Comparative Evaluation of Lung Cancer CT Image Synthesis with Generative Adversarial Networks

Alexander Semiletov[1], Aleksandra Vatian[1], Maksim Krychkov[1], Natalia Khanzhina[1], Anton Klochkov[1], Aleksey Zubanenko[2], Roman Soldatov[2] , Anatoly Shalyto[1] and Natalia Gusarova[1]

[1] ITMO University, 49 Kronverksky Pr., St. Petersburg 197101, Russia
[2] Ltd International Diagnostic Center, 140 Leninsky Pr., St. Petersburg 198216, Russia
alexvatyan@gmail.com

Abstract. Generative adversarial networks have already found widespread use for the formation of artificial, but realistic images of a wide variety of content, including medical imaging. Mostly they are considered to be used for expanding and augmenting datasets in order to improve accuracy of neural networks classification. In this paper we discuss the problem of evaluating the quality of computer tomography images of lung cancer, which is characterized by small size of nodules, synthesized using two different generative adversarial network, architectures – for 2D and 3D dimensions. We select the set of metrics for estimating the quality of the generated images, including Visual Turing Test, FID and MRR metrics; then we carry out a problem-oriented modification of the Turing test in order to adapt it both to the actually obtained images and to resource constraints. We compare the constructed GANs using the selected metrics; and we show that such a parameter as the size of the generated image is very important in the development of the GAN architecture. We consider that with this work we have for the first time shown that for small neo-plasms, direct scaling of the corresponding solutions used to generate large neo-plasms (for example, gliomas) is ineffective. Developed assessment methods have shown that additional techniques like MIP and special combinations of metrics are required to generate small neoplasms. In addition, an important conclusion can be considered that it is very important to use GAN networks not only, as is usually the case, for augmentation and expansion of the datasets, but for direct use in clinical practice by radiologists.

Keywords: generative adversarial networks, 2D 3D GAN, CT image synthesis, evaluation metrics, lung cancer.

1 Introduction

Generative adversarial networks (GANs), first proposed in [1], have already found widespread use for the formation of artificial, but realistic images of a wide variety of content, including in medicine [2, 3]. Initially, GANs in the field of medicine were used as an aid for augmentation of datasets for processing medical images based on machine learning, primarily deep neural networks. But over the past two or three

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_49

years, the range of scenarios for the use of GANs in medicine has dramatically expanded. With the help of GANs, it is possible to form medical images of various natures, which can be used as reference images when setting up automated classifiers of corresponding diseases, as well as for training less experienced pathologists and radiologists. For example, the paper [4] reports on Amyloid Brain PET Image Synthesis, which simulates changes in brain tissue in Alzheimer's disease. In the work [5], using GAN, images of plaques in coronary arteries, which are the main cause of atherosclerosis, are simulated. The authors [6], using GAN to simulate histological sections of liver tissue, experimentally confirmed the possibility of not storing natural tissue sections in glass and completely switching to digital histopathology, which is important for definitive diagnosis of non-fatty liver damage.

Accordingly, the requirements for assessing the effectiveness of the use of GAN in medicine have expanded. Of the wide variety of metrics proposed for the GAN Evaluation [7], only a few of them are used in medical applications. Initially, this role was played only by indicators of the effectiveness of training deep neural networks on datasets augmented with the help of GAN - such as ROC, FROC, AUC ROC etc. But now, great importance is attached to the visual qualities of the generated images and the degree of their similarity to the simulated prototype. Therefore, assessments of visual similarity entered the everyday life of the medical GAN developers. For this, both model-based methods, such as the visual Turing test and t-SNE, and model-agnoctic metrics, such as MMD, LOO, FID, DFD etc, are used. (A more detailed description of the listed methods and metrics is given in the next section of the article).

As practice shows and as literature reviews [2, 3, 8, 9] confirm, in recent years there has been an active and rapid development of GAN models, both in terms of improving architectural solutions and in terms of taking into account the specifics of the target area of medicine

Adequate selection and problem-oriented adaptation of methods for assessing the quality of medical images generated with the help of GAN will allow not only assessing the prospects of a particular development at a fairly early stage, but also identifying key influence parameters that need to be highlighted during its further development. In oncological practice as a whole, the objects of interest are the neoplasms themselves and their structure, as well as the border zones between them and the surrounding tissues.

This article discusses the problem of evaluating the quality of CT images of lung cancer synthesized using GAN. The GAN methods are needed due to the fact that some specific cancer nodules are not presented well in the datasets like really small lung cancer nodules and etc. We describe our development of two problem-oriented GANs, which solve the problem of imitating malignant pulmonary nodes in 2D and 3D dimensions, respectively. We justify the selection of metrics for estimating the quality of the generated images, and of the parameters needed for their adaptation for lung tissue. For qualitative estimating we use Visual Turing test as well as d t-distributed Stochastic Neighbor Embedding (t-SNE). As quantitative metrics we use inception distance (FID). To evaluate the performance of the classifier trained on the basis of the augmented dataset, we use ROC-metrics. We present the results of a comparative evaluating of the

DOI: 10.1007/978-3-030-77967-2_49

constructed GAN using the selected metrics and show that from this point of view the development of the 2D approach seems to be more effective.

