Interpreting Latent Spaces of Generative Models for Medical Images using Unsupervised Methods

Julian Schön[1] JULIAN.E.S@DI.KU.DK
Raghavendra Selvan[1,2] RAGHAV@DI.KU.DK
Jens Petersen[1,3] PHUP@DI.KU.DK
[1] *Department of Computer Science, University of Copenhagen, Denmark*
[2] *Department of Neuroscience, University of Copenhagen, Denmark*
[3] *Department of Oncology, Rigshospitalet, Denmark*

Editors: Under Review for MIDL 2022

Abstract

Generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) play an increasingly important role in medical image analysis. They are used to synthesize, de-noise, super-resolve, and augment medical images. The latent spaces of these models often show semantically meaningful directions corresponding to human-interpretable image transformations. However, until now, their exploration for medical images has been limited due to the requirement of supervised data. Recently, several methods for unsupervised discovery of interpretable directions in GAN latent spaces have shown interesting results on natural images. This work explores the potential of applying these techniques on medical images by training a deep convolutional GAN and a VAE on thoracic CT scans and using an unsupervised method to discover interpretable directions in the resulting latent space. We find several directions corresponding to non-trivial image transformations, such as rotation or breast size, as well as directions showing that the generative models capture 3D structure despite being presented only with two-dimensional data. The results show that unsupervised methods to discover interpretable directions in generative model latent spaces generalize to VAEs and can be applied to medical images. This could open a wide array of future work using these methods in medical image analysis.

Keywords: Generative models, unsupervised learning, interpretability, CT

1. Introduction

The combination of deep learning and medical images has emerged as a promising tool for diagnostics and treatment. Though the amount of available data is increasing, one of the main limitations is the often small dataset sizes available to learn from. This is due to reasons such as the high costs of collecting and labeling data, adverse effects of radiation exposure from imaging procedures, and protection of sensitive patient data. Generative models can be used to synthesize or augment medical images (Nie et al., 2017; Frid-Adar et al., 2018; Hiasa et al., 2018; Zhu et al., 2019), mitigating some of these factors.
Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have emerged as the prominent generative model for image synthesis in recent years. Consequently, an extensive line of research focusing on the interpretability of GANs has unfolded. Lately, there has been a focus on the structure and interpretability of the latent space learned by GANs. Radford et al. (2016) showed that there is meaningful vector arithmetic in the latent space

of Deep Convolutional Generative Adversarial Networks (DCGANs). This led to the investigation of interpretable directions in GAN latent spaces. For several years, the methods used for discovering interpretable directions in latent spaces have been supervised (Goetschalckx et al., 2019; Shen et al., 2020; Jahanian et al., 2020; Plumerault et al., 2020). Especially in medical image analysis, supervision is expensive as it typically involves radiologist's or other expert's time. Recently, several unsupervised methods for the discovery of interpretable directions in GAN latent spaces were presented (Voynov and Babenko, 2020; Härkönen et al., 2020; Shen and Zhou, 2021). Since these methods do not require supervision, they seem more promising for the medical domain. However, at present it is unclear if they work with the often more homogeneous images and the smaller dataset sizes encountered in this field.
Variational Autoencoders (VAEs) (Kingma and Welling, 2014) are another popular class of generative models that explicitly approximate the data distribution (Goodfellow, 2016). Research on the interpretability of VAEs has mainly focused on obtaining disentangled latent space representations (Higgins et al., 2017; Kim and Mnih, 2018). While this shows promising results, there are limitations to unsupervised learning of disentangled representations, in that it might not be possible without introducing inductive biases (Locatello et al., 2019). The approach of Voynov and Babenko (2020), does not restrict the latent space representation as it is trained post-hoc. Thus, this and similar methods developed for GANs, when applied to VAEs, allow for latent spaces that need not incorporate additional inductive biases. This allows for easier applications of discovered directions to real images. If the same methods that have shown promising results on GANs are effective on VAEs, then VAEs can be trained without restrictions on the latent space while still having the benefit of interpretability and applicability to real images.
Contributions: We employ an unsupervised technique to explore the latent spaces of DCGANs and VAEs trained on Computed Tomography (CT) images[1]. We show that these methods previously used to interpret the latent spaces of GANs generalizes to VAEs. Further, our results provide insights into the applicability of these methods for medical image analysis. We evaluate the directions obtained and show that there are non-trivial and semantically meaningful directions encoded in the latent space of the generative models under consideration. These directions include both anatomical transformations specific to our dataset choice and geometric transformations that likely generalize to other data.

2. Background

2.1. Generative Adversarial Networks

GANs (Goodfellow et al., 2014) are a class of deep generative models that implicitly model the data generating distribution (Goodfellow, 2016). They optimize a zero-sum game between a generator neural network that synthesizes new data from random samples and a discriminator neural network that classifies real and synthesized data. The generator, $G : z \in \mathbb{R}^L \rightarrow x \in \mathcal{D}$, maps from a L-dimensional latent variable z to the data space $\mathcal{D}$. The discriminator, $D : x \in \mathcal{D} \rightarrow \{0, 1\}$, presented with $G(z)$ and real samples, is tasked with classifying them as real or fake.
Given the latent distribution p_z, the data distribution p_{data}, and binary cross-entropy as

1. https://github.com/julschoen/Latent-Space-Exploration-CT

the loss we can define the min-max game as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]. \tag{1}$$

The introduction of DCGANs (Radford et al., 2016) has permitted more effective application of GANs on images. In particular, there is a vast amount of work showing the effectiveness of DCGANs in medical imaging (Zhang et al., 2018; Diaz-Pinto et al., 2019; Alyafi et al., 2020; Bushra and Shobana, 2020; Fujioka et al., 2019).

