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

Julian Schön[1,2(✉)], Raghavendra Selvan[1,3], and Jens Petersen[1,2]

[1] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
julian.e.s@di.ku.dk
[2] Department of Oncology, Rigshospitalet, Copenhagen, Denmark
[3] Department of Neuroscience, University of Copenhagen, Copenhagen, Denmark

Abstract. Generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) play an increasingly important role in medical image analysis. 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. 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 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. Furthermore, the directions show that the generative models capture 3D structure despite being presented only with 2D data. The results show that unsupervised methods to discover interpretable directions in GANs generalize to VAEs and can be applied to medical images. This opens a wide array of future work using these methods in medical image analysis. The code and animations of the discovered directions are available online at https://github.com/julschoen/Latent-Space-Exploration-CT.

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. One of the main limitations is the often

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-18576-2_3.

A. Mukhopadhyay et al. (Eds.): DGM4MICCAI 2022, LNCS 13609, pp. 24–33, 2022.
https://doi.org/10.1007/978-3-031-18576-2_3

small dataset sizes available for deep learning. Generative models can be used to mitigate this by synthesizing and augmenting medical images [12].

Generative Adversarial Networks (GANs) [6] have emerged as the prominent generative model for image synthesis. Consequently, research focusing on the interpretability of GANs has unfolded. At their inception, Radford et al. [20] showed meaningful vector arithmetic in the latent space of Deep Convolutional Generative Adversarial Networks (DCGANs). For several years, the methods used for discovering interpretable directions in latent spaces have been supervised [4,11,19] or based on simple vector arithmetic [20]. Especially in medical image analysis, supervision is expensive as it typically involves radiologists or other experts' time. Recently, several unsupervised methods for discovering interpretable directions in GAN latent spaces were proposed [7,23,25]. Due to being unsupervised, they seem more promising for the medical domain. However, it is still unclear if they work with the often more homogeneous images and the smaller dataset sizes encountered in this field.

Next to GANs, the interpretability of Variational Autoencoders (VAEs) [15] has also been studied extensively. However, the investigation has mainly focused on obtaining disentangled latent space representations [10,13]. While this shows promising results, it might not be possible without introducing inductive biases [17]. Applying the approaches for the unsupervised discovery of interpretable directions in latent spaces developed for GANs to VAEs might yield an alternative route for the investigation of interpretability in VAEs. Thus, 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, therefore not incorporating inductive biases while still having the benefit of interpretability and explicit data approximation.

Contributions: We employ a technique for the unsupervised discovery of interpretable directions in the latent spaces of DCGANs and VAEs trained on Computed Tomography (CT) scans. We show that these methods used to interpret the latent spaces of GANs generalize 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 non-trivial and semantically meaningful directions are encoded in the latent space of the generative models under consideration. These directions include both transformations specific to our dataset choice and ones that likely generalize to other data. In particular, this allows for future work considering semantic editing of medical images in latent spaces of generative models.

2 Background

2.1 Generative Latent Models

As the backbone of this work we use generative latent models. We employ two of the most popular model types in GANs [6] for implicit and VAEs [15] for explicit approximation of the data distribution [5].

Given the discriminator D, the generator G, the latent distribution p_z, the data distribution p_{data}, and binary cross-entropy as the loss the GAN optimization is given by:

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

We optimize the VAE using the Evidence Lower Bound (ELBO) with additional scaling factor β [10] given by:

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

where the first term is referred to as the reconstruction loss, with p_ϕ giving the likelihood parameterised by ϕ, and the second term as the regularization loss given by the Kullback-Leibler Divergence (KLD), with q_θ giving the approximate posterior parameterised by θ and $p(z)$ is the prior given by $p(z) \sim \mathcal{N}(0, I)$.