2 Background and Related Works

2.1 GAN's Specifics for Lung Tissue Imitation

Initially [1], the GAN scheme was a generator of random objects and a discriminator that distinguishes these objects from real ones, and a feedback loop was organized between both blocks through a back propagation mechanism. Modern GANs used in medical imaging, are far more sophisticated and more diverse in architecture [2], however, a problem-oriented analysis of the literature of recent years allows us to single out a number of main dominants.

Although the literature presents scenarios for generating medical images using GANs "from scratch", i.e. from random noise without any other conditional information [5, 10], however, in the last years the use of GANs for lung tissue imitation prevails in domain transformation scenarios, i.e. in image-to-image translation frameworks [11–16].

In general, in order to generate high-quality medical images, 3D GANs are applied, i.e. 3D fragments containing nodules and surrounding tissues are used for training [6, 13–15, 17]. This approach, of course, makes rather high demands on the complexity of the architecture and the level of computing resources. Meanwhile, the specifics of pulmonary nodes on CT scans is their rather small dimensions, i.e. a few (up to 2–3) slices of the CT image are enough to display an particular node. Therefore, in recent studies, there is also a 2D approach to generating images of pulmonary nodes [18]. The Maximum Intensity Projection (MIP) approach [19, 20] looks promising here. MIP is a method that projects 3-D voxels with maximum intensity to the plane of projection, thereby providing a transition to a 2D task.

As for up-to-date architecture for simulating pulmonary nodes, a plethora of variants are proposed here. Compared to the vanilla GAN [1], they apply various modifications of the loss function and approaches to normalization, as well as their combinations. For example, [11] employs WGAN, where the loss function is defined using the Wasserstein distance instead of the Jensen–Shannon divergence, [16] uses WGAN-GP, i.e. Wasserstein GAN with gradient penalty. The work of [15] is based on Conditional GAN (CGAN), where a discriminator is conditioned on an additional input. In [13] 3D-Multiconditional GAN is proposed containing two discriminators with different loss functions tailored for nodules and for context (surrounding tissues). In [17] a specialized CT-GAN based on a Conditional GAN is proposed.

[10] proposed to use Deep Convolutional Generative Adversarial Networks (DC-GANs), augmenting the standard GAN by using convolutional layers along with batch normalization. A wide spread solution for the generation of lung tissue is Progressive GAN, where the GAN is sequentially trained to create images of increasing dimension. For example, [12] implemented a Progressive Growing WGAN (PGGAN), with sliced Wasserstein distance loss, progressive growing, and pixel-wise normalization. [18] proposes a combined solution named Conditional Progressive

Growing of GANs (CPGGANs), incorporating highly-rough bounding box conditions incrementally into PGGANs. Based on Progressive GAN and adding the Adaptive Image Normalization (AdaIN), [21] developed StyleGAN architecture, which has been successfully used to generate medical images of different kind [6, 22], including pulmonary nodes [14].

2.2 Metrics for Evaluating the Quality of Synthesized Images

A short list of metrics fetched in the current literature for evaluating the quality of medical images, was presented earlier in the Introduction section. Now we present their consideration in more detail with an emphasis on their applicability to lung tissue imaging assessment.

Measures for integral effectiveness of CNN-based classifier. There are well-known indicators of the effectiveness of training deep neural networks on datasets augmented with the help of GAN - such as F-measure and its components, ROC, FROC, AUC ROC etc. Typically, they answer slightly different research questions. For example, the ROC method only involves stating the presence of an anomaly in the image, while the FROC method additionally requires the observer to detect anomalies [23]. However, as the analysis shows, in relation to the classification of lung cancer, these metrics are spread almost evenly in the literature. For example, as concerning to F-measure components, [10] uses False Recognition Rate (FRR) and True Recognition Rate (TRR), [11] applies accuracy, [24] and [15] estimate F-measure as a whole, [18] uses sensitivity in diagnosis with clinically acceptable additional False Positives (FPs), [16] ajustes False Positive per Scan vs Sensitinity. Such a diversity undoubtedly complicates a comparative assessment of the proposed solutions. At the same time, most researchers of the GUN-assisted lung cancer classification use ROC-curve [11, 12, 24] or FROC-curve as its variation [13].

Model-based methods. For an expert assessment of the synthetic images' realism, the Visual Turing Test is proposed in [25]. The full Visual Turing Test involves presenting real and virtual (generated) images to experienced radiologists, comparing them with the ground truth, and constructing a contingency table. However, the full procedure assumes a large amount of available statistical material for comparison [6], which is obviously lacking in pilot studies. Therefore, in most studies related to generative imaging of lung tissue, the Turing test is performed with different truncations without special justification [10, 12, 13, 15, 18] or not performed at all [11, 16].

