Information Technology and Management Science
ISSN 2255-9094 (online)
2023, vol. 26, pp. 41–48
https://doi.org/10.7250/itms-2023-0006
https://itms-journals.rtu.lv

1862
RIGA TECHNICAL UNIVERSITY

Towards Explainability of the Latent Space by Disentangled Representation Learning

Ivars Namatēvs[1*], Kaspars Sudars[2], Artūrs Ņikuļins[3], Anda Slaidiņa[4], Laura Neimane[5], Oskars Radziņš[6]
[1]*Riga Technical University, Riga, Latvia*
[1–3]*Institute of Electronics and Computer Science, Riga, Latvia*
[4–6]*Rīga Stradiņš University, Riga, Latvia*

***Abstract* – Deep neural networks are widely used in computer vision for image classification, segmentation and generation. They are also often criticised as "black boxes" because their decision-making process is often not interpretable by humans. However, learning explainable representations that explicitly disentangle the underlying mechanisms that structure observational data is still a challenge. To further explore the latent space and achieve generic processing, we propose a pipeline for discovering the explainable directions in the latent space of generative models. Since the latent space contains semantically meaningful directions and can be explained, we propose a pipeline to fully resolve the representation of the latent space. It consists of a Dirichlet encoder, conditional deterministic diffusion, a group-swap and a latent traversal module. We believe that this study provides an insight into the advancement of research explaining the disentanglement of neural networks in the community.**

***Keywords* – Diffusion modelling, disentangled representation learning, explainability, latent space.**

I. Introduction

Existing end-to-end black-box deep learning models [1] are truncated and unable to extract the hidden attributes contained in representations with a generalisation capacity like that of humans. To fill this gap, the paradigm of representation learning is proposed [2]. Learning representations, in which various semantic aspects of the data are structurally disentangled, occupy a central place in training robust machine learning models [3]–[5]. A good latent representation should ideally reflect and disentangle the underlying mechanisms of data generation, ensure interpretability of the learned representations, and be usable for efficient classification and prediction tasks [6]. However, one problem with learning latent representations is their explainability. The projection into a latent space makes it difficult to investigate why the decision was made [7].

This means that we are dealing with the decomposability of latent space (the representation is decomposable) and it is, therefore, intelligible to man. More specifically, Rudin et al. [8] discuss major challenges related to supervised and unsupervised disentanglement of neural networks to design an interpretable model and explain a black box.

Learning independent and semantic representations whose individual dimensions have interpretable meaning is usually referred to as disentangled representations learning (DRL) [2], [9], [10]. On the other hand, disentangled representation learning [11] aims to learn the representation of the underlying explainable factors behind the observed data and it is considered one of the possible ways for AI to fundamentally understand the world.

Based on the motivation and requirements of DRL, there are numerous works on DRL and its applications to various computer vision tasks. DRL methods are usually based on generative models, such as VAE [12]–[15], GAN [16], [17], which initially have great potential for learning explainable representations for visual images. If VAE-based methods have an inherent trade-off between the ability to disentangle and the quality of generation [13], GAN-based methods suffer from the problem of reconstructing the difficulty of GAN inversion [18]. In addition, there are other DRL methods based on group theory [11] and causal inference [4]. However, little attention has been paid to representation learning [19] based on diffusion probabilistic models (DPM). There are only a few methods recently proposed for representation learning to reconstruct images in the context of DPM, such as Diff-AE [20] and PADE [19].

In this paper, we propose a novel pipeline for explainable DRL that combines for the first time the Dirichlet semantic autoencoder, conditional deterministic diffusion, the group-swap and a latent traversal.

Our main contribution can be summarised as follows:

- We explore the disentanglement capability of the latent space extracted from semantic autoencoder, diffusion modelling, group-based concept swapping and latent traversal, which can improve the explainability of the model.
- We propose a novel explainable pipeline for representation learning that is able to disentangle and explain the latent space (latent code) and has the potential for better image reconstruction in the context of multiple attribute disentanglement.

* Corresponding author. E-mail: ivars.namatevs@gmail.com
Article received 29.10.2023; accepted 13.11.2023

The rest of this paper is organised as follows: Section 2 gives theoretical background on disentanglement representation learning, diffusion probabilistic modelling, group-swap disentanglement, Dirichlet distribution and latent traversal. Section 3 proposes a novel pipeline to explain latent space by disentangled representation learning. Section 4 makes concluding remarks.

II. Background

A. Disentangled Representation Learning

Disentangled representation learning aims to learn a model that is able to identify and disentangle the underlying factors (concepts) hidden in the observable data in the form of representations. The core concept of the DRL architecture is to encourage the latent factors to learn disentangled representations while optimising the task objective, e.g., generating a discrimination objective [21].