2.2 Discovery of Interpretable Directions in Latent Spaces

Several unsupervised methods to find interpretable directions in GAN latent spaces have been proposed [7,23,25]. In Härkönen et al.; Shen et al. [7,23] the directions are orthogonal. This constraint is relaxed in Voynov and Babenko [25]. As interpretable directions do not have to be orthogonal, we employ the method suggested by Voynov and Babenko [25]. The proposed method can be applied to any pretrained latent generative model G. The objective is to learn distinct directions in the latent space of G by learning a matrix A containing directions and a reconstructor R to distinguish between them. Since A and R are learned jointly, the directions of A are likely to be interpretable, semantically meaningful, and affect all images equally. Otherwise, distinguishing between the directions would be hard, and consequently, the accuracy of R would suffer.

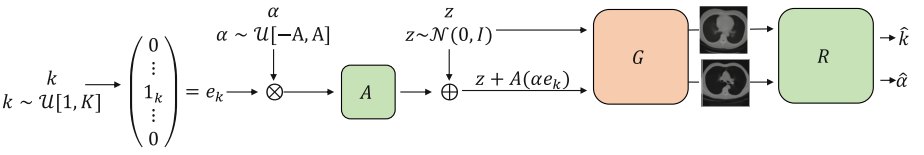

Fig. 1. Schematic overview of the learning protocol suggested by Voynov and Babenko. The upper path corresponds to the original latent code $z \sim \mathcal{N}(0, I)$ and the lower path corresponds to the shifted code $z + A(\alpha e_k)$ (Adapted from [25]).

Formally, the method learns 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. Thus, the columns of A correspond to discovered directions and are optimized during the training process to be easily distinguishable. Further, let $z \sim \mathcal{N}(0, I)$ be a latent code, e_k an axis-aligned unit vector with $k \in [1, ..., K]$

and α a scalar. Then, we can define the image pair $(G(z), G(z + A(\alpha e_k)))$ where $G(z)$ is the original image generated by latent code z and $G(z + A(\alpha e_k))$ is a shifted image from the original latent code z shifted along the kth discovered direction by amount α. Thus, α is a 'knob' controlling the magnitude of the shift. Given such an image pair, the method optimizes the reconstructor R presented with that pair to predict the shift direction k and amount α. Figure 1 illustrates the architecture. The optimization objective is given by:

$$\min_{A,R} \mathbb{E}_{z,k,\alpha}[L_{cl}(k, \hat{k}) + \gamma L_s(\alpha, \hat{\alpha})] \tag{3}$$

where k and α are the direction and amount respectively, and $\hat{k}$ and $\hat{\alpha}$ are the predictions. The classification term L_{cl} is given by cross-entropy. Further, we can use the classification term to get the Reconstructor Classification Accuracy (RCA), i.e., the accuracy of predicting the direction. Finally, the shift term L_s is given by the mean absolute error, and the regularization factor γ.

3 Material and Methods

3.1 Data

We use Lung Image Database Consortium image collection (LIDC-IDRI) [2] provided by The Cancer Imaging Archive (TCIA). 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 slice as an individual image. Thus, our dataset consists of $246,016$ CT slices. We resize the images to 128×128 pixels to limit computational demands and limited the data to a range of $[-1000, 2000]Hu$ to reduce the amount of outlier values and normalized using min-max scaling.

3.2 Models and Training

Since this study focuses on the potential of unsupervised exploration of latent spaces for medical images, we use simple generative models. We use a DCGAN based on Radford et al. [20], improving training stability by introducing one-sided label smoothing [22], replacing the fixed targets 1 of the real labels with smoothed values randomly chosen from the interval $[0.9, 1]$. Additionally, we add 0-mean and 0.1 standard deviation Gaussian noise to the discriminator input [1], incrementally reducing the standard deviation and finally removing it at the midpoint of training. The encoder and decoder of the VAE are based on ResNet [8], and we use $\beta = 0.01$ to improve reconstruction quality. For both generative models, we use a latent space size of $d = 32$ 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 train the GAN and the VAE for 50 epochs selecting the best weights out of the last 5 by considering the models Fréchet Inception Distance (FID) [9] on test data. We

use binary cross-entropy as loss for the GAN and log mean squared error [28] as reconstruction loss for the VAE. We use Adam [14] with a learning rate of 0.0002 and 0.0001 to optimize the GAN and VAE, respectively. The best model weights yield a FID of 33.4 for the GAN and 93.9 for the VAE on the test data.