One more model-based method is visualizing the data distribution via t-Distributed Stochastic Neighbor Embedding (t-SNE) [26]. The t-SNE method allows you to visually compare the distributions of real and generated images by translating high-dimensional data into a lower-dimensional space. Despite the rather low requirements for experimental and computational resources, as well as high information content [27, 28], the method is still used relatively rarely to assess the quality of lung imaging [12, 18].

Model-agnoctic metrics. According to [7, 29], in principle, the following model-independent metrics can be used to assess the quality of medical images generated by GANs.

DOI: 10.1007/978-3-030-77967-2_49

The 1-NN classifier [30] assesses the similarity of the real image (labeled as 0) and the generated images (labeled as 1) by Leave-one-out cross-validation (LOOCV). However, this technique involves preliminary labeling of the images synthesized, for which, as the authors [30] themselves note, one must employ a “naturalness discriminator”. For the time being, only a human expert can play this role for medical images, i.e. the technique becomes very resource-intensive and of little use for pilot projects.

The Maximum Mean Discrepancy (MMD) [31] is a distance-measure between two distributions which is defined as the squared distance between their embeddings in the a reproducing kernel Hilbert space F, i.e. is the distance between feature means μ_P and μ_Q of compared image data X and Y having probability measures P and Q respectively:

$$MMD^2(P,Q) = \|\mu_P - \mu_P\|_F^2. \tag{1}$$

The lower the result the more evidence that distributions are the same. Note that in CNN practice, when fulfilling the empirical estimation of MMD one usually limits to simple kernel functions [32].

The Fréchet inception distance (FID) [33] calculates the distance between the feature vectors of real images $N(\mu, \mathrm{C})$ and of images generated by the GAN $N(\mu_w, \mathrm{C}_w)$. FID is based on the assumption of multidimensional Gaussian distributions of real and generated images, i.e. the mean and standard deviation is compared, so the formula is as follows:

$$d^2((m,C),(m_w,C_w)) = \|m - m_w\|_2^2 + Tr\left(C + C_w - 2\left(CC_w\right)^{1/2}\right), \tag{2}$$

where $Tr(.)$ is the trace of the covariance matrices of the feature vectors C and C_w.

Note that the comparison both in (1) and in (2) is not performed on the image itself, but on one of the deeper layers of CNN, which allows us to get away from the human perception of similarity in images. On the other hand, when using the Gaussian kernel function, expression (1) coincides with the first term of expression (2) up to notation, that is, FID and MMD metrics seem to some extent interchangeable. However, in the case of a strong discrepancy in the distributions, the use of the FID metric is criticized [5, 22], but for the assessment of lung imaging by the GANs, it is still used [12] as opposed to MMD.

Summarizing the review of the use of GAN for generating images of lung tissue, we see a motley picture of various architectures and approaches of GAN implementing, as well as of assessing the quality of the images obtained. This complicates the comparability of different projects and the possibility of assessing the prospects of the newly conducted research even at the pilot stage. In this regard, the authors of the article set themselves the following tasks:

- to carry out a pilot development of two GANs with different architectures and dimension approach, designed to generate images of lung tissue with cancerous nodes, under resource constraints;
- to select the set of metrics for estimating the quality of the generated images, and of the parameters needed for their adaptation for lung tissue;

- to compare the constructed GANs using the selected metrics and to identify key influence parameters that need to be highlighted during its further development.

3 Methods and Materials

3.1 Developing and Training GAN's Models

For comparison, we have developed two variants of GANs, which differ both in architecture and in approach of image forming (in 3D or 2D projection).

For 3D-GAN we chose CT-GAN from [17] as the baseline, but made some changes to it, which involved generator and discriminator weight updates frequency ratioing and combining Wasserstein Loss (WL) with a baseline Mean Square Error (MSE) loss. Besides, we used AdaIN instead of Batch Normalization after each convolutional block, the normalization parameters being

$$AdaIN(x,y) = \sigma(y)\frac{x-\mu(x)}{\sigma(y)}, \quad (3)$$

where x is the previous layer output, y is the affine transformation. The modified blocks of 3D-GAN are depicted on Figure 1a. As a detection model we used the solution of [34].

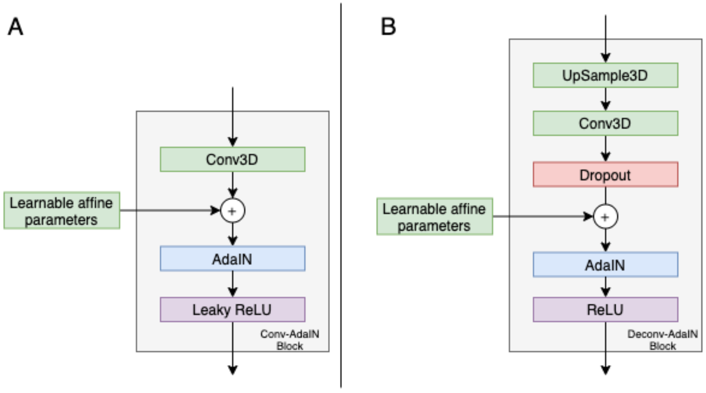


Fig. 1. The blocks of 3D-GAN modified in comparison of [Mirsky]: (A) Convolutional block, (B) Deconvolutional block.