2.2. Variational Autoencoders

VAEs (Kingma and Welling, 2014) are another popular class of deep latent variable models used for generative modelling and approximating the data generating distribution explicitly (Goodfellow, 2016). The standard VAE uses a probabilistic encoder-decoder architecture. The probabilistic encoder, parameterised by θ, is given by $q_\theta(\cdot) : x \in \mathcal{D} \mapsto z \in \mathbb{R}^L$, where z is the L-dimensional latent variable. The encoder approximates the true posterior distribution, $p(z|x)$, with a Gaussian density with mean $\mu_\theta \in \mathbb{R}^L$ and variance $\sigma^2_\theta \in \mathbb{R}^L$, i.e., $q_\theta(z|x) \sim \mathcal{N}(\mu_\theta, \sigma^2_\theta)$. The decoder, given by $p_\phi(x|z)$ and parameterised by ϕ, is trained to reconstruct the input x based on the latent variable z.
The VAE is optimized using the Evidence Lower Bound (ELBO) as the objective (Kingma and Welling, 2014). The ELBO is given by:

$$\mathcal{L}_{VAE} = -\mathbb{E}_{q_\theta}[\log\ p_\phi(x|z)] + D_{KL}[q_\theta(z|x)||p(z)] \tag{2}$$

where the first term is referred to as the reconstruction loss $\mathcal{L}_{rec}$, and the second term as the regularization loss $\mathcal{L}_{reg}$ given by the Kullback-Leibler Divergence (KLD) and $p(z)$ is the prior given by $p(z) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. The regularization loss forces the encoder density to match the prior during training, enabling generative sampling from the prior density at inference. To balance latent space regularization and reconstruction quality, an additional scaling factor β, suggested by Higgins et al. (2017), is introduced, giving:

$$\mathcal{L}_{VAE} = \mathcal{L}_{rec} + \beta\mathcal{L}_{reg} \tag{3}$$

VAEs often suffer from 'posterior collapse' (Bowman et al., 2016; Kingma et al., 2016) which can be alleviated when using a $\beta < 1$.

2.3. Unsupervised Discovery of Interpretable Directions in Latent Spaces

Recently, several unsupervised methods to find interpretable directions in GAN latent spaces have been proposed (Voynov and Babenko, 2020; Härkönen et al., 2020; Shen and Zhou, 2021). In Härkönen et al. (2020); Shen et al. (2020) the interpretable directions are constrained to be orthogonal, whereas this constraint is relaxed in Voynov and Babenko (2020). As interpretable directions do not necessarily have to be orthogonal, we employ the method suggested by Voynov and Babenko (2020). The proposed method is model agnostic and can be applied to any latent generative model G. The goal of this unsupervised method is to learn distinct directions from a trained latent generative model, by learning a directions matrix A and a reconstructor R to distinguish between the directions. Since the method

jointly learns a set of directions and a model that distinguishes the resulting transformations, the directions are likely to be interpretable if the model can distinguish between them with high accuracy. If the directions do not affect all images the same or are semantically not meaningful, distinguishing them would be hard. Therefore, if the accuracy of the model is high, the directions are likely meaningful.
Formally, the unsupervised direction discovery method learns two things: First, a matrix $A \in \mathbb{R}^{d \times K}$, where d is the dimensionality of the latent space of G, and K is the number of directions that will be discovered. I.e., the columns of A correspond to the directions. Second, a reconstructor R, mapping from an image pair $(G(z), G(z + A(\epsilon e_k)))$, with the shifted latent vector $z + A(\epsilon e_k)$, where e_k is an axis-aligned unit vector and ϵ is a scalar. The reconstructor predicts the direction k by determining e_k and the scalar ϵ. In other words, the reconstructor is given an image and a shifted version and tries to determine the amount, and direction, the latent vector is shifted by. The optimization objective is given by:

$$\min_{A,R} \mathbb{E}_{z,k,\epsilon}[L_{cl}(k, \hat{k}) + \gamma L_s(\epsilon, \hat{\epsilon})] \tag{4}$$

where k and ϵ are the actual direction and amount respectively, and $\hat{k}$ and $\hat{\epsilon}$ are the predictions, L_{cl} is the classification loss based on the Reconstructor Classification Accuracy (RCA), L_s is the shift loss, and γ is a regularization factor.

3. Material & Methods

3.1. Data

The Lung Image Database Consortium image collection (LIDC-IDRI) (Armato III et al., 2011) provided by The Cancer Imaging Archive (TCIA) is used. It consists of clinical thoracic CT scans of 1010 patients, collected from diagnostic and lung cancer screenings and is assembled by seven academic centers and eight medical imaging companies. We consider each axial CT slice as an individual image. This results in a dataset of $246,016$ 512×512 pixel CT images, which are resized to 128×128 pixels to limit computational demands.

3.2. Models & Training

We use a DCGAN based on the original paper (Radford et al., 2016), introducing the following slight changes. We improve training by introducing one-sided label smoothing (Salimans et al., 2016). One-sided label smoothing replaces the fixed targets 1 of the real labels with smoothed values randomly chosen from the interval $[0.9, 1]$. Additionally, we add noise to the discriminator input (Arjovsky and Bottou, 2017). We add Gaussian noise with a mean of 0 and start with a standard deviation of 0.1. During training we incrementally reduce the standard deviation and finally stop adding noise to the discriminator input at epoch 25. The encoder and the decoder of the VAE are based on a ResNet architecture, and we use $\beta = 0.01$ to improve reconstruction quality. For both generative models, we use a latent space size of $L = 32$ as it showed the best trade-off between image quality and compactness of the latent space. We refer to the provided GitHub repository for implementation details. We choose to train the GAN and the VAE for 50 epochs selecting

the best performing weights out of the last 5 by evaluating the models on test data using Fréchet Inception Distance (FID) (Heusel et al., 2017). We use binary cross-entropy as loss for the GAN and log mean squared error (Yu, 2020) as reconstruction loss for the VAE. As suggested by Radford et al. (2016) we use Adam (Kingma and Ba, 2015) with a learning rate of 0.0002 and 0.0001 to optimize the GAN and VAE, respectively. The best GAN and VAE weights achieved a FID of 33.3941 on and 93.8956 on the test data, respectively.
To find interpretable latent directions, we use two different reconstructors, based on LeNet (LeCun et al., 1998) and ResNet (He et al., 2016). We experiment with A having columns of unit length or orthonormal columns as suggested by Voynov and Babenko (2020). We set the number of directions K equal to the size of the latent space, i.e. $K = 32$, and experiment with increasing it to $K = 100$. We observe significantly faster convergence when using the ResNet reconstructor. Thus, when using $K = 32$, we train the model for $25,000$ iterations when using LeNet and $3,000$ iterations when using the ResNet reconstructor. When using $K = 100$ we train the VAE for $75,000$ and $4,000$ iterations with the LeNet and ResNet reconstructors respectively. For the GAN we observe slower convergence. Therefore, we train the GAN for $250,000$ and $10,000$ iterations with the LeNet and ResNet reconstructors, respectively. Since we cannot have $K > L$ when finding orthonormal directions, we only use A with columns of unit length for $K = 100$. We evaluate direction models using the RCA and the shift loss L_s from Equation 4. Additionally, after training the unsupervised model, we manually examine and label the resulting directions to determine the quality of the directions. Note that the training is still completely unsupervised.