DRL methods derive latent factors from visible features, with the generative assumption that each latent factor is responsible for generating a semantic attribute, e.g., lesion [22]. Let us suppose that an image can be generated by a set of semantically significant features such as colour, objects, shapes, etc. If these variation factors are captured separately in latent space and in an interpretable way, the image generation process becomes understandable and controllable by humans [8].

Algorithms [12]–[14], [23] which focus on learning disentangled representations map visual samples to a latent space where information belonging to different attributes is separated. The idea behind these algorithms is disentanglement by interpolation between attribute values, e.g., pose interpolation. However, these methods usually process one sample at a time rather than linking or contrasting a group of samples [22]. However, similar to human reasoning, we try to process samples in relation to the object of interest.

Wang et al. [21] propose a DRL taxonomy, which the authors divide into four main categories: (i) dimensional-wise vs. vectorial-wise, (ii) unsupervised vs. supervised, (iii) flat vs. hierarchical, and (iv) independent vs. causal. The first category can also be divided into vanilla VAE-based, group theory-based and GAN-based methods.

B. Dimension-wise DRL vs. vector-wise DRL

This group of DRLs depends on the structure of the latent space. Dimension-wise methods are characterised by having a single dimension or several dimensions, where each dimension represents a fine-grained generative factor. In vector-based methods, a vector represents a coarse-grained generative factor, but different vectors represent different types of semantic meaning [24], [25]. A typical architectural example of dimension-wise DRL is Variational Auto-encoder (VAE) based methods [12], [13], [24], [26], where different dimensions of the latent vector represent different factors. These methods adapt the idea of modelling data distributions from the maximum likelihood perspective by using variational inference, i.e., maximising $\log p_\theta(x)$. However, the vanilla VAE shows a weak disentanglement capability for relatively complex datasets [21]. To address this problem and improve the disentanglement capability, various inductive biases (explicit or implicit) have been integrated into the network architecture [9]. A common core point of the methods in this category is the implementation of disentanglement by designing specific loss objectives, e.g., by using different regularisers or specially designed supervised signals [21].

C. Unsupervised DRL vs. supervised DRL

This group of DRLs depends on the learning scheme. Unsupervised disentanglement is when we do not know the concepts or the case, or when the concepts are numerous and we do not know how to parameterise them [8]. Interpretability can be achieved by automatic factorisation of latent representations. For example, even the original VAE model [12] shows the possibility of unsupervised learning of the latent space by using Bayesian inference [21]. Supervised disentanglement is characterised by which concepts in the latent space are to be disentangled [8]. Recent work in this area aims to disentangle the latent space with respect to a collection of predefined concepts [27], [28].

D. Flat DRL vs. hierarchical DRL

Flat DRL can be characterised by the assumption that the architecture of generative factors is flat, i.e. the disentangled factors are parallel, at the same level of abstraction and there is no hierarchical structure between them [21]. Dimension-wise and vector-wise methods belong to the flat DRLs [13], [29]. When there are factors with different semantic abstraction levels among the latent representations, which are either independent [30] or dependent, we call them hierarchical [31].

E. Independent DRL vs. causal DRL

DRL models in which the latent factors are statistically independent are called independent DRL. Disentanglement can be completed by independent or factorial regularisation [13], [15] or various disentanglement losses [32], [33]. Causal DRL is characterised by the integration of causal mechanisms via the disentanglement of causal factors [34], [35].

Most disentanglement approaches aimed to separate content and increase the performance of networks, with the problem of explaianbility of latent space helping to overcome the black box nature of networks.

F. Diffusion Probabilistic Models

Machine learning methods can be divided into two types: *discriminative* and *generative* [36]. Generative models are a class of machine learning methods that learn a representation of the data on which they are trained and model the data itself [37]. They are usually based on deep neural networks. In practice, there are three mainstream generative models, namely, Generative Adversarial Networks (GAN), Variational Autoencoders (VAEs) and Normalizing flows (NF). GAN [38] is an end-to-end pipeline consisting of two networks: a generative one and a discriminator that trains the generator in an adversarial manner to produce samples that the discriminator can distinguish from the real data samples. VAE [12], [39] consists of an encoder and a decoder that can be operated independently. The encoder follows a projection of a data

sample onto a low-dimensional latent space and generates samples (reconstruction) from it via a decoding path [40]. Normalizing flows [39], [41] use an invertible flow function to transform the input into the latent space and generate samples with the inverse flow function parameterising the data distribution $p_\theta(x)$. More recently, there is a fourth generative model class, i.e., diffusion-based models (DPMs) and score-based generative models that model the target distribution by learning a denoising process with varying levels of noise [42].