For 2D-GAN we chose Syle-GAN [21] as the baseline fulfilling the MIP approach within. Since implementation of [21] is designed to work with 3-channel images, we converted the number of channels in the input and output layers to work with single-channel images. We also changed the Loss function type to BinaryCrossEntropy. We used the VGG11 model [35] being undemanding in terms of computing resources as a classifier.

To train both GAN models, we used the open Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset [36]. For 3D-GAN we employed its subset, namely LUNA-16 having the following properties compared to the base LIDC dataset: all the nodules are calibrated more precisely (the average size

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_49

of nodules in LUNA16 is 8.3 mm with a standard deviation of 4.8 mm) compared to LIDC-IDRI (12.8 mm and 10.6 mm, respectively); each nodule is already labeled with a bounding box.

For 2D-GAN training we used LIDC-IDRI in general, but selecting the DICOM series only with tumor nodules. In order to extract the nodule we formed a circumscribing cube containing the nodule itself and the surrounding (context) tissues. The absolute dimensions of the cube were chosen in accordance with the size of the extracted nodule (from 1 mm to 40 mm), but were resampled to a single size of 128^3 pixels and to pixel values according to Hounsfield scale [37]:

$$x = \frac{2 \cdot (x_{in} - in_{\min})}{((in_{\max} - in_{\min}) - 1)}, \qquad (4)$$

with boundary values of in_{max}=800, in_{min} = –1000. In order to pass from 3D to 2D nodule image we performed a MIP lookup operation using the *numpy* package.

3.2 Evaluating the Quality of Synthesized Images of Lung Tissues

Measures for integral effectiveness. Although both GANs form images of nodules in the lungs, 3D-GAN is more focused on augmentation of datasets used in training DNN-based neoplasm detectors, while 2D-GAN is more focused on augmentation of datasets used in training DNN-based classifiers of neoplasm types. In this regard, in order to evaluate the classification model in both cases, we use ROC AUC (Area Under ROC Curve) metrics as measures for integral effectiveness of training DNN on datasets augmented with the help of GAN. In addition, for evaluation of 2D-GAN-based classifier we additionally use PR AUC (Precison-Recall Area Under Curve), and for 3D-GAN-based detection we use FROC in a modified form: instead of per-scan FROC calculations [34] 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.

Model-based methods. We used the Visual Turing Test, but made its problem-oriented modification. As with the traditional Visual Turing Test (see above), real and virtual (generated by GAN) images are demonstrated to *N* experienced radiologists. Each radiologist is presented with *S* sets containing 20 randomly selected images generated by a specific GAN, and we inform the radiologist that the exposed set can contain any mixture of real and generated images. Examples of the presented sets are shown in Fig. 2. The radiologist is asked to answer the following questions:

- If the presented set contains nodules, then which ones are real?
- If the presented set contains nodules, then which of them are solid (single), and which are subsolid (parietal)?

Based on the test results, the False Recognition Rate (FRR) is calculated as the proportion of nodules correctly identified by radiologists as generated among all generated nodes.

We calculated the t-SNE metric using the *scikit* package with default parameters[1].

[1] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_49

We selected FID and MMD as **model-agnoctic metrics.** For implementation the FID metric, we used the code[2] with default parameters. The implementation of MMD is made according to the work[3]; as kernel functions we used radial basis function (instead of Gaussian kernel function, which is a standard practice for empirical estimates for CNN) and multiscale function.

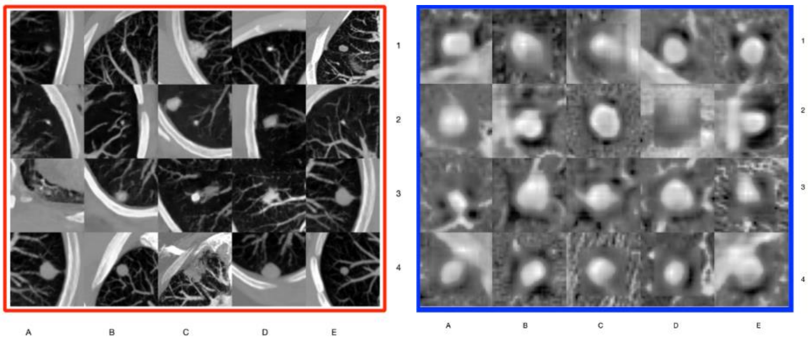


Fig. 2. a - An example of presentation for the evaluation of 2D-GAN. Images 1-10 contain nodules generated by 2D-GAN, images 11-20 are completely real. b - An example of presentation for the evaluation of 3D-GAN. All images contain generated nodules.