4. Experiments & Results

We perform several experiments to investigate the unsupervised exploration of latent spaces of deep generative models. First, we train using orthonormal directions and directions of unit length. We also experiment with increasing the number of directions. Finally, we perform all experiments both with a DCGAN and a VAE as generative model.
All results are obtained without supervision, with the exception of the labeling of the selected directions. The RCA and L_s of the different experiments are presented in Tables 1(b) and 1(a) for the LeNet and ResNet reconstructors respectively. There are several key observations. First, the VAE always outperforms the GAN with respect to both RCA and L_s. Second, using directions of unit length achieves higher RCA than orthonormal directions and lower L_s in all but one case. We also observe higher RCA when using ResNet over LeNet as a reconstructor. In contrast, LeNet achieves a lower L_s when K is set to 100. Voynov and Babenko (2020) mention that a larger K does not harm interpretability but alleviates entanglement, and may, lead to more duplicate directions. We observe the same behaviour with $K = 100$ as opposed to $K = 32$.
We consistently observe eight key directions in our results. They can be categorized into five geometrical directions: width, height, size, y-translation, and rotation, and three anatomical directions: breast size, thickness, and z-Position. All model configurations find all eight directions with varying degrees of entanglement. In this work, we omit directions entangled to such a degree that there is no clear interpretation dominating the image transformation. Thus, all configurations find at least a subset of the aforementioned directions in a sufficiently disentangled manner. We present animations of all discovered directions in the

Table 1: Reconstructor Classification Accuracy (RCA) and L_s for all model configurations for ResNet (a) and LeNet (b) as reconstructor.

(a) RCA and L_s for ResNet reconstructor.

	Orthogonal		**Unit Length**		**100 Directions**	
	RCA	L_s	RCA	L_s	RCA	L_s
GAN	0.9236	0.2538	0.9383	0.1949	0.9522	0.1560
VAE	0.9939	0.1040	0.9947	0.1086	0.9861	0.1117

(b) RCA and L_s for LeNet reconstructor.

	Orthogonal		**Unit Length**		**100 Directions**	
	RCA	L_s	RCA	L_s	RCA	L_s
GAN	0.8559	0.3317	0.9062	0.2439	0.9305	0.1406
VAE	0.9800	0.1421	0.9895	0.1090	0.9791	0.0962

provided GitHub repository[2]. Figures 1(a) and 1(b) show all eight directions for the VAE and GAN respectively. The directions presented in 1 are obtained using LeNet as reconstructor and $K = 100$. For images of the other model configurations, we refer to Appendix A. Our results show that enforcing orthonormal directions increases entanglement. Finally, we observe that when using a LeNet reconstructor, more of the obtained directions are easily interpretable compared to using a ResNet reconstructor.

5. Discussion

Influence of K: The main effects of variations in the number of directions, K, is that lower K likely makes the reconstructor classification task easier. It is possible that this lessens the need for disentanglement. If so, when increasing K to 100, the increasing classification difficulty forces the model to disentangle the directions to make them easier to classify.
Orthonormal Directions: While constraining the directions to be orthonormal still leads to the same subset of interpretable directions being discovered, the quality of the directions suffers. This aligns with the observations of Voynov and Babenko (2020). However, their results show that some datasets benefit from orthonormal directions, leading to more interesting directions. We do not observe this on our data, and the lack of disentanglement is also clear from the lower RCA of the methods using orthonormal directions. Thus, we conclude that for our purpose, directions of unit length are preferable.
Choice of Reconstructor: When $K = 32$ both reconstructors show similar qualitative results, more entangled directions, L_s is larger and ResNet quantitatively outperforms LeNet. For $K = 100$, LeNet produces better qualitative results than ResNet. This is also evident in the quantitative results with LeNet and $K = 100$ achieving the lowest L_s. While ResNet has a higher RCA, RCA gives a measure of duplicate directions and only describes interpretability to a certain degree. Since LeNet performed best when using $K = 100$ and the increased number of directions benefited disentanglement we prefer LeNet as reconstructor.

