







This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Interpretable Representation Learning of Cardiac MRI via Attribute Regularization

Maxime Di Folco^{1,2}, Cosmin I. Bercea^{1,2,3}, Emily Chan^{1,2}, and Julia A. Schnabel^{1,2,3,4}

¹ Institute of Machine Learning in Biomedical Imaging, Helmholtz Munich, Germany

² School of Computation, Information and Technology,
Technical University of Munich, Germany

³ Helmholtz AI, Helmholtz Munich, Germany

⁴ School of Biomedical Engineering & Imaging Sciences, King's College London, UK
maxime.difolco@helmholtz-munich.de

Abstract. Interpretability is essential in medical imaging to ensure that clinicians can comprehend and trust artificial intelligence models. Several approaches have been recently considered to encode attributes in the latent space to enhance its interpretability. Notably, attribute regularization aims to encode a set of attributes along the dimensions of a latent representation. However, this approach is based on Variational AutoEncoder and suffers from blurry reconstruction. In this paper, we propose an Attributed-regularized Soft Introspective Variational Autoencoder that combines attribute regularization of the latent space within the framework of an adversarially trained variational autoencoder. We demonstrate on short-axis cardiac Magnetic Resonance images of the UK Biobank the ability of the proposed method to address blurry reconstruction issues of variational autoencoder methods while preserving the latent space interpretability.

Keywords: Cardiac Imaging · Interpretability

1 Introduction

Interpretability is crucial for transparent AI systems in medical imaging to build clinician trust and advance AI adoption in clinical workflows. As highlighted by Rudin [19], it is essential that models are inherently interpretable and not black-box models to ensure their relevance. Latent representation models like Variational AutoEncoder (VAE) have emerged as potent models capable of encoding crucial hidden variables within the input data [8,2,12]. Especially when dealing with data that contain different interpretable features (data attributes), self-supervised [7] or supervised approaches can encode those attributes in the latent space [13,11,5,14].

In this context, Pati et al. [14] introduced an attribute-regularized method based on VAEs that aims to regularize each attribute, added as extra input, along a dimension of the latent space and, therefore, ensure the latent space

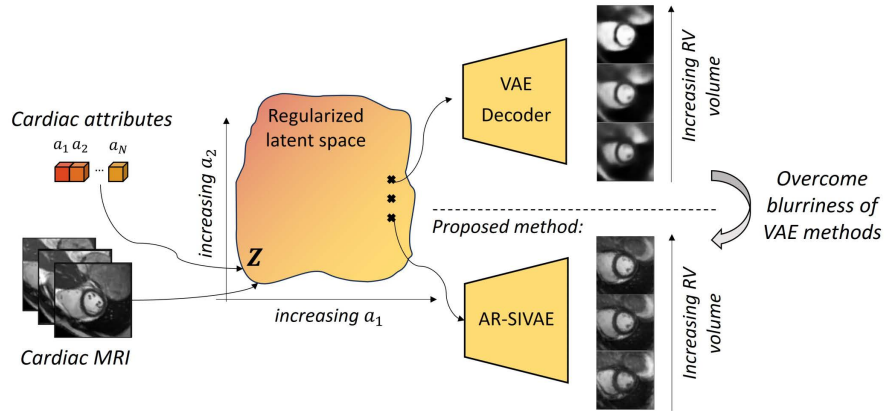


Fig. 1: **AR-SIVAE**: Attribute Regularized Soft Introspective Variational AutoEncoder (AR-SIVAE) combines attribute regularization within the SIVAE framework (an adversarially trained encoder-decoder network) to enhance the interpretability of the latent space while being able to generate non-blurry samples.

interpretability. Notably, Cetin et al. [3] applied this architecture for cardiac attributes on MRI data, demonstrating a significant improvement in the interpretability of the latent representation and its relevance for a downstream cardiac disease classification task. Nevertheless, VAE-based methods may suffer from blurry reconstruction, which could be problematic for any downstream task.

The advancement of VAE generation capabilities to overcome this limitation can be categorised into approaches that focus on enhancing the network’s architecture [20,21], integrating more robust priors [17,10], introducing regularisation techniques [6,22] or integrating adversarial objectives [9,4]. The latter has the benefits of combining the generative capability of GANs and the inference capability of VAEs, which is needed to add attribute regularisation. Notably, Daniel et al. [4] introduced a novel approach called Soft Introspective Variational AutoEncoder (SIVAE) that incorporates an adversarial loss into VAE training. In contrast to earlier methods that used additional discriminator networks [16], SIVAE utilizes the encoder and decoder of VAE in an adversarial manner and demonstrated good generation quality and strong inference capabilities on several vision datasets.