4 Experimental Results

Figure 3 shows the ROC-curves for the best learning epochs for 2D-GAN (a) and for 3D-GAN (b) respectively. For 3B-GAN we also give the FROC metric (Table 1).

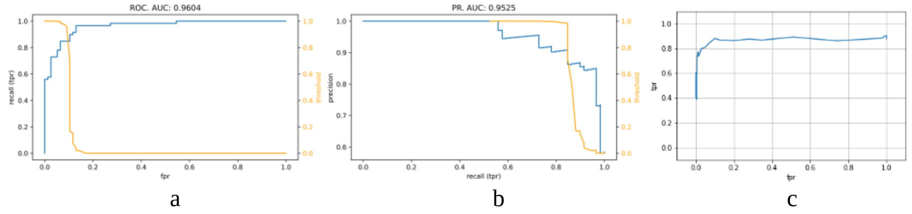


Fig. 3. ROC- curves for for 2D-GAN (a, b) and 3D-GAN (c)

Using the proposed approach of 2D-GAN for dataset augmentation we obtained the best values of ROCAUC=0.9604 and PRAUC=0.9625. This is better than the result of [11] on a comparable dataset and is only slightly inferior to the result of the same authors, obtained on much more powerful computing resources [38]. For 3D-GAN, the best value of ROCAUC=0.95. So (see also Table 1), the proposed 3D-GAN also surpasses the most modern of similar GAN implementations [13], chosen as a baseline.

[2] https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch/

[3] https://www.kaggle.com/onurtunali/maximum-mean-discrepancy

Table 1. FROC metrics obtained for 3D-GAN 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

When performing the Visual Turing Test, we recruited N = 6 radiologists, each of whom were presented with S = 20 test sets of 20 images each. Tables 2, 3 show examples of experimental data obtained from two radiologists. Table 4 contains the results of calculating the FRR metric, carried out by averaging over 6 radiologists, over 2 radiologists, and also for each of the two selected radiologists separately. Table 5 contains the analogous results concerning identifying of solid and subsolid nodules.

Table 2. Examples of the results of the Visual Turing Test.

2D-GAN	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Radiologist 1	30	15	20	20	15	10	10	20	20	20	40	25	20	10	10	20	40	15	30	20
Radiologist 2	20	15	10	20	20	20	15	20	15	10	20	10	20	40	30	20	30	25	30	25
Real / generated nodules in the set	10/ 10	0/ 20	10/ 10	10/ 10	0/ 20	10/ 10	0/ 20	10/ 10	0/ 20	10/ 10	10/ 10	0/ 20	0/ 20	10/ 10	10/ 10	10/ 10	10/ 10	0/ 20	10/ 10	0/ 20
3D-GAN	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Radiologist 1	25	20	40	30	25	20	20	30	50	40	15	20	25	10	20	15	20	30	40	30
Radiologist 2	30	25	70	80	70	80	20	30	50	40	55	30	35	20	40	25	50	60	40	60
Real / generated nodules in the set	0/ 20	0/ 20	10/ 10	10/ 10	0/ 20	10/ 10	0/ 20	10/ 10	10/ 10	10/ 10	0/ 20	10/ 10	0/ 20	0/ 20	10/ 10	0/ 20	10/ 10	10/ 10	10/ 10	0/ 20

Table 3. FRR metrics calculated for different groups of radiologists.

	Radiologist 1	Radiologist 2	Average for 2 radiologists	Average for 6 radiologists
2D-GAN, S=20	20,5±8,6	20,7±7,4%	20,6±8%	19,4±5,4%
2D-GAN, S=5	20±5,4%	17±4%	18,5±4,7%	17,9±7,2%
3D-GAN, S=20	45,5±18,9%	26,25±9%	35,8±14%	41 ±11%
3D-GAN, S=5	28±6,7%	55±22%	41,5±14,3%	47±15,3%

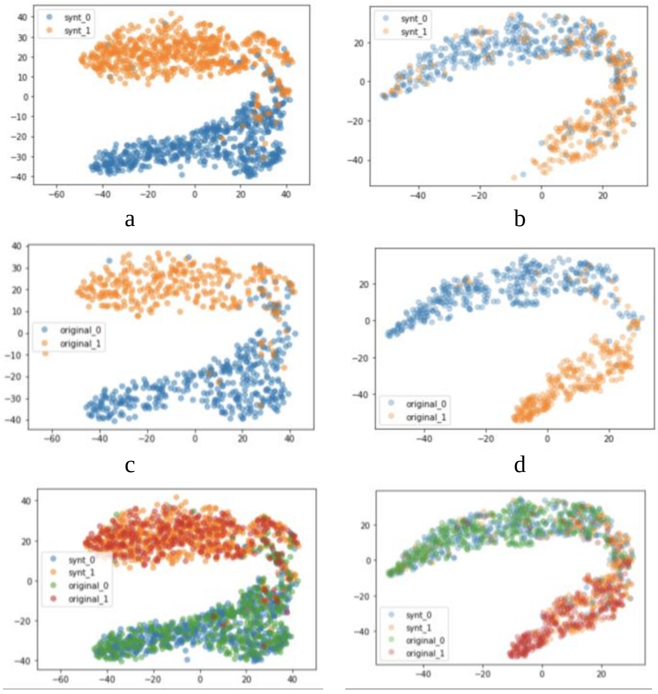
Table 4. FRR metrics calculated for identifying of solid and subsolid nodules