G. Denoising Diffusion Probabilistic Models (DDPMs)

DDPMs aim to learn how to reverse the diffusion process and reconstruct the desired data patterns to generate new data. During training, a diffusion model transforms the original data samples *x*_0 by adding and removing noise ϵ. Recently, there have been several attempts to integrate the DRL paradigm into diffusion models [43]–[47]. Typical probabilistic diffusion model consists of four main components: the forward process, the reverse process, the training (optimisation) and the inference phase [48].

In standard DDPMs [42], in forward process q the original data are corrupted with Gaussian noise using a Markov chain of diffusion process that gradually perturbs the data x into random noise ϵ. The posterior converts the original data distribution $p_\theta(x)$ to the latent variable distribution $q(x_T)$ by gradually adding noise to the data according to a variance schedule (noise schedule) $\beta_1, \dots, \beta_T (\beta_t \in (0,1), 1 \le t \le T)$.

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \tag{1}$$

and

$$q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \boldsymbol{I}\right), \forall t \{1, \dots T\}, \tag{2}$$

where $\boldsymbol{I}$ is an identity matrix, t is a timescale. Setting $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_t$, the diffusion process allows sampling x_t at an arbitrary timestep t in closed form:

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}), \tag{3}$$

which can be further reparametrized:

$$\boldsymbol{x}_t(\boldsymbol{x}_0, \epsilon) = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon. \tag{4}$$

In the reverse process p_θ, the latent variable distribution $p_\theta(x_T)$ is transformed back to the data distribution $p_\theta(x_0)$ parametrized by θ.

$$p_\theta(\boldsymbol{x}_0, \dots, \boldsymbol{x}_{T-1}|\boldsymbol{x}_T) := \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t), \tag{5}$$

and

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, x_0) = \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t, t)\sigma_\theta(x_t, t)^2\boldsymbol{I}), \tag{6}$$

where $\mu_\theta(x_t,t)$ is learned mean and $\sigma_\theta(x_t,t)$ is variance. To generate a new image with the learned transition distribution p_θ from the reverse process, we first sample x_T from the standard Gaussian distribution and then sample x_{t-1} from $p_\theta(x_{t-1}|x_t)$ for $t = T, T-1, \dots, 1$.

H. Denoising Diffusion Implicit Model (DDIM)

Some of the main problems and limitations of diffusion models are their slow speed and the computational cost required. Several methods have been developed to address these drawbacks. DDIM [46] is one of the advancements that aims to speed up the sampling process. DDIM extends the DDPM by replacing the Markovian process with a non-Markovian process, resulting in a faster sampling process with negligible quality degradation. Song et al. [46] derive $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$, (6) using undetermined coefficient method, replacing x_0 with the following:

$$x_0 = \frac{1}{\alpha_t}(x_t - \bar{\beta}_t \epsilon_\theta(x_t, t) \tag{7}$$

and reuse DDPM parameter because $p(x_t|x_0)$ remains unchanged. Thus, we obtain:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, x_0) \approx \mathcal{N}\left(\frac{1}{\alpha_t}\left[x_t\left(\bar{\beta}_t - \alpha_t\sqrt{\bar{\beta}_{t-1}^2 - \sigma_t^2}\epsilon_\theta(x_t, t)\right)\right]\sigma_t^2 I\right). \tag{8}$$

When $\sigma_t = 0$, better quality was empirically observed with other σ settings, and the generation process p became deterministic, which was a stable mapping from the noise to reconstructed image.

I. Group-Based Disentanglement

Bengio et al. [2] introduced statistically independent disentangled representation learning as follows.

Definition 1. *Disentangled representation should separate the distinct, independent and informative generative factors of variation in the data. Single latent variables are sensitive to changes in single underlying generative factors, while being relatively invariant to changes in other factors.*

Based on this definition, the first studies on DRL methods were applied to independent component analysis (ICA) [49] and principal component analysis (PCA) [50]. Later, Higgins et al. [11] proposed a mathematically more rigorous DRL definition, formalised by analogy with physics and based on group representations of symmetry transformations.

Definition 2. *Assume we have the symmetry group G which can be decomposed as a direct product $G = G_1 \times G_2 \times \dots \times G_n$, world state space W (i.e., ground true factors which generate observations or an observation space), and the representation (latent) space Z. Representation Z is disentangled with respect to G if:*

(i) group action G on Z exists: $G \times Z \to Z$. Maps symmetry group on G to a linear representation (in algebraic sense) on Z.

(ii) the mapping f between the actions on W and Z exists $f: W \to Z$. This means finding a homonorphism $\rho : G \to GL(Z)$

between the symmetry group G and the general linear group of the latent space GL(Z) so that the map is equivariant.