2. https://github.com/julschoen/Latent-Space-Exploration-CT

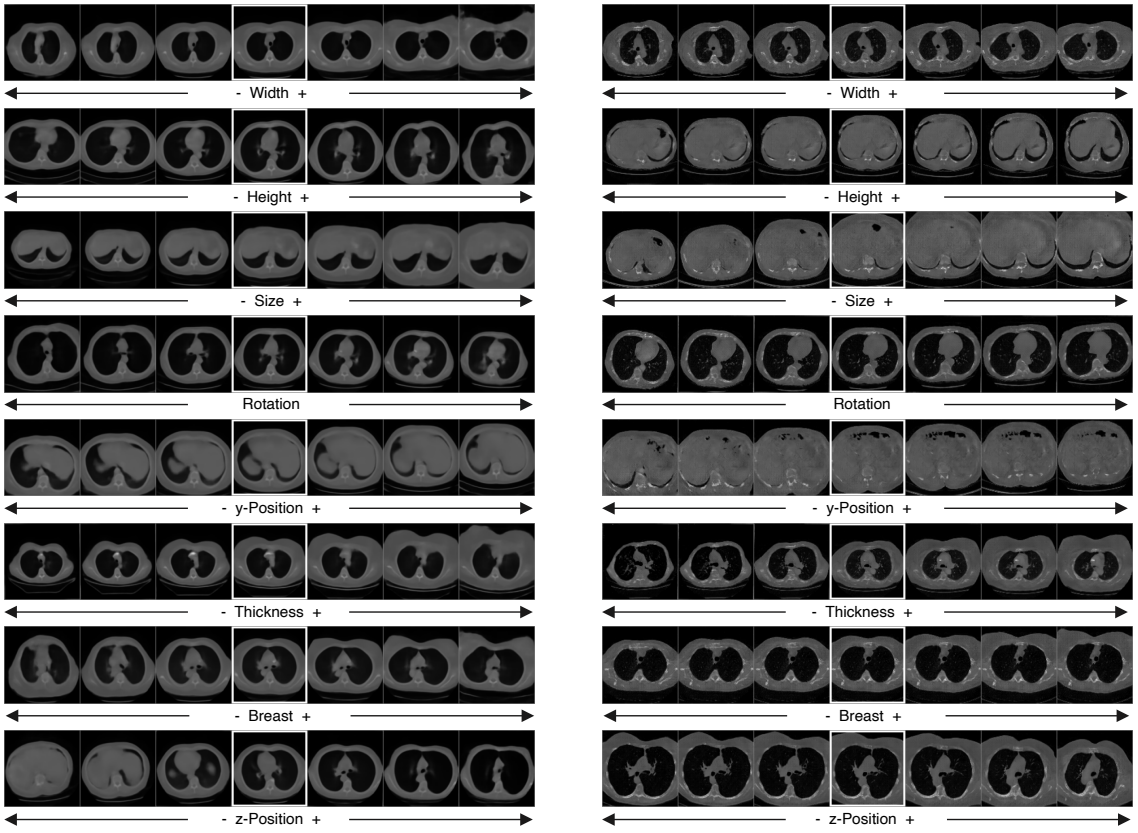


Figure 1: Example of interpretable directions using $A^{32\times100}$ with columns of unit length, LeNet as reconstructor, and the VAE (a) and GAN (b) as underlying generative models. The central images correspond to the original latent vector. The images to the left/right correspond to shifts.

Consistent Discovery of the Same Human Interpretable Directions: The same subset of interpretable directions appears in the results of all models. However, the degree of entanglement varies. Further, the directions are validated by showing that the same set of directions is discovered in the latent space of both the DCGAN and the VAE as shown in Figure 1. The resulting directions we discover show non-trivial image transformations. In particular, the directions changing the z-Position of the latent vector are interesting. It demonstrates that the models learn the 3-dimensional structure of the data despite being trained on 2-dimensional images. We provide further illustrations of the z-Position changes and a comparison of random directions with the discovered directions in Appendix B.
While the focus of discovery of directions in latent spaces has mainly been on GANs in recent years, we see that the same methods apply to VAEs. Since VAEs allow for easy mapping of real images to latent vectors, they have a practical benefit over GANs when considering the usefulness of these methods as covered in the following.

Benefits and Purpose: Improving interpretability of GANs and VAEs is an important task, which is at least partially addressed if we can interpret the learned latent representation. Finding meaningful directions in the latent space provides insights into the learned representations. We show that the method we use generalizes to VAEs, indicating that the latent spaces of VAEs and GANs can be interpreted in similar ways. However, we observe shorter convergence times on the VAE when learning interpretable directions. This indicates that the latent space of VAEs could be inherently easier to interpret. Unsupervised exploration further benefits the medical image domain due to the lack of well supervised datasets and, maybe, more importantly, it could lead to surprising results that lie outside of what we are explicitly supervising methods to find.
Aside from obvious applications in unsupervised exploration and interpretation of existing trained generative models, the discovered directions could be used for data augmentation, to perform semantic editing of medical images and as has been shown with natural images, perhaps some of these directions can be used to allow for training of unsupervised segmentation models (Voynov et al., 2021; Melas-Kyriazi et al., 2021; Voynov and Babenko, 2020).
Limitations: The main limitations we observe in our work are based on the methodology for unsupervised exploration. First, there are limitations concerning convergence. While the classification accuracy and shift loss of the reconstructor gives some insights into convergence, the implications of overfitting to these methods need to be investigated further. In particular, since the method does not rely on data, techniques such as evaluation on hold-out data are not available. This makes the decision of how many training iterations to use difficult as model performance can not be assessed on independent data. Further, the lack of evaluation metrics makes the choice of reconstructor difficult. We tried to mitigate this by using RCA and L_s for quantitative and human interpretation for qualitative analysis. But further investigation is needed to find good evaluation metrics. Second, the large amount of directions resulting from these methods makes evaluation difficult and time-consuming. This is particularly challenging in medical image analysis since evaluation of the directions may need to rely on trained evaluators such as radiologists. Automating this process or introducing a hierarchy of interpretability could be a focus of future work.

6. Conclusion

In this work, we set out to address the question of what deep generative models know when operating on medical images. We have demonstrated for the first time that recent techniques for unsupervised discovery of interpretable directions in the latent space of generative models yield good results on medical images. The methods employed can be used for dataset sizes typically encountered in this field. In addition, we show that they generalize well to two popular latent variable generative models. The results show that generative models, such as a DCGAN and VAE learn non-trivial, semantically meaningful directions when trained on CT images of the thorax. We encounter directions with the same semantic meaning regardless of the generator or direction discovery model, indicating a general structure of the latent spaces. Further, our results show that the latent spaces of the generative models capture the 3-dimensional structure of the CT scans even though they have only been trained on individual slices. The work opens up the possibility of exploring these techniques for unsupervised medical image segmentation, interpolation, augmentation, and more.

Acknowledgments

The authors acknowledge the National Cancer Institute and the Foundation for the National Institutes of Health, and their critical role in the creation of the free publicly available LIDC/IDRI Database used in this study. The authors would like to thank Anna Kirchner and Arnau Morancho Tardà for help in preparation of the manuscript. Jens Petersen is partly funded by research grants from the Danish Cancer Society (grant no. R231-A13976) and Varian Medical Systems.

References

Basel Alyafi, Oliver Diaz, and Robert Martí. Dcgans for realistic breast mass augmentation in x-ray mammography. In *Medical Imaging 2020: Computer-Aided Diagnosis*, volume 11314, page 1131420. International Society for Optics and Photonics, 2020.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, 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*, 38(2):915–931, 2011.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1002. URL https://aclanthology.org/K16-1002.

S. Nikkath Bushra and G. Shobana. A survey on deep convolutional generative adversarial neural network (dcgan) for detection of covid-19 using chest x-ray/ct-scan. In *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pages 702–708, 2020. doi: 10.1109/ICISS49785.2020.9316125.