To find interpretable latent directions, we use two different reconstructor architectures, based on LeNet [16] and ResNet18. We experiment with A having unit length or orthonormal columns as suggested by Voynov and Babenko [25]. 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 using LeNet and $3,000$ iterations using the ResNet reconstructor. When $K = 100$, 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. Thus, we train for $250,000$ and $10,000$ iterations with the LeNet and ResNet reconstructors, respectively. Since we cannot have $K > d$ for 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 Eq. 3. Further, we follow the ablation provided by Voynov and Babenko [25] and use a regularization factor $\gamma = 0.25$. To evaluate the directions, preliminary labeling was done by the first author with eight animations, each showing different latent vectors per direction. Next, each direction and preliminary label was considered on eight static images. The evaluator does not have formal training in medical image interpretation, and it is possible that more experienced evaluators could have discovered more interesting directions.

4 Experiments and 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 models. All results are obtained without supervision, except the labeling of the selected directions. The RCA and Ls of the different experiments are presented in Table 1. We observe that the VAE always outperforms

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

	Orthogonal		Unit length		100 directions	
	RCA	L_s	RCA	L_s	RCA	L_s
GAN ResNet	0.9236	0.2538	0.9383	0.1949	0.9522	0.1560
GAN LeNet	0.8559	0.3317	0.9062	0.2439	0.9305	0.1406
VAE ResNet	0.9939	0.1040	0.9947	0.1086	0.9861	0.1117
VAE LeNet	0.9800	0.1421	0.9895	0.1090	0.9791	0.0962

the GAN with respect to both RCA and L_s. Further, 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 [25] mention that a larger K does not harm interpretability but alleviates entanglement and may lead to more duplicate directions. We observe the same behavior with $K = 100$ as opposed to $K = 32$.

Our results show eight consistent directions: width, height, size, rotation, y-position, thickness, breast size, 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 directions above in a sufficiently disentangled manner. We present animations of all discovered directions in the provided GitHub repository. Figure 2 shows all eight directions for the VAE and GAN. The directions presented are obtained using LeNet as reconstructor and $K = 100$. Directions obtained using different model configurations are presented in the supplementary material. Our results show that enforcing orthonormal directions increases entanglement. Finally, we