(iii) if there is a subgroup decomposition of G such as that $G = G_1 \times G_2 \times \ldots \times G_n$ *. We would like to decompose the representation* (ρ, Z) *in subrepresentations* $Z_1 \oplus Z_2 \ldots \oplus Z_n$ *such that the restricted subrepresentations* $(\rho_{|G_i}, V_i)_i$ *are non-trivial and the restricted subrepresentations* $(\rho_{|G_i}, V_j)_{j \neq i}$ *are trivial.*

However, these representations cannot be learned without some form of interaction with the environment [51]. Another definition was proposed by Suter et al. [4] to define DRL as a property of a causal process instead of independence. They considered disentanglement as a property of a causal process [52] responsible for data generation, rather than just a heuristic property of encoding. Of the above DRL definitions, only Higgins et al. [11] provide a group-based definition; however, the authors do not propose a specific learning method based on their definition. Moreover, as pointed out by Quessard et al. [6], it is not easy to reconcile probabilistic inference methods with the group-based definition framework. However, it has been argued that for effective representation learning, one should not only consider static data, but also how these data can be transformed [11] or interacted with [53].

J. Dirichlet Distribution

The Dirichlet distribution is a continuous multivariate probability distribution defined over a set of discrete distributions [54]. Dirichlet distributions are often used as prior distributions in Bayesian learning [55]. The Dirichlet distribution is a composition of multiple Gamma random variables [56]. It is parameterised by a *K*-dimensional vector, typically referred to as the class concentration $\{\alpha_1, \ldots, \alpha_K > 0\}$ and a derived precision value $\alpha_0 = \sum_{c=1}^{K} \alpha_c$, where K corresponds to the set of discrete distributions. Thus, the probability density function (PDF) of a Dirichlet distribution is given by:

$$Dir(\mathbf{x}; \boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k x_k^{\alpha_k - 1}, \tag{9}$$

where Γ is the gamma function, α_k is a class concentration. The PDF of a Gamma distribution is given by:

$$\text{Gamma}(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \tag{10}$$

where β is a rate parameter and $\alpha_k, \alpha, \beta > 0$. If there are K independent variables following the Gamma distribution $X_k \sim Gamma(\alpha_k, \beta)$ or $\mathbf{X} \sim \text{MultiGamma}(\alpha, \beta \cdot \mathbf{1}_K)$, where $\mathbf{1}_K$ is all-one dimensional vector, then we have $\mathbf{Y} \sim \text{Dirilecht}(\boldsymbol{\alpha})$ where:

$$Y_k = \frac{X_k}{\sum X_i}. \tag{11}$$

It should be noted that the rate parameter, β, should be the same for every Gamma distribution in the composition. In our proposed pipeline, the set of discrete distributions represent the latent space.

K. Latent Traversal

The latent space of images is rich in semantics and traversing latent codes according to carefully selected trajectories and it provides the opportunity to make semantically meaningful transformations in the generated images [57]. The idea is the following: one value (dimension) of the row vector of the latent space z is adjusted, while the other values remain unchanged. Therefore, if you change one value incrementally, but keep the others fixed, you can generate a bunch of latent variations. One research direction uses interpretable directions without any prior knowledge (unsupervised) [58]. For example, [59] proposed learning a set of semantic concepts via an auxiliary classifier. Other studies [60], [61] use explicit human annotations to define semantic labels for interpretable directions.

III. Explainability of Disentanglement of the Latent Space

We assume that diffusion-based properties in combination with Dirichlet distribution, group-based disentanglement and latent traversal lead to a disentangled latent space, which enables the explainability of the latent space. The proposed pipeline consists of a Dirichlet probability-based semantic encoder, a diffusion-based conditional DDIM that includes a stochastic encoder and decoder, a group swap and a latent traversal module (see Fig. 1).

A. Semantic Encoder

The goal of semantic encoder $q_\phi(\mathrm{z}|x)$ is to map an input image into a subcode (semantic descriptive vector) $\boldsymbol{z}_{\text{sem}} \sim = Dir(\alpha)$ with the necessary information to help the decoder $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \boldsymbol{z}_{\text{sem}})$ decode and predict the output image. The architecture for the semantic encoder would be similar to of Dirichlet variational auto encoder (DirVA) [62]. We assume that latent space $\boldsymbol{z}_{\text{sem}}$ could be information-rich by using the DirVA and enable a more efficient denoising process. The proposed encoder should reconstruct the image linkage relation semantic concepts and visual features. Reconstruction loss is calculated for each image $L_r = \sum_{i=1}^{n} |x_i - f(x_i)|$, where f denotes the encoder-decoder functions. In the next step, the encoder provides a regularization, which regulates the training of the model by minimising the divergence between the approximated posterior distribution of the latent factors and prior distribution. The KL divergence is used, and the loss accordingly is calculated as $L_{KL} = D_{KL}(q_\phi(z|x)||p_\theta(z|x)$. Using latent representations as inputs each given concept is independently optimised with binary cross entropy loss (SGD). Once each concept is fully optimised, the ensemble of concepts $\{c_1, c_2, \ldots, c_m\}$ is trained together.