In this paper, we propose the Attributed Regularized Soft Introspective Variational Autoencoder (AR-SIVAE) by combining an attribute regularization loss in the SIVAE framework to preserve the interpretability of the latent space while having better image generation capabilities. To the best of our knowledge, we are the first to introduce this loss in an adversarially trained VAE. We compare our method to the one described in Cetin et al. [3] on a healthy population of cardiac

MRI from the UK Biobank. Our method overcomes the limitations associated with blurry reconstruction while maintaining latent space interpretability.

2 Methods

2.1 Preliminaries:

Attributed regularized VAE : Attr-VAE, proposed by Pati et al. [14], aims to encode an attribute a along a dimension k of a \mathbb{D} -dimensional latent space $\mathbf{z} : z^k, k \in [0, \mathbb{D})$, such that the attribute value increases when we traverse the dimension k . This is achieved by adding a loss to the VAE training objective. This attribute regularization loss based on an attribute distance matrix D_a and a similar distance matrix D_k computed from the regularized dimension k . They are defined as follows and computed for each batch of the training data:

$$D_a(i, j) = a(x_i) - a(x_j); D_k(i, j) = z_i^k - z_j^k \quad (1)$$

where $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^N$ are two high-dimensional samples of dimension N (with $N \gg \mathbb{D}$). The attribute regularization loss term is then computed for each attribute k as follows:

$$L_{k,a} = MAE(\tanh(\delta D_k) - \text{sgn}(D_a)) \quad (2)$$

where $MAE(\cdot)$ is the mean absolute error, $\tanh(\cdot)$ is the hyperbole tangent function, $\text{sgn}(\cdot)$ is the sign function and δ is a tunable hyperparameter which decide the spread of the posterior distribution. The sum for each attribute is then added to the β -VAE loss term and weighted by the hyperparameter γ_{reg} :

$$\mathcal{L} = \mathcal{L}_r(x) + \beta \mathcal{L}_{KL}(x) + \gamma_{reg} \mathcal{L}_{attr} \quad \text{with} \quad \mathcal{L}_{attr} = \sum_{l=0}^{\mathbb{D}-1} L_{k_l, a_l} \quad (3)$$

where $\mathcal{A} : a_l, l \in [0, D)$ a set of attributes, \mathcal{L}_r the reconstruction loss and \mathcal{L}_{KL} the Kullback Leibler divergence controlled by the parameter β .

Soft Introspective Variational Autoencoder (SIVAE) The SIVAE framework proposed by [4] is an adversarially trained VAE. Its encoder is trained to distinguish between real and generated samples by minimizing the KL divergence between the latent distribution of real samples and the prior while maximizing the KL divergence of generated samples. Conversely, the decoder aims to deceive the encoder by reconstructing real data samples using the standard Evidence Lower Bound (ELBO) and minimizing the KL divergence of generated samples embedded by the encoder. The optimization objectives for the encoder, E_ϕ , and decoder, D_θ , to be maximized are formulated as follows:

$$\begin{aligned} \mathcal{L}_{E_\phi}(x, z) &= ELBO(x) - \frac{1}{\alpha} (\exp(\alpha ELBO(D_\theta(\mathbf{z}))), \\ \mathcal{L}_{D_\theta}(x, z) &= ELBO(x) + \eta ELBO(D_\theta(z)) \end{aligned} \quad (4)$$

