems*, 2015.

[20] J. Hu, L. Shen, and G. Sun. "Squeeze-and-excitation networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.

[21] K. He, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[22] G. Wang, et al. "Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks." *Neurocomputing,* vol. 338, pp. 34-45, 2019.

[23] A. Odena, V. Dumoulin, and C. Olah. "Deconvolution and checkerboard artifacts." *Distill* 1.10: e3, 2016.