Andres Diaz-Pinto, Adrián Colomer, Valery Naranjo, Sandra Morales, Yanwu Xu, and Alejandro F. Frangi. Retinal image synthesis and semi-supervised learning for glaucoma assessment. *IEEE Transactions on Medical Imaging*, 38(9):2211–2218, 2019. doi: 10.1109/TMI.2019.2903434.

Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2018.09.013. URL https://www.sciencedirect.com/science/article/pii/S0925231218310749.

Tomoyuki Fujioka, Mio Mori, Kazunori Kubota, Yuka Kikuchi, Leona Katsuta, Mio Adachi, Goshi Oda, Tsuyoshi Nakagawa, Yoshio Kitazume, and Ukihide Tateishi. Breast ultrasound image synthesis using deep convolutional generative adversarial networks. *Diagnostics*, 9(4):176, 2019.

Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5743–5752, 2019. doi: 10.1109/ICCV.2019.00584.

Ian Goodfellow. Neurips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9841–9850. Curran Associates, Inc., 2020. URL https://tinyurl.com/4rycanab.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a1d694707eb0fefe65871369074926d-Paper.pdf.

Yuta Hiasa, Yoshito Otake, Masaki Takao, Takumi Matsuoka, Kazuma Takashima, Aaron Carass, Jerry L. Prince, Nobuhiko Sugano, and Yoshinobu Sato. Cross-modality image synthesis from unpaired data using cyclegan. In Ali Gooya, Orcun Goksel, Ipek Oguz, and Ninon Burgos, editors, *Simulation and Synthesis in Medical Imaging*, pages 31–41, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00536-8.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Ali Jahanian, Lucy Chai, and Phillip Isola. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18b.html.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

D.P Kingma and L.J Ba. Adam: A method for stochastic optimization. In *ICLR 2015*. arXiv.org, 2015.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/ddeebdeefdb7e7e7a697e1c3e3d8ef54-Paper.pdf.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an Unsupervised Image Segmenter in Each of Your Deep Generative Models. *arXiv e-prints*, art. arXiv:2105.08127, May 2021.

Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In Maxime Descoteaux, Lena Maier-Hein, Alfred Franz, Pierre Jannin, D. Louis Collins, and Simon Duchesne,

editors, *Medical Image Computing and Computer Assisted Intervention MICCAI 2017*, pages 417–425, Cham, 2017. Springer International Publishing. ISBN 978-3-319-66179-7.

Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. In *International Conference on Machine Learning (ICLR)*, 2020. URL https://openreview.net/forum?id=H1laeJrKDB.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021.

Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9243–9252, 2020.

Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *International Conference on Machine Learning*, pages 9786–9796. PMLR, 2020.

Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10596–10606. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/voynov21a.html.

Ronald Yu. A tutorial on vaes: From bayes' rule to lossless compression, 2020.

Qianqian Zhang, Haifeng Wang, Hongya Lu, Daehan Won, and Sang Won Yoon. Medical image synthesis with generative adversarial networks for tissue recognition. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 199–207, 2018. doi: 10.1109/ICHI.2018.00030.

Jin Zhu, Guang Yang, and Pietro Lio. How can we make gan perform better in single medical image super-resolution? a lesion focused multi-scale approach. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1669–1673, 2019. doi: 10.1109/ISBI.2019.8759517.

Appendix A. Discovered Directions for all Model Configurations.

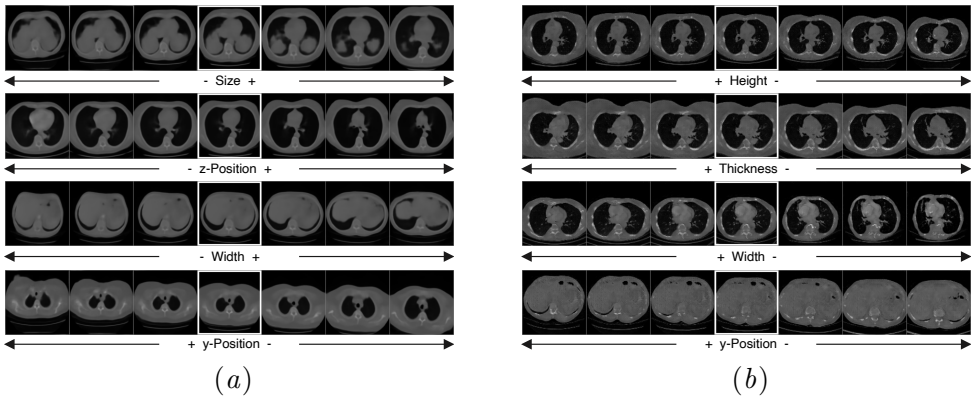


Figure 2: Example of interpretable directions using $A^{32\times32}$ with orthonormal columns, LeNet as reconstructor, the VAE (a) and GAN (b) as underlying generative models. The central images correspond to the original latent vector. The images to the left/right of that correspond to a negative/positive shift. We observed fewer disentangled directions than with other methods.