	Radiologist 1	Radiologist 2	Average for 2 radiologists	Average for 6 radiologists
2D-GAN	92,5±7%	94,6±5%	93,6±6%	94,5±3,5%
3D-GAN	27,5±19,5%	30,3±17.4%	28,9±18.5%	25.4±19.6%

Figure 4 shows the visualized t-SNE metrics for different combinations of pulmonary nodes (benign and malignant; real and virtual), as well as their summary, for 2D and 3D GANs. Table 6 summarizes the experimental data for the FID metrics, as well as for the MMD-metrics plotted for various kernel functions. The following designations are adopted in the Table 6: r-v is the value of the metric between real and synthesized data as a whole; rb-vb is the value of the metric between real and synthesized images of benign nodules; rm-vm is the value of the metric between real and synthesized images of malignant nodules. The advantages of using FID and MRR metrics is that we can make a comprehensive assessment of the generated nodules. By using Visual Turing Test we can carry out an integral assessment according to the context, taking into account the opinion of the doctor.

Table 5. FID and MMD metrics calculated for different groups of images (designations decoding - in the text)

	2D-GAN			3D-GAN		
	r-v	rb-vb	rm-vm	r-v	rb-vb	rm-vm
FID	51.61	74.98	37.35	171.10	130.93	670.66
MMD, radial basis kernel	0.0169	0.0214	0.0287	0.0193	0.0252	0.0375
MMD, multiscale kernel	0.0118	0.0208	0.0271	0.0123	0.0212	0.0280



a b c d e f

Fig. 4. Visualized t-SNE metrics for various combinations of embeddings of images with pulmonary nodes: a – benign vs malignant nodules, generated by 2D-GAN; b - benign vs malignant nodules, generated by 3D-GAN; c- benign vs malignant nodules, real images; d- benign vs malignant nodules, real images; e - all the nodules of dataset augmented by 2D-GAN; f - all the nodules of dataset augmented by 3D-GAN;

DOI: 10.1007/978-3-030-77967-2_49

5 Discussion

The experimental data (see Fig. 4 and Table 1), as well as a comparison of these data with baselines show that from the point of view of augmentation of datasets used for CNN training, the developed GANs are at a completely conventional level, i.e. the chosen architectural and parametric solutions are quite successful. Note that 3DGAN was trained in obviously better conditions, since the nodule detection on LUNA16 is considered less challenging than on LIDC-IDRI due to the greater reliability of the data and smaller scatter in the size of nodules [17].

Let's move on to the analysis of the visual quality of the generated images (Fig. 2, 3 and Tables 2-5). The number of radiologists performing a Visual Turing Test and their professional experience will obviously affect test results. Despite this, in most works related to the visual assessment of the quality of images generated by the GAN, the authors limit themselves, at best, to a statement of the number of experts involved, and, as a rule, there are only two [12, 13] or one [22] of them, and in [28], even an expert radiologist and a non-specialist are used. The results of our experiments, presented in Table 4, show that when passing from averaging the FRR values over 6 experts to averaging over 2 experts, the mathematical expectation, although it changed, remained within the standard deviation, and for some experts it was the significant change. This allows us to say that in pilot studies of the prospect of the GAN architecture, one can limit ourselves to two experts-radiologists, and the use of only one radiologist and, moreover, a non-specialist does not guarantee against the presence of emissions in the assessment and cannot be justified. At the same time, for large-scale studies, it is necessary to use stronger statistical measures of similarity, for example, the Fleiss kappa [39].

Analyzing the model-agnoctic metrics of image quality (Table 6 in comparison with Fig. 4), we can state that the MMD and PID metrics demonstrate different rankings of the quality of the resulting images. In this case, a change in the kernel function, although it affects the absolute values of the MMD metric, does not change its ratio for different groups of images. The gradation of the metrics is different - the FID metric, unlike MMD, takes into account not only average values, but also data variance. For example, on the t-SNE visualization (Figure 4, Figure c), it can be seen that the real data has several outliers. Because of them, the average of real and generated data is different, but this practically does not affect the variance. And therefore, in this case, the FID metric can be considered fairer than MMD metric. In addition, it is necessary to emphasize the independent role of the t-SNE metric in assessing the quality of the generated images of pulmonary nodules, since it allows you to visually assess the degree of interpenetration of the generated objects of different classes (compare Fig. 4, a and b).