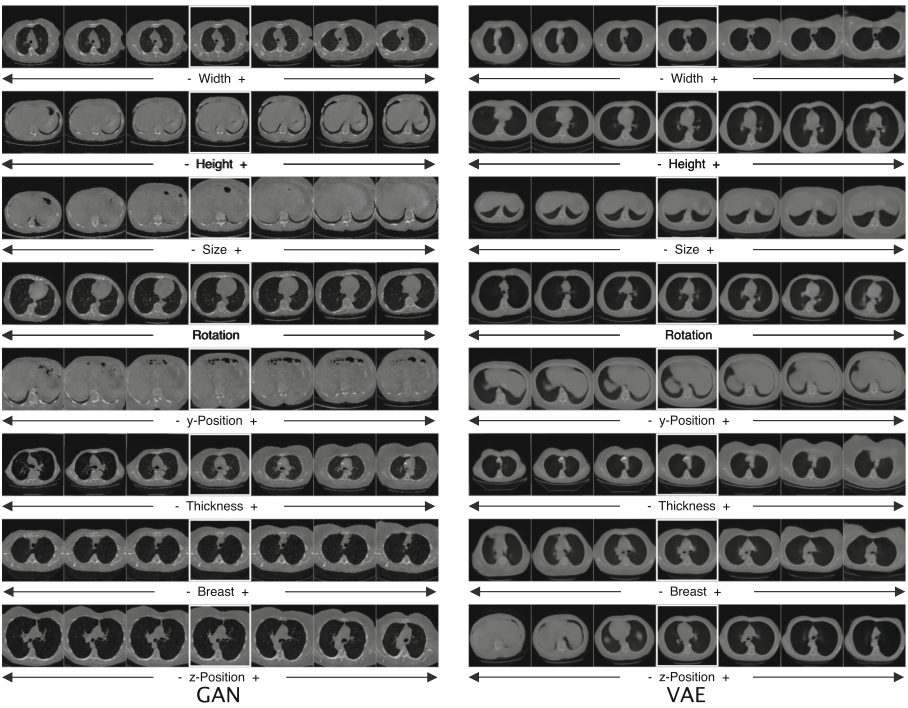

Fig. 2. Interpretable directions using $A^{32\times100}$ with unit length columns, LeNet as reconstructor, and the GAN and VAE as generative models. The central images correspond to the original latent vector. The left/right images correspond to shifts.

observe that when using a LeNet reconstructor, more of the obtained directions are easily interpretable compared to using a ResNet reconstructor.

5 Discussion

In this work, we explored the latent spaces of deep generative models to discover semantically meaningful directions. We next elaborate on some of the findings of our experiments.

Influence of K: We observe less entanglement when increasing K. Thus, we hypothesize that lower K likely makes the reconstructor classification task easier, as there are less classes, lessening the need for disentanglement. If so, when increasing K, the increasing classification difficulty forces the model to disentangle the directions more.

Orthonormal Directions: While constraining the directions to be orthonormal still leads to the same subset of interpretable directions being discovered, their quality suffers. This aligns with the observations of Voynov and Babenko [25]. 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, it seems likely that directions offering semantic meaning are not necessarily orthonormal, strengthening our reasoning for choosing this method over Härkönen et al.; Shen et al. [7,23].

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 partially describes interpretability. Since LeNet performed best when using $K = 100$ and the increased number of directions benefited disentanglement, we prefer LeNet as reconstructor.

Consistent Discovery of Interpretable Directions: The same subset of human interpretable directions appears for all models with varying degrees of entanglement. Recent work has shown non-linear directions to be less entangled [24] which could be studied further. The directions are validated by showing that the same set is discovered in the latent space of both the generative models. The resulting directions we discover show non-trivial image transformations. In particular, the directions changing the z-Position of the latent vector demonstrates that the models learn the 3D structure of the data despite being trained on 2D images. While the focus of discovering 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 explicit data approximation, they have a practical benefit over GANs when considering the usefulness of these methods.

Impact and Applications: Improving interpretability of GANs and VAEs is important and addressed in this work by finding and visualizing meaningful latent space directions and providing novel insights into the learned representations. The method is shown to generalize to VAEs, indicating that the latent spaces of VAEs and GANs can be interpreted in similar ways. However, shorter convergence times on the VAE when learning the directions indicate that VAEs latent spaces could be inherently easier to interpret. Unsupervised exploration further benefits the medical image domain due to the lack of well-supervised datasets, and more importantly, it could lead to surprising results outside of what we are explicitly supervising methods to find.

Our work can further be used for context-aware image augmentation and editing. Image augmentation using synthetic data improves downstream machine learning tasks on medical images [3] and can alleviate both the small dataset sizes and imbalance inherent to medical imaging [12,27]. Our results could be used to explore more diverse augmentations, e.g., adjusting for sex and weight imbalances. Additionally, our work might offer an alternate unsupervised approach to disease-aware image editing [21].

We see further applications needing more investigation, such as exploring the potential in consistency regularization and multi-modal datasets. For example, finding directions corresponding to adding or removing contrast in scans. Further, the approach we use has been shown to be effective in unsupervised saliency detection and segmentation on natural images [18,25,26].

Limitations: The main limitations we observe in our work are based on the methodology for unsupervised exploration. First, while the RCA and shift loss give some insights into convergence, the implications of overfitting need to be investigated. In particular, deciding how many training iterations to use is 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. Nevertheless, further investigation is needed to find good evaluation metrics. Second, the large amount of resulting directions makes evaluation difficult and time-consuming. This is particularly challenging in medical image analysis as evaluation may involve trained evaluators such as radiologists. Further automation or introducing a hierarchy of interpretability could be a focus of future work. Next to the methodological limitations, we see further potential for expanding our work to 3D generative models and more datasets in the future.