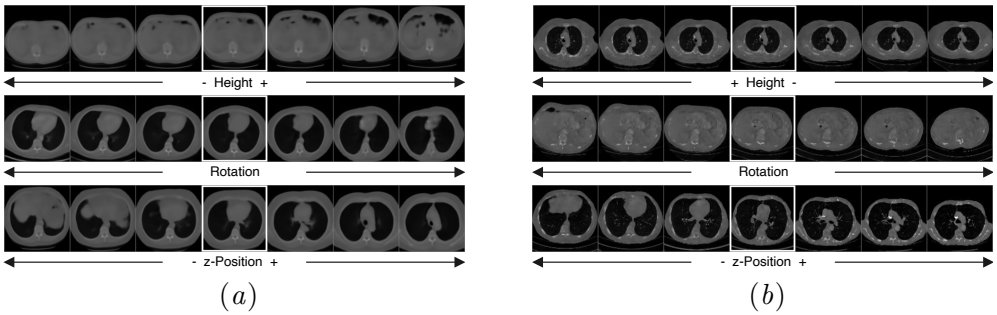
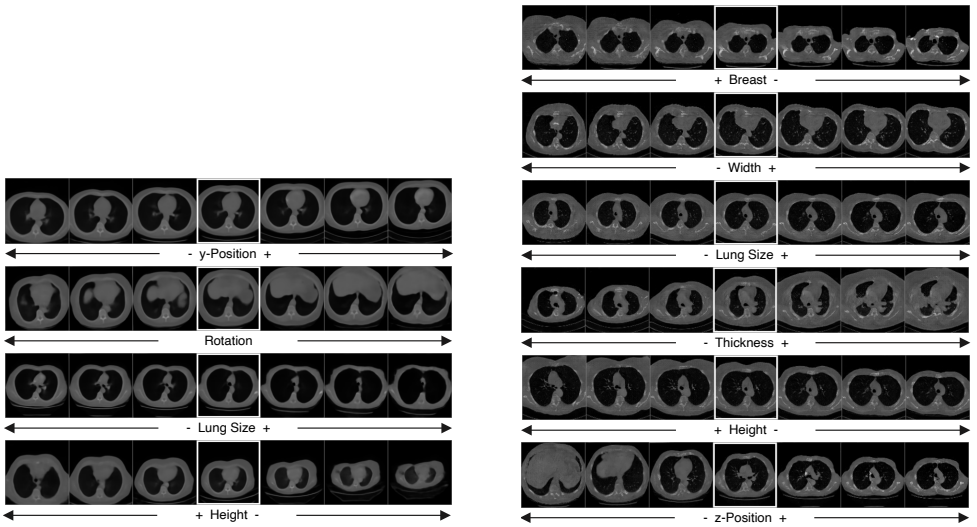
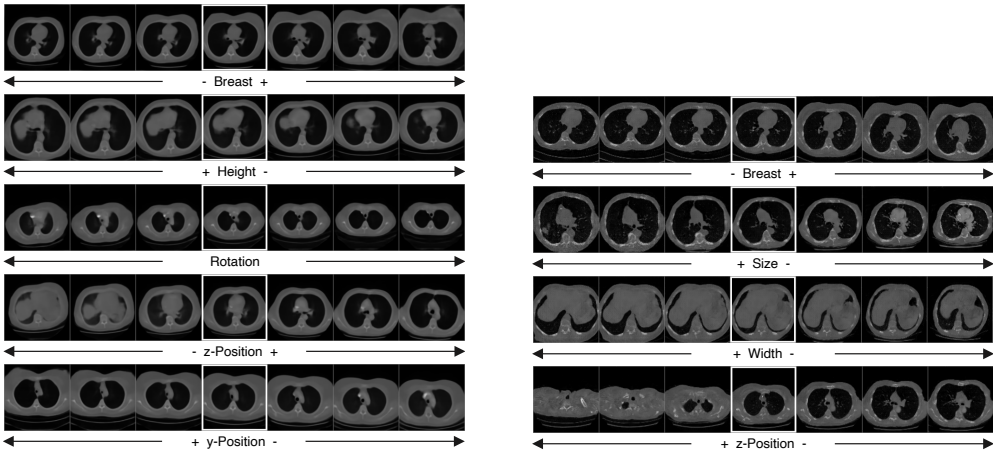


Figure 3: Example of interpretable directions using $A^{32\times32}$ with orthonormal columns, ResNet as reconstructor, the VAE (a) and GAN (b) as underlying generative models. The central images correspond to the original latent vector. The images to the left/right of that correspond to a negative/positive shift. Again, we observe far fewer disentangled directions compared to the other methods limiting the amount of directions we report.



(a) (b)

Figure 4: Example of interpretable directions using $A^{32\times32}$ with columns of unit length, LeNet as reconstructor, the VAE (a) and GAN (b) as underlying generative models. The central images correspond to the original latent vector. The images to the left/right of that correspond to a negative/positive shift.



(a) (b)

Figure 5: Example of interpretable directions using $A^{32\times32}$ with columns of unit length, ResNet as reconstructor, the VAE (a) and GAN (b) as underlying generative models. The central images correspond to the original latent vector. The images to the left/right of that correspond to a negative/positive shift.

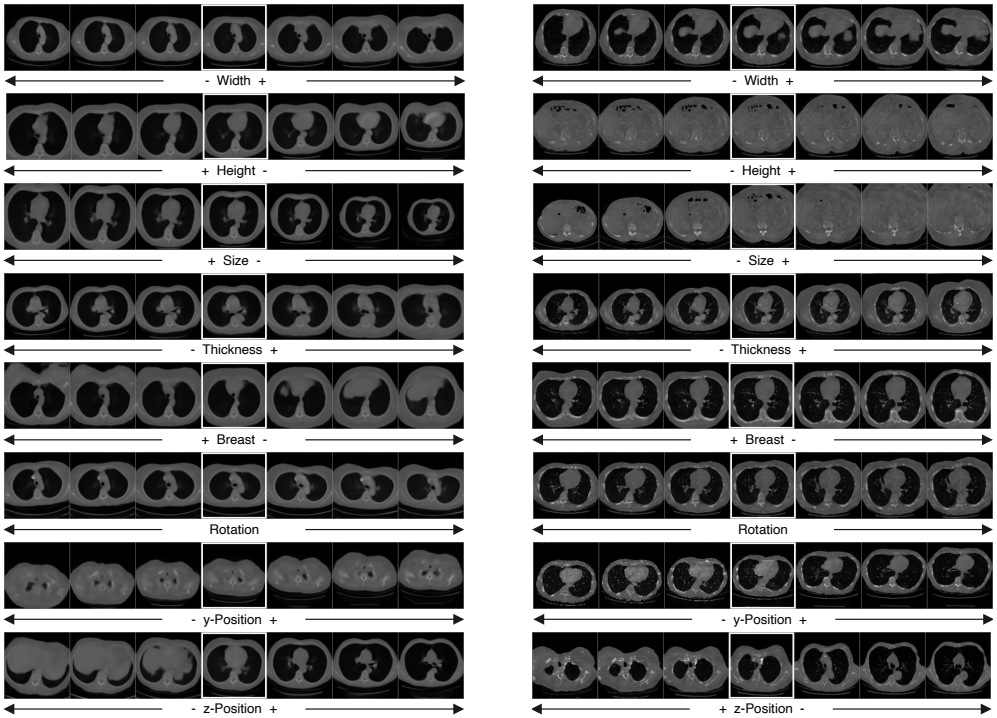


Figure 6: Example of interpretable directions using $A^{32\times100}$ with columns of unit length, ResNet as reconstructor, the VAE (a) and GAN (b) as underlying generative models. The central images correspond to the original latent vector. The images to the left/right of that correspond to a negative/positive shift.