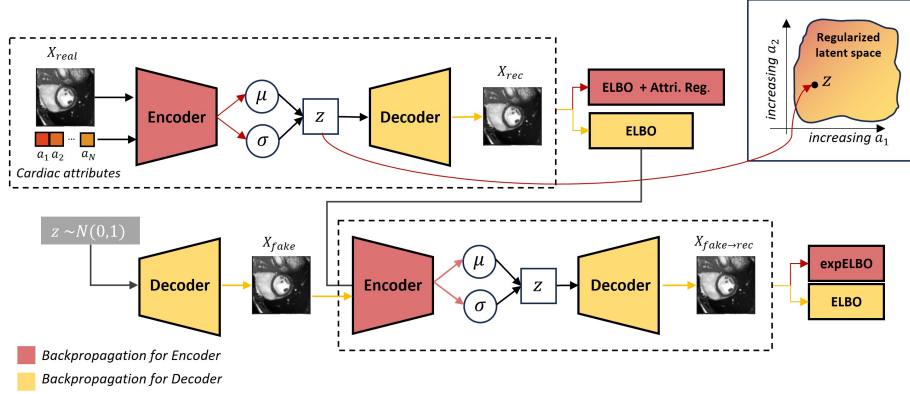


Fig. 2: Illustration of the AR-SIVAE framework. The encoder aims to distinguish, through the ELBO value, between real and generated samples and regularize the cardiac attributes in the latent space, while the decoder is induced to generate samples to fool the encoder. Attr. Reg: Attribute regularization.

where $\alpha \geq 0$ and $\eta \geq 0$ are hyper-parameters.

2.2 Proposed method: AR-SIVAE

In this work, we propose an Attribute Regularized Soft Introspective Variational Autoencoder (AR-SIVAE) (Fig. 2) by adding the attribute regularization loss defined in Eq. 2 to the encoder loss of the VAE. The optimization objective of the encoder becomes:

$$\mathcal{L}_{E_\phi}(x, z) = \gamma_{reg} \mathcal{L}_{attr} + ELBO(x) - \frac{1}{\alpha} (\exp(\alpha ELBO(D_\theta(z))) \quad (5)$$

where γ_{reg} is a hyperparameter that weights the attribute regularization loss term. Accordingly to the guidelines described in [4], we always set $\alpha = 2$. In practice, the optimization objectives are computed as follows:

$$\begin{aligned} \mathcal{L}_{E_\phi}(x, z) &= \gamma_{reg} \mathcal{L}_{attr} + s \cdot (\beta_{rec} \mathcal{L}_r(x) + \beta_{kl} \mathcal{L}_{KL}(x)) \quad (6) \\ &\quad + \frac{1}{2} \exp(-2s \cdot (\beta_{rec} \mathcal{L}_r(D_\theta(z)) + \beta_{neg} \mathcal{L}_{KL}(D_\theta(z)))) \end{aligned}$$

$$\mathcal{L}_{D_\theta}(x, z) = s \cdot \beta_{rec} \mathcal{L}_r(x) + s \cdot (\eta \cdot \beta_{rec} \mathcal{L}_r(D_\theta(z)) + \beta_{kl} \mathcal{L}_{KL}(D_\theta(z))) \quad (7)$$

where s is a normalizing constant set to the size of the images. \mathcal{L}_r is the reconstruction loss computed as a combination of the Mean Squared Error (MSE) and perceptual loss weighted by an hyperparameter α_{pl} . The training process, detailed in [4], is composed of two steps: first, the decoder is frozen, and the encoder is updated, and then the encoder is frozen, and the decoder is updated.

2.3 Dataset and Implementation details

Dataset: In this work, we processed the Cine MRI acquisitions of the short-axis view of the UK Biobank study [15]. We selected all of the 5392 subjects who self-declared to not have any cardiovascular disease using the UK Biobank field 20002. We preserve only the 5360 cases where the segmentation mask contains at end-diastole (ED) and end-systole (ES) the following: more than 10 pixels per region (left ventricle (LV), right ventricle (RV) and myocardium), more than six slices segmented with no discontinuity (no missing segmentation between the slices) and the mid-cavity slice has the LV and the RV segmented (similar to [1] and detailed in the associated code). Per subject, we selected the basal slice at ED and ES. The barycenter of the left ventricle was centred for each image and was aligned with the right ventricular barycenter along the horizontal axis. The images were cropped around the centre at a size of 128x128 pixels. The dataset was split into 3752, 804 and 804 subjects for training, validation and testing, respectively.

Attributes: We computed cardiac morphometric attributes for the regularization based on the methods described in [1] and using the associated public code ⁵. We used the volume at ED and ES of the LV, RV and the myocardium denoted LVEDV, RVEDV, MyoEDV, LVESV, RVESV and MyoESV respectively.