Comparing the developed GANs using the selected metrics, we can draw the following conclusions. Although, as noted above, both GANs have a fairly high efficiency in terms of augmentation of datasets for CNN training, nevertheless, as can be seen from Table 3 and 4, as well as from Fig. 1 and 2, the visual quality of the images generated by 3D-GAN differ significantly. We can associate this with the fact that most of the existing GAN are focused on the generation of relatively large neoplasms, and

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_49

the problem of their inserting into existing images is reduced to eliminating defects at the boundaries between the formed image and the substrate (native tissue) (see, for example, [6].

In the case of small nodules, as shown in Fig. 1 and 2, this effect is not observed, i.e. it is not possible to clearly identify the zones of the generated image that are most critical for visual assessment. Thus, direct scaling of solutions for large neoplasms to small sizes is hardly effective, and imitation of small neoplasms should be considered as a separate task when building a GAN. In our case, this was done using the MIR approach, but, of course, other options are also possible here. So, we consider the relative size of the simulated medical neoplasm to be one of the important influence parameters that need to be highlighted and estimated during the development of GUN architecture for medical purposes.

6 Conclusion

The specificity of lung cancer is that neoplasms are malignant nodules, which have small size (10-30 mm) and high morphological similarity with benign nodules normally present in the lungs. In this regard, according to the Lung Image Database Consortium, LIDC) [36], even experienced radiologists correctly classify only 75.1% of the nodes when compared with the results of biopsy. Therefore, there is a great need for an early assessment of the prospects for the development of one or another technology for the implementation of GAN for this area of medicine. This article is a contribution to solving this problem.

All the tasks set in the article have been solved, namely:

- we carried out a pilot development of two GANs of up-to-date level with different architectures and dimension approach, designed to generate images of lung tissue with cancerous nodes, under resource constraints;
- we selected the set of metrics for estimating the quality of the generated images, including Visual Turing Test, FID and MRR metrics; we carried out a problem-oriented modification of the Turing test in order to adapt it both to the actually obtained images and to resource constraints;
- we compared the constructed GANs using the selected metrics; we showed that such a parameter as the size of the generated image was very important in the development of the GAN architecture.

The proposed work may be useful for providing a baseline for future studies and implementing GANs for medical purposes, as well as for determining the direction of next future experiments in practical studies.

The work was supported by Russian Science Foundation, Grant 19-19-006-96

References

1. Goodfellow I.J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative Adversarial Nets. arXiv:1406.2661v1 [stat.ML] 10 Jun 2014.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_49

2. Tschuchnig M. E., Oostingh G.J., and Gadermayr M. Generative Adversarial Networks in Digital Pathology: A Survey on Trends and Future Potential. Patterns REVIEW. Vol. 1, Is. 6, 100089, Sept. 11, 2020.
3. Yi X., Walia E., Babyn P. Generative adversarial network in medical imaging: A review. Medical Image Analysis Volume 58, December 2019, 101552
4. Kang H., et al. Visual and Quantitative Evaluation of Amyloid Brain PET Image Synthesis with Generative Adversarial Network. Appl. Sci. 2020, 10, 2628.
5. Bargsten L., and Schlaefer A. SpeckleGAN: a generative adversarial network with an adaptive speckle layer to augment limited training data for ultrasound image processing. International Journal of Computer Assisted Radiology and Surgery (2020) 15:1427–1436.
6. Levy J.J., Azizgolshani N., Andersen M.J.Jr., et al. A large-scale internal validation study of unsupervised virtual trichrome staining technologies on nonalcoholic steatohepatitis liver biopsies. Springer Nature. Modern Pathology. Published online: 09 December 2020
7. Borji A. Pros and Cons of GAN Evaluation Measures. arXiv:1802.03446v5 [cs.CV] 24 Oct 2018
8. Kazeminia S., Baur C., Kuijper A., van Ginneken B., Navab N., Albarqouni S., Mukhopadhyay A. GANs for Medical Image Analysis. arXiv.org > cs > arXiv:1809.06222v3. 9 Oct 2019.
9. Wang T., Lei Y., Fu Y., Wynne J.F., Curran W.J., Liu T., Yang X. A review on medical imaging synthesis using deep learning and its clinical applications. Journal of Applied Clinical Vedical Physics, Vol. 22, Is. 1, Pp. 11-36, January 2021.
10. Chuquicusma M.J.M., Hussein S., Burt J., Bagci U. How to Fool Radiologists with Generative Adversarial Networks? A Visual Turing Test for Lung Cancer Diagnosis. IEEE International Symposium on Biomedical Imaging (ISBI) 2018.
11. Onishi Y., Teramoto A., Tsujimoto M., Tsukamoto T., Saito K., Toyama H., Imaizumi K., and Fujita H. Automated Pulmonary Nodule Classification in Computed Tomography Images Using a Deep Convolutional Neural Network Trained by Generative Adversarial Networks. Hindawi BioMed Research International. Vol. 2019, Article ID 6051939.
12. Wang Y., Zhou L., Wang M., et al. Combination of generative adversarial network and convolutional neural network for automatic subcentimeter pulmonary adenocarcinoma classification. Quantitative Imaging in Medicine and Surgery 2020;10(6):1249-1264
13. Han C., 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), Quebec City, QC, Canada, 2019 pp. 729-737. doi: 10.1109/3DV.2019.00085
14. Shi H., Lu J. and Zhou Q., 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.
15. Jin D., Xu Z., Tang Y., Harrison A.P., and Mollura D.J. CT-realistic lung nodule simulation from 3D conditional generative adversarial networks for robust lung segmentation. Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp.732–740, 2018
16. Gao C., 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
17. Mirsky Y. et al. CT-GAN: Malicious tampering of 3D medical imagery using deep learning. 28th {USENIX}. Security Symposium ({USENIX} Security 19), pp. 461-478, 2019.
18. Han C., 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.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77967-2_49