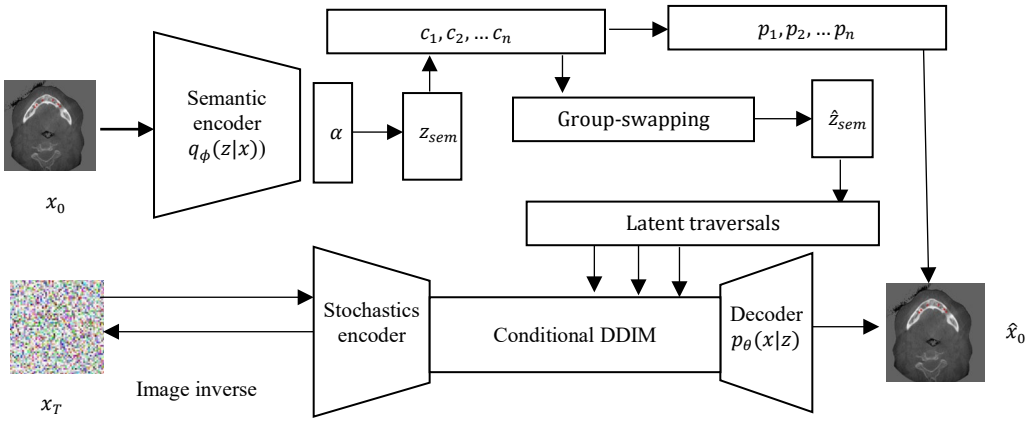

Fig. 1. The proposed latent space representative disentanglement explainability pipeline.

B. Stochastic Encoder

The aim of the DDIM stochastic encoder is to encode only the information left by $\boldsymbol{z}_{\text{sem}}$. There may be a possibility that not all information is compressed due to the stochasticity of $\boldsymbol{z}_{\text{sem}}$. The information left out by $\boldsymbol{z}_{\text{sem}}$ is encoded into the stochastic subcode $\mathbf{x}_T$ by running the deterministic generative process backwards. By using both semantic and stochastic encoders, our auto-encoder branch can capture an input image in great detail while providing a high-level representation for downstream tasks $\boldsymbol{z}_{\text{sem}}$. During training, the stochastic encoder is not used. Only for those tasks that require exact reconstruction or inversion, such as real image manipulation. To transform images into related noise codes, a stochastic encoder should be with the same posterior distribution $q(x_t|x_0)$ as DirVA.

C. Decoder

To obtain a meaningful latent code, the DDIM image decoder is chosen similarly to [46]. The conditional diffusion-based decoder receives as input $\boldsymbol{z} = (\boldsymbol{z}_{\text{sem}}, \mathbf{x}_T)$ to generate an image reconstruction, where $\boldsymbol{z}_{\text{sem}}$ comes from DirVA but $\mathbf{x}_T$ is from the stochastic encoder. We assume that a latent variable $\boldsymbol{z}_{\text{sem}}$ consists of the high-level semantic subcode $\boldsymbol{z}_{\text{sem}}$ and the low-level stochastic subcode x_T. To integrate $\boldsymbol{z}_{\text{sem}}$ into DDIM, a group normalisation (AdaGN) similar to [44], [62] is used by applying channel-wise scaling and shifting to the normalized feature map $\boldsymbol{h} \in \mathbb{R}^{c \times h \times w}$. Using the KL divergence, the distance between the variational posterior distribution $q_\phi(\boldsymbol{z}|x)$ and prior distribution $p_\theta(\boldsymbol{z})$ is reflected.

D. Latent space explainability

To ensure semantic consistency of concepts, extract attributes of concepts by leveraging semantic links between input images, and make a qualitative assessment of explainability, an implicit group-swap exchange module between concepts is used. We implement this by extracting image attributes from the descriptive latent vector. We assume that a fully disentangled descriptive vector will map image attributes with non-overlapping and rich attribute-based explanations, thus improving explainability in attribute extraction and decision making. More specifically, we divide the latent space into several semantically specific parts. For each input x, the encoder embeds the data into a low-dimensional vector by encoder. We then link d_i units of the vector to a particular concept c_i. Using the swapping operations, we enforce semantic consistency of N concept and extract features of N concepts by leveraging the semantic links between the input images. Finally, using the trained network, we calculate the loss function between two latent discriptive vectors. Our goal is to swap concepts between images by exchanging the corresponding entries in the latent representations. To do this we create and use a group-based dataset D_z (group latent codes).