Implementation details: For the β -VAE-based methods, we followed the public implementation associated ⁶. We trained for up to 1000 epochs (with a patience of 100 epochs) and used ADAM optimizer with a learning rate of $5e-5$. For the SIVAE-based methods, we followed the public implementation⁷ associated with the publication [4]. We trained for up to 750 epochs (with a patience of 100 epochs), and we used two ADAM optimizers (one for the encoder and the decoder) with a learning rate of $2e-4$. For all methods, the size of the latent space was fixed to 128 dimensions and we trained with a batch size of 128. The hyperparameters were chosen empirically and detailed in Appendix A. An ablation study of the influence of the weight given to the attribute regularization is also conducted. More details on the implementation are available on the code repository: <https://github.com/compai-lab/2024-miccai-di-folco>.

3 Experiments and results

We compared the proposed method AR-SIVAE to evaluate the reconstruction performance and the interpretability of the learned representation against β -VAE, SIVAE and Attri-VAE. The β -VAE and SIVAE methods do not include

⁵ https://github.com/baiwenjia/ukbb_cardiac/

⁶ <https://github.com/1Konny/Beta-VAE/>

⁷ <https://taldatech.github.io/soft-intro-vae-web/>

Table 1: Evaluation in terms of reconstruction performance of the compared methods with or without attributed regularization (Reg. column) when trained using only ES, only ED or both at the same time (ED/ ES column).

	Reg.	SSIM \uparrow			LPIPS \downarrow		
		All (ED/ES)	ED	ES	All (ED/ES)	ED	ES
β -VAE	\times	0.51 / 0.44	0.54	0.48	0.36/ 0.41	0.34	0.38
Attri-VAE	\checkmark	0.51 / 0.44	0.54	0.48	0.37/ 0.43	0.34	0.38
SIVAE	\times	0.46/ 0.40	0.45	0.36	0.17 / 0.18	0.19	0.21
AR-SIVAE	\checkmark	0.47/ 0.40	0.46	0.40	0.17 / 0.20	0.17	0.18

attribute regularization, so will be used as baselines for the reconstruction performance. Attri-VAE [14] is considered as the reference for attribute regularization and interpretability of the latent space.

3.1 Reconstruction performance

We first assessed the reconstruction performance of the compared methods. We experimented using as input only the ES or ED images, or both at the same time as two different channels. Figure 3 illustrates the capacity of the SIVAE-based methods to overcome the blurry reconstruction of VAE-based methods for two samples (ED and ES are considered together and displayed for each sample). We also report in Table. 1 the Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) metrics (the latter is a measure of the perceptual similarity between two images) when considering ED and ES together (denoted as All) and individually. Both of the VAE- and SIVAE-based methods achieved similar performance with and without attribute regularization, suggesting that the addition of the regularization term has minimal influence on reconstruction quality. Despite obtaining blurry reconstructions (illustrated in Fig. 3), the VAE-based methods achieved slightly higher SSIM scores. While it is a widely used metric to assess the similarity between two images, the SSIM often fails to detect nuances of human perception [23]. We employ the LPIPS metric to address this limitation, revealing a significant improvement for the SIVAE-based methods. Furthermore, as expected, the reconstruction of shape variability is more challenging for ES than for ED, resulting in lower performance globally for ES. Finally, in contrast to the VAE-based methods, reconstructing ED and ES together has very little impact on the metrics compared to reconstructing individually them when using the SIVAE-based methods.

3.2 Interpretability of the latent space

In this section, we evaluate the interpretability of the latent space when reconstructing ES and ED together. Table. 1 shows the *Interpretability* score, which