6 Conclusion

In this work, we have demonstrated for the first time that techniques for unsupervised discovery of interpretable directions in the latent space of generative models yield good results on medical images. While the interpretability of latent spaces is arguably an abstract concept depending on those interpreting, our results show that generative models 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 generative models' latent spaces capture the 3D structure of the CT scans despite only being trained on 2D slices. The work opens up the possibility of exploring these techniques for unsupervised medical image segmentation, interpolation, augmentation, and more.

Acknowledgements. 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

1. Arjovsky, M., Bottou, L.: Towards principled methods for training generative adversarial networks. In: 5th International Conference on Learning Representations (2017)
2. Armato III, S.G., et al.: The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Med. Phys. **38**(2), 915–931 (2011)
3. Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., Haworth, A.: A review of medical image data augmentation techniques for deep learning applications. J. Med. Imaging Radiat. Oncol. **65**(5), 545–563 (2021)
4. Goetschalckx, L., Andonian, A., Oliva, A., Isola, P.: GANalyze: toward visual definitions of cognitive image properties. In: IEEE/CVF International Conference on Computer Vision, pp. 5743–5752 (2019)
5. Goodfellow, I.J.: NIPS 2016 tutorial: generative adversarial networks. arXiv (2016)
6. Goodfellow, I.J., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, pp. 2672–2680 (2014)
7. Härkönen, E., Hertzmann, A., Lehtinen, J., Paris, S.: GANspace: discovering interpretable GAN controls. In: Advances in Neural Information Processing Systems, vol. 33, pp. 9841–9850. Curran Associates, Inc. (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE Computer Society (2016)
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
10. Higgins, I., et al.: beta-VAE: learning basic visual concepts with a constrained variational framework. In: 5th International Conference on Learning Representations (2017)
11. Jahanian, A., Chai, L., Isola, P.: On the "steerability" of generative adversarial networks. In: 8th International Conference on Learning Representations (2020)

12. Kazeminia, S., et al.: GANs for medical image analysis. Artif. Intell. Med. **109**, 101938 (2020)
13. Kim, H., Mnih, A.: Disentangling by factorising. In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2649–2658 (2018)
14. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations (2015)
15. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: 2nd International Conference on Learning Representations (2014)
16. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
17. Locatello, F., et al.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 4114–4124 (2019)
18. Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A.: Finding an unsupervised image segmenter in each of your deep generative models. arXiv (2021)
19. Plumerault, A., Le Borgne, H., Hudelot, C.: Controlling generative models with continuous factors of variations. In: International Conference on Machine Learning (2020)
20. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: 4th International Conference on Learning Representations (2016)
21. Saboo, A., Ramachandran, S.N., Dierkes, K., Keles, H.Y.: Towards disease-aware image editing of chest X-rays. arXiv (2021)
22. Salimans, T., et al.: Improved techniques for training GANs. In: Advances in Neural Information Processing Systems, vol. 29. Curran Associates, Inc. (2016)
23. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in GANs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1532–1540 (2021)
24. Tzelepis, C., Tzimiropoulos, G., Patras, I.: WarpedGANSpace: finding non-linear RBF paths in GAN latent space. In: IEEE/CVF International Conference on Computer Vision, pp. 6393–6402 (2021)
25. Voynov, A., Babenko, A.: Unsupervised discovery of interpretable directions in the GAN latent space. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 9786–9796 (2020)
26. Voynov, A., Morozov, S., Babenko, A.: Object segmentation without labels with large-scale generative models. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 10596–10606 (2021)
27. Yi, X., Walia, E., Babyn, P.S.: Generative adversarial network in medical imaging: a review. Med. Image Anal. **58**, 101552 (2019)
28. Yu, R.: A tutorial on VAEs: From bayes' rule to lossless compression. arXiv (2020)