Quality assessment of the explainability of the latent space is achieved by a series of latent traversals. Once the model is trained, we input the image from the test set as input. We first identify a single latent factor with the largest gradient activation, respectively, the row vector value responsible for a particular concept. The identified latent factor is adjusted or allowed to pass, while all other factors are preserved. If changing a single factor results in class-specific structural changes in the image reconstruction, we can assume that the latent space has been successfully disentangled into visual features for that class. We visualise the impact of changing a single attribute on the decoder image reconstructions. A pixel-wise variance is created to summarise the changes across a traversal and identify the features controlled by a particular latent attribute.

IV. Conclusion

We have explored the disentangling ability of the diffusion autoencoder and its potential to extract and explain high-level semantics by disentangling representations in latent space. We have presented a novel pipeline consisting of a conditional diffusion modelling approach capable of disentangling and explaining multiple attributes. The Dirichlet variational autoencoder has been presented as a novel approach for better concept disentangling representation in latent space. Group-swapping and latent traversal are the key to promote representation explainability, better understand the semantic information of the image and provide overlap-free concept representations. We believe that future research can explore the practical implications of our pipeline in image processing.

ACKNOWLEDGMENT

The research has been funded by the Latvian Council of Science, "A Deep Learning Approach for Osteoporosis Identification using Cone-beam Computed Tomography", grant number lzp-2021/1-0031.

INSTITUTIONAL REVIEW BOARD STATEMENT

The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of Rīga Stradiņš University (2-PĒK-4/336/2022, and 28/05.10.2017).

REFERENCES

[1] J. Egger, A. Pepe, C. Gsaxner, Y. Jin, J. Li, and R. Kern, "Deep learning – a first meta-survey of selected reviews across scientific disciplines, their commonalities, challenges and research impact", *PeerJ Computer Science*, vol. 7, 2021, Art. no. e773. https://doi.org/10.7717/peerj-cs.773

[2] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives", *TPAMI*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013. https://doi.org/10.1109/TPAMI.2013.50

[3] K. Ridgeway and M.C. Mozer, "Learning deep disentangled embeddings with the F-statistic loss", in *32nd Conference on Neural Information Processing Systems (NeurIPS2018)*, Montréal, Canada, 2018, pp. 1–10. https://proceedings.neurips.cc/paper_files/paper/2018/file/2b24d495052a8ce66358eb576b8912c8-Paper.pdf

[4] R. Suter, D. Miladinovic, B. Schölkopf, and S. Bauer, "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness," in *International Conference on Machine Learning*, PMLR, vol. 97, 2019, pp. 6056–6065. https://proceedings.mlr.press/v97/suter19a/suter19a.pdf

[5] D. Friede, C. Reimers, H. Stuckenschmidt, and M. Niepert, "Learning disentangled discrete representations", in *Machine Learning and Knowledge Discovery in Databases: Research Track. ECML PKDD 2023. Lecture Notes in Computer Science*, D. Koutra, C. Plant, M. Gomez Rodriguez, E. Baralis, and F. Bonchi, Eds., vol. 14172. Springer, Cham. https://doi.org/10.1007/978-3-031-43421-1_35

[6] R. Quessard, T. D. Barrett, and W. R. Clements, "Learning group structure and disentangled representations of dynamical environments", *arXiv*:2002.06991, 2020. https://doi.org/10.48550/arXiv.2002.06991

[7] M. Cerrato, A.V. Coronel, M. Köppel, A. Segner, R. Esposito, and S. Kramer, "Fair interpretable representation learning with correction vectors", *arXiv*:2202.03078v1, 2022. https://doi.org/10.48550/arXiv.2202.03078

[8] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges", *arXiv*:2103.11251v2, 2021. https://doi.org/10.48550/arXiv.2103.11251

[9] F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem, "Disentangling factors of variations using few labels," in *International Conference on Learning Representations (ICLR)*, 2020.