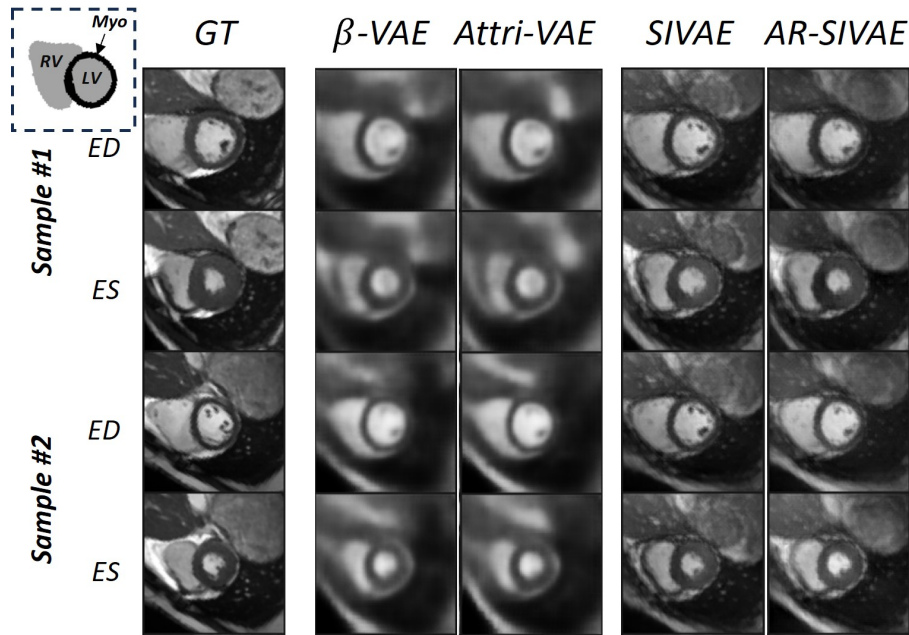


Fig. 3: Qualitative evaluation of the reconstruction of two samples at ES and ED for β -VAE, SIVAE, Attri-VAE and the proposed method AR-SIVAE. The first column corresponds to the ground truth (GT). The SIVAE-based methods (4th and 5th columns) overcome the blurry reconstruction of VAE-based methods (2nd and 3rd columns) especially for the cardiac regions of interest (illustrated on the top left): Left Ventricle (LV), Right Ventricle (RV) and Myocardium (Myo).

measures the ability to predict a given attribute using only one dimension of the latent space; the Separated Attribute Predictability (SAP), which calculates the difference in *Interpretability* score between the two most predictive dimensions; the *Modularity* metric [18], which quantifies whether each dimension of the latent space depends on only one attribute; and the Spearman Correlation Coefficient (SCC) which is the maximum value of the Spearman’s correlation coefficient between an attribute and each dimension of the latent space. We observed that adding the regularization to VAE- and SIVAE-based methods improved the performance of SCC (improvement of 0.15 for VAE and 0.14 for SIVAE) and *Interpretability* score (improvement of 0.4 VAE and 0.25 for SIVAE). Compared to Attri-VAE, AR-SIVAE achieves similarly for the SCC, the *Modularity* and *Interpretability* scores (Appendix B reported all of the *Interpretability* scores per attribute). By regularizing correlated attributes such as the cardiac volumes at ES and ED, several regularized dimensions are able to predict another attribute and this results in a lower SAP score.

Figure 4 illustrates a walk in the latent dimensions showing the evolution of the regularized attribute. The proposed method is able to generate non-blurry

samples contrary to Attri-VAE, while still displaying the variation of the considered attributes.

Table 2: Assessment of the interpretability of the latent space using the Spearman’s Correlation Coefficient (SCC), the *Modularity* score (Mod.), and the mean *Interpretability* score (Interp.). For the latter, results are shown for all attributes, as well as for the EDV- and ESV-specific ones. All the metrics are between 0 and 1, with 1 being the best performance.

	Reg.	SCC.	Mod.	SAP	Interp.
					All (EDV/ ESV)
β -VAE	✗	0.67	0.80	0.30	0.48 (0.51/ 0.48)
Attri-VAE	✓	0.96	0.85	0.09	0.88 (0.89/ 0.87)
SIVAE	✗	0.78	0.86	0.40	0.60 (0.62/ 0.58)
AR-SIVAE	✓	0.92	0.86	0.06	0.85 (0.88/ 0.82)

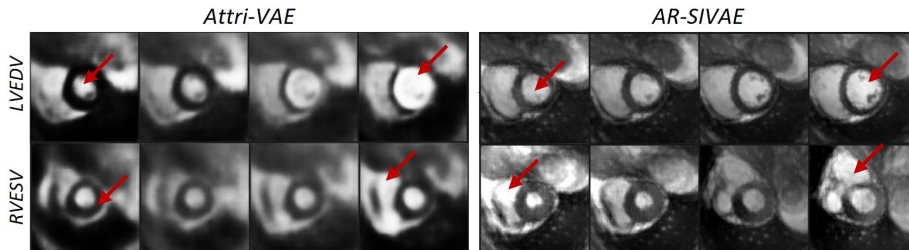


Fig. 4: Walk in the regularized latent dimensions of LV end-diastolic volume (first row) and RV end-systolic volume (second row) for Attri-VAE (left) and AR-SIVAE (right). The red arrows indicate the region of interest for each considered attribute.