Appendix B. Supplementary Images

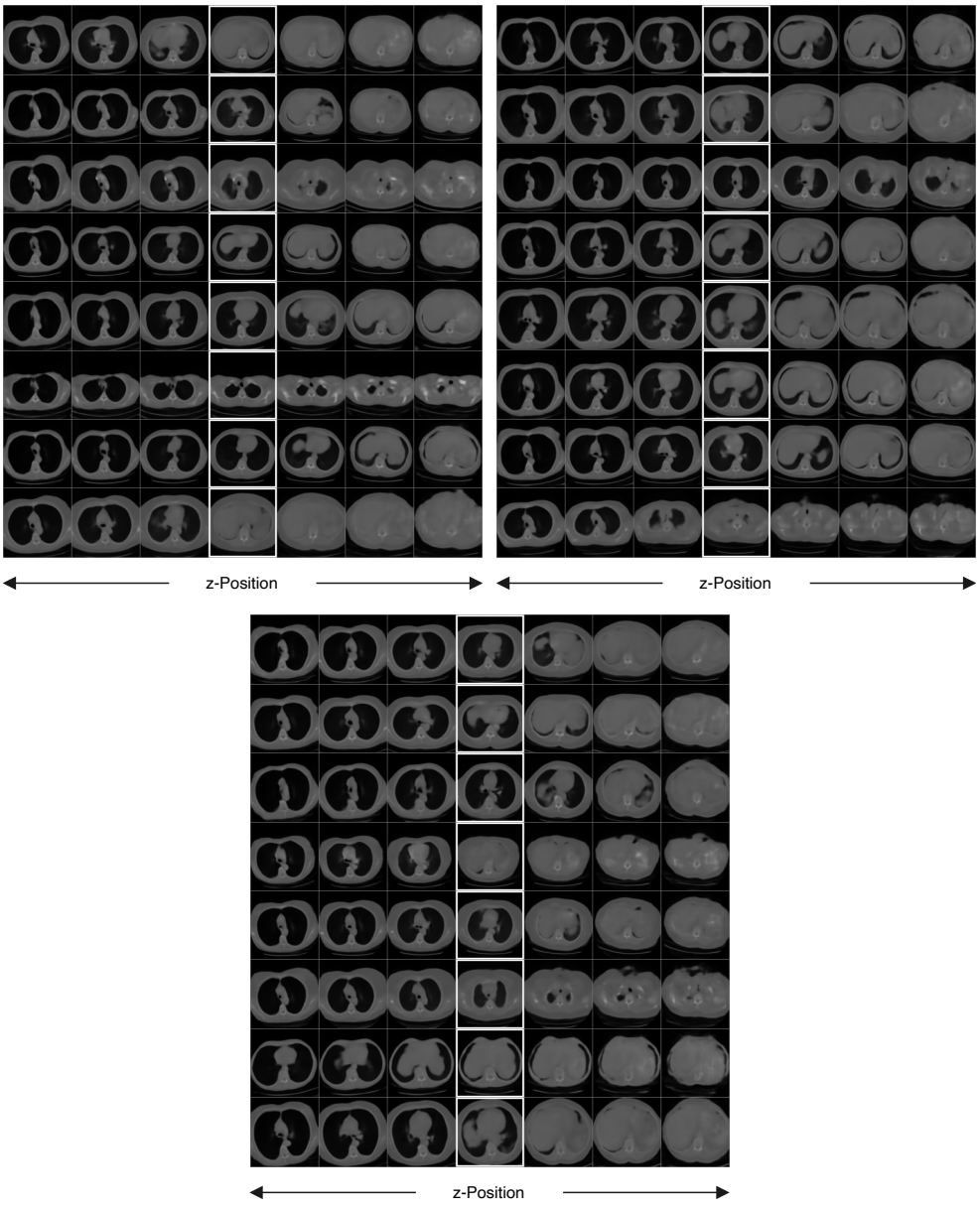


Figure 7: 24 Randomly sampled latent vectors shifted along the directions corresponding to z-Position. The central images correspond to the original latent vector. The images to the left/right of that correspond to a negative/positive shift. Each latent vector shows biological variation and all shifts show realistic changes in anatomy corresponding to different z-positions in order of the amount of shift, such as different anatomical areas of the airways, heart, lungs, and liver.