[10] X. Liu, P. Sanchez, S. Thermos, A.Q. O'Neil, and S.A. Tsaftaris, "Learning disentangled representations in the imaging domain", *arXiv*:2108.12043, 2021. https://doi.org/10.48550/arXiv.2108.12043

[11] I. Higgins, D. Amos, D. Pfau, S. Racaniere, L. Matthey, D. Rezende, and A. Lerchner, "Towards a definition of disentangled representations", *arXiv*:1812.02230, 2018. https://doi.org/10.48550/arXiv.1812.02230

[12] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes", *arXiv*:1312.6114, 2013. https://doi.org/10.48550/arXiv.1312.6114

[13] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *International Conference on Learning Representations (ICLR)*, 2016. https://openreview.net/forum?id=Sy2fzU9gl

[14] R. T. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in VAEs", *arXiv*:1802.04942, 2018. https://doi.org/10.48550/arXiv.1802.04942

[15] H. Kim and A. Mnih, "Disentangling by factorising", *arXiv*:1802.05983, 2018. https://doi.org/10.48550/arXiv.1802.05983

[16] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2016, pp. 2180–2188. https://proceedings.neurips.cc/paper_files/paper/2016/file/7c9d0b1f96aebd7b5eca8c3edaa19ebb-Paper.pdf

[17] Z. Lin, K. Thekumparampil, G. Fanti, and S. Oh, "InfoGAN-CR and ModelCentrality: Self-supervised model training and selection for disentangling GANs", *arXiv*:1906.06034, 2020. https://doi.org/10.48550/arXiv.1906.06034

[18] T. Wang, Y. Zhang, Y. Fan, J. Wang, and Q. Chen, "High fidelity GAN inversion for image attribute editing", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, USA, Jun. 2022, pp. 11379–11388. https://doi.org/10.1109/CVPR52688.2022.01109

[19] Z. Zhang, Z. Zhao, and Z. Lin, "Unsupervised representation learning from pre-trained diffusion probabilistic models", in *36th Conference on Neural Information Processing Systems*, 2022, pp. 1–14. https://openreview.net/pdf?id=IiCsx9KNVa0

[20] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation", in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022. https://doi.org/10.1109/CVPR52688.2022.01036

[21] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled representation learning", *arXiv*:2211.11695. 2023. https://doi.org/10.48550/arXiv.2211.11695

[22] Y. Ge, S. Abu-El-Haija, G. Xin, and L. Itti, "Zero-shot synthesis with group-supervised learning", in *International Conference on Learning Representations (ICLR)*, 2021.

[23] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-VAE", *arXiv*:1804.03599, 2018. https://doi.org/10.48550/arXiv.1804.03599

[24] H. Cheng, Y. Wang, H. Li, A. C. Kot, and B. Wen, "Disentangled feature representation for few-shot image classification," *arXiv*:2109.12548, 2021. https://doi.org/10.48550/arXiv.2109.12548

[25] S. Lee, S. Cho, and S. Im, "Dranet: Disentangling representation and adaptation networks for unsupervised cross-domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 15 252–15 261. https://doi.org/10.1109/CVPR46437.2021.01500

[26] R. T. Chen, X. Li, R. Grosse, and D. Duvenaud, "Isolating sources of disentanglement in VAEs," in *Proceedings of the 32nd Int. Conference on Neural Information Processing Systems*, 2019, pp. 2615–2625.

[27] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang, "Concept bottleneck models", in *Proceedings of International Conference on Machine Learning (ICML)*, 2020, pp 5338–5348.

[28] M. Losch, M. Fritz, and B. Schiele, "Interpretability beyond classification output: Semantic bottleneck networks", *arXiv*:1907.10882, 2019. https://doi.org/10.48550/arXiv.1907.10882

[29] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 1415–1424. https://doi.org/10.1109/CVPR.2017.141

[30] Z. Li, J. V. Murkute, P. K. Gyawali, and L. Wang, "Progressive learning and disentanglement of hierarchical representations," *arXiv*:2002.10549, 2020. https://doi.org/10.48550/arXiv.2002.10549

[31] A. Ross and F. Doshi-Velez, "Benchmarks, algorithms, and metrics for hierarchical disentanglement," in *International Conference on Machine Learning*, vol. 139, Jul. 2021, pp. 9084–9094. https://proceedings.mlr.press/v139/

[32] L. Liu, J. Li, L. Niu, R. Xu, and L. Zhang, "Activity image-to-video retrieval by disentangling appearance and motion," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1–9.

[33] H. Chen, Y. Zhang, X.Wang, X. Duan, Y. Zhou, and W. Zhu, "DisenBooth: Disentangled parameter-efficient tuning for subject driven text-to-image generation," *arXiv*:2305.03374, 2023. https://www.catalyzex.com/paper/arxiv:2305.03374

[34] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang, "Causal VAE: Disentangled representation learning via neural structural causal models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 9593–9602. https://doi.org/10.1109/CVPR46437.2021.00947