4 Conclusion

This paper introduces the Attribute Regularized Soft Introspective Variational Autoencoder (AR-SIVAE), which combines attribute regularization with the SIVAE framework to enhance the interpretability of the latent space while improving image generation/reconstruction capabilities. We demonstrated its effectiveness in overcoming the issue of blurry generation inherent in VAE-based methods. Nonetheless, the proposed method is limited by having a large number of hyperparameters, which makes it challenging to achieve convergence. Future

efforts will focus on extending attribute regularization to non-morphometric attributes and leveraging the interpretable latent space in downstream cardiac MRI classification tasks.

Acknowledgments. This research has been conducted using the UK Biobank Resource under Application Number 87065 and was supported by the German Federal Ministry of Health on the basis of a decision by the German Bundestag, under the frame of ERA PerMed. C.I.B. is in part supported by the Helmholtz Association under the joint research school “Munich School for Data Science - MUDS”.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bai, W., Sinclair, M., Tarroni, G., et al.: Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. *Journal of Cardiovascular Magnetic Resonance* **20**(1), 1–12 (2018)
2. Biffi, C., Cerrolaza, J.J., Tarroni, G., et al.: Explainable anatomical shape analysis through deep hierarchical generative models. *IEEE Trans Med Imaging* (2020)
3. Cetin, I., Stephens, M., Camara, O., et al.: Attri-vae: Attribute-based interpretable representations of medical images with variational autoencoders. *Computerized Medical Imaging and Graphics* (2023)
4. Daniel, T., Tamar, A.: Soft-introvae: Analyzing and improving the introspective variational autoencoder. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2021)
5. Engel, J., Hoffman, M., Roberts, A.: Latent constraints: Learning to generate conditionally from unconditional generative models. *arXiv:1711.05772* (2017)
6. Ghosh, P., Sajjadi, M.S., Vergari, A., Black, M., Schölkopf, B.: From variational to deterministic autoencoders. *arXiv preprint arXiv:1903.12436* (2019)
7. Hager, P., Menten, M.J., Rueckert, D.: Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In: *Conference on Computer Vision and Pattern Recognition* (2023)
8. Higgins, I., Matthey, L., Pal, A., Burgess, C., et al.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: *International Conference on Learning Representations* (2017)
9. Huang, H., He, R., Sun, Z., Tan, T., et al.: Introvae: Introspective variational autoencoders for photographic image synthesis. *Advances in Neural Information Processing Systems* **31** (2018)
10. Kalatzis, D., Eklund, D., Arvanitidis, G., Hauberg, S.: Variational autoencoders with riemannian brownian motion priors. *arXiv preprint arXiv:2002.05227* (2020)
11. Lample, G., Zeghidour, N., Usunier, N., et al.: Fader networks: Manipulating images by sliding attributes. *Advances in Neural Information Processing Systems* (2017)
12. Liu, W., Li, R., Zheng, M., et al.: Towards visually explaining variational autoencoders. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2020)
13. Mirza, M., Osindero, S.: Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784* (2014)

14. Pati, A., Lerch, A.: Attribute-based regularization of latent spaces for variational auto-encoders. *Neural Computing and Applications* (2021)
15. Petersen, S.E., Matthews, P.M., Francis, J.M., et al.: UK biobank’s cardiovascular magnetic resonance protocol. *J Cardiovasc Magn Reson* (2016)
16. Pidhorskyi, S., Almohsen, R., Doretto, G.: Generative probabilistic novelty detection with adversarial autoencoders. *Advances in Neural Information Processing Systems* (2018)
17. Razavi, A., Van den Oord, A., Vinyals, O.: Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems* **32** (2019)
18. Ridgeway, K., Mozer, M.C.: Learning deep disentangled embeddings with the f-statistic loss. *Advances in Neural Information Processing Systems* (2018)
19. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* (2019)
20. Sønderby, C.K., Raiko, T., Maaløe, L., Sønderby, S.K., Winther, O.: Ladder variational autoencoders. *Advances in Neural Information Processing Systems* **29** (2016)
21. Vahdat, A., Kautz, J.: Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems* **33**, 19667–19679 (2020)
22. Xu, H., Luo, D., Henao, R., Shah, S., Carin, L.: Learning autoencoders with relational regularization. In: *International Conference on Machine Learning*. pp. 10576–10586. PMLR (2020)
23. Zhang, R., Isola, P., Efros, A.A., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018)