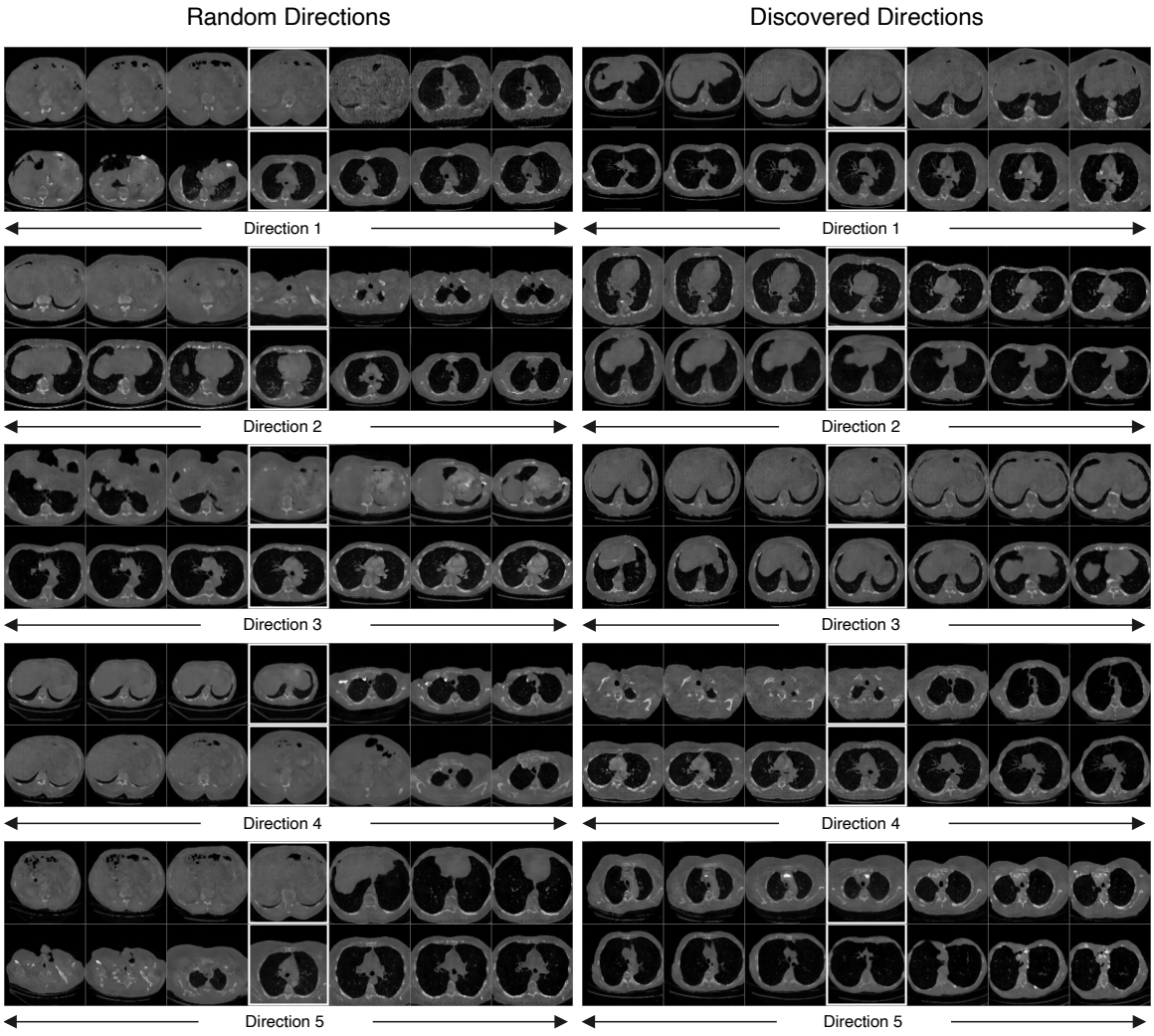


Figure 8: 5 Random directions (left) next to the first 5 discovered directions (right) using $A^{32\times100}$ with columns of unit length, LeNet as reconstructor and the GAN as underlying generative model. The central images correspond to the original latent vector. The images to the left/right of that correspond to a negative/positive shift. The results show a marked difference in interpretability of the discovered directions in contrast to random directions.

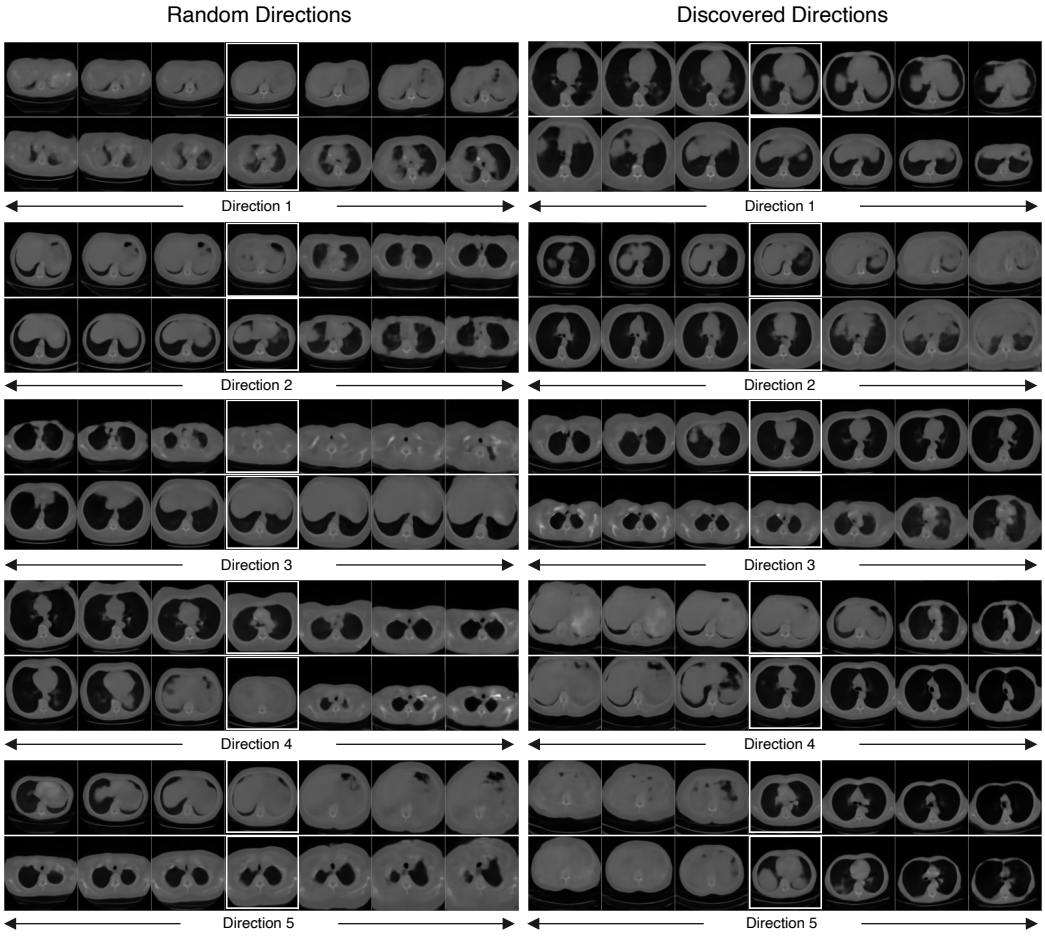


Figure 9: 5 Random directions (left) next to the first 5 discovered directions (right) using $A^{32\times100}$ with columns of unit length, LeNet as reconstructor and the VAE as underlying generative model. The central images correspond to the original latent vector. The images to the left/right of that correspond to a negative/positive shift. We can see that the random directions are continues and some of them somewhat interpretable. In particular, the random direction results are better than what we observe with the GAN. This is most likely due to the regularization of the latent space as well as the VAE learning a structured latent space which is in contrast to the GAN.