19. Zhang J., Xia Y., Zeng H., Zhang Y. NODULe: Combining constrained multi-scale LoG filters with densely dilated 3D deep convolutional neural network for pulmonary nodule detection. Neurocomputing, vol. 317, pp. 159-167, 2018.
20. Zheng S., Guo J., Cui X., Veldhuis R.N.J., Oudkerk Matthijs, van Ooijen P.M.A. Automatic Pulmonary Nodule Detection in CT Scans Using Convolutional Neural Networks Based on Maximum Intensity Projection. arXiv:1904.05956 [cs.CV] 10 Jun 2019.
21. Karras T., Laine S., Aila T. A Style-Based Generator Architecture for Generative Adversarial Networks. arXiv:1812.04948v3 [cs.NE] 29 Mar 2019
22. Chang A., Suriyakumar V.M., Moturu A., et al. Using Generative Models for Pediatric wbMRI. Medical Imaging with Deep Learning 2020 1–7
23. Hillis S.L., Chakraborty D.P., Orton C. G. ROC or FROC? It depends on the research question. Medical Physics, Vol. 44, Is. 5, May 2017, Pp 1603-1606.
24. Ghosal S.S., Sarkar I., Hallaoui I.E. Lung nodule classification using Convolutional Autoencoder and Clustering Augmented Learning Method (CALM). http://ceur-ws.org/Vol-2551/paper-05.pdf, last access 05.02.2021.
25. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training GANs. In Advances in Neural Information Processing Systems (NIPS), pages 2234–2242, 2016.
26. van der Maaten L. and Hinton G. Visualizing data using t-SNE. J. Mach. Learn. Res., 9:2579–2605, 2008.
27. Kang H., et al. Visual and Quantitative Evaluation of Amyloid Brain PET Image Synthesis with Generative Adversarial Network. Appl. Sci. 2020, 10, 2628.
28. Haarburger C., Horst N., Truhn D., Broeckmann M., Schrading S., Kuhl C. and Merhof D. Multiparametric Magnetic Resonance Image Synthesis using Generative Adversarial Networks. Eurographics Workshop on Visual Computing for Biology and Medicine (2019)
29. Xu Q.,; Huang G., Yuan Y, Guo C., Sun Y., Wu F.,Weinberger K. An empirical study on evaluation metrics of generative adversarial networks. arXiv 2018, arXiv:1806.07755.
30. Lopez-Paz D. Oquab M. Revisiting classifier two-sample tests. arXiv 2016, arXiv:1610.06545.
31. Gretton A., Borgwardt K.M., Rasch, M.J., Schölkopf B., Smola A. A kernel two-sample test. J. Mach. Learn. Res. 2012, 13, 723–773.
32. Dziugaite G.K., Roy D.M., and Ghahramani Z. Training generative neural networks via maximum mean discrepancy optimization. arXiv preprint arXiv:1505.03906, 2015.
33. Dowson D.C., Landau B V (1 September 1982). The Fréchet distance between multivariate normal distributions. Journal of Multivariate Analysis. 12 (3): 450–455.
34. Li Y., Fan Y., DeepSEED: 3D squeeze-and-excitation encoder-decoder convnets for pulmonary nodule detection. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, 2020.
35. Simonyan K., Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556 [cs.CV] 10 Apr 2015
36. Armato S.G. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scan. Med Phys. 2011 Feb; 38(2): 915–931.
37. Feeman T.G. (2010). The Mathematics of Medical Imaging: A Beginner's Guide. Springer Undergraduate Texts in Mathematics and Technology. Springer. ISBN 978-0387927114
38. Onishi Y., Teramoto A., Tsujimoto M., et al. Multiplanar analysis for pulmonary nodule classification in CT images using deep convolutional neural network and generative adversarial networks. Int J Comput Assist Radiol Surg. 2020 Jan;15(1):173-178. Epub 2019 Nov 16.
39. Gwet K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, Vol. 61, pp. 29–48