[35] X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang, "Disentangled generative causal representation learning," *arXiv*:2010.02637, 2020.
[36] J. Fragemann, L. Ardizzone, J. Egger, and J. Kleesiek, "Review of disentanglement approaches for medical applications", in *MICCAI MAD Workshop*, 2022. https://arxiv.org/ftp/arxiv/papers/2203/2203.11132.pdf
[37] M. Fan, C. Chen, C. Wang, J. Huang, "On the trustworthiness landscape of state-of-the-art generative models: A comprehensive survey", *arXiv*:2307.16680, 2023. https://doi.org/10.48550/arXiv.2307.16680
[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks", *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, Nov. 2020. https://doi.org/10.1145/3422622
[39] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models", in *International Conference on Machine Learning*, vol. 32, no. 2, 2014, pp. 1278–1286. https://proceedings.mlr.press/v32/rezende14.html
[40] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using real NVP", in *International Conference on Learning Representations*, 2017. [Online]. Available: https://arxiv.org/pdf/1605.08803.pdf
[41] G. Papamakarios, E.T. Nalisnick, D.J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference", *J. Mach. Learn. Res*., vol. 22, no. 57, pp. 1–64, 2021. https://jmlr.org/papers/volume22/19-1028/19-1028.pdf
[42] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoder: Towards a meaningful and decodable representation", *arXiv*:2111.12640, 2022. https://doi.org/10.48550/arXiv.2111.15640
[43] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", in *Proceedings of NeurIPS*, vol. 33, Vancouver, Canada, 2020, pp. 6840–6851. https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
[44] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis", *Advances in Neural Information Processing Systems*, vol. 34, 2021. https://openreview.net/pdf?id=AAWuCvzaVt
[45] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, Jun. 2022, pp. 10 684–10 695. https://doi.org/10.1109/CVPR52688.2022.01042
[46] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models", in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/pdf?id=St1giarCHLP
[47] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*, Montreal, QC, Canada, Oct. 2021, pp. 8162–8171. https://doi.org/10.1109/ICCV48922.2021.01410
[48] Z. Chang, G.A. Koulieris, and H.P.H. Shum, "On the design of fundamentals of diffusion models: A survey", *arXiv*:2306.04542v1, 2023. https://doi.org/10.48550/arXiv.2306.04542
[49] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications", *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, May–Jun. 2000. https://doi.org/10.1016/S0893-6080(00)00026-5
[50] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments", *Philos. Trans. A Math. Phys. Eng. Sci*., vol. 374, no. 2065, Apr. 2016, Art. no. 20150202. https://doi.org/10.1098/rsta.2015.0202
[51] H. Caselles-Dupré, M. Garcia Ortiz, and D. Filliat, "Symmetry-based disentangled representation learning requires interaction with environments", *Advances in Neural Information Processing Systems*, pp. 4608–4617, Jan. 2019.
[52] O. Oreshkov and C. Giarmatzi, "Causal and causally separable processes", *arXiv*:1506.05449, 2016. https://doi.org/10.48550/arXiv.1506.05449
[53] V. Thomas, J. Pondard, E. Bengio, M. Sarfati, P. Beaudoin, M.-J. Meurs, J. Pineau, D. Precup, and Y. Bengio, "Independently controllable factors", *arXiv*:1708.01289, 2017. https://doi.org/10.48550/arXiv.1708.01289
[54] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks", *arXiv*:1802.10501, 2018. https://doi.org/10.48550/arXiv.1802.10501
[55] J. Gawlikowski, S. Saha, A. Kruspe, and X. X. Zhu, "An advanced Dirichlet prior network for out-of-distribution detection in remote sensing", *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, Jan. 2022, Art. no. 5616819. https://doi.org/10.1109/TGRS.2022.3140324
[56] W. Joo, W. Lee, S. Park, and I.-C. Moon, "Dirichlet variational autoencoder", Pattern Recognition, vol. 107, 2020, Art. no. 107514. https://doi.org/10.1016/j.patcog.2020.107514
[57] Y. Song, T. Anderson Keller, N. Sebe, and M. Welling, "Latent traversals in generative models as potential flows", *arXiv*:2304.12944, 2023. https://doi.org/10.48550/arXiv.2304.12944
[58] X. Ren, T. Yang, Y. Wang, and W. Zeng, "Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view", in *ICLR*, 2022.
[59] A. Voynov and A. Babenko, "Unsupervised discovery of interpretable directions in the GAN latent space", in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119, pp. 9786–9796. https://proceedings.mlr.press/v119/voynov20a.html
[60] A. Plumerault, H. L. Borgne, and C. Hudelot, "Controlling generative models with continuous factors of variations", *arXiv*:2001.10238, 2020. https://doi.org/10.48550/arXiv.2001.10238
[61] Y. Shi, X. Yang, Y. Wan, and X. Shen, "SemanticStyleGAN: Learning compositional generative priors for controllable image synthesis and editing", in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 11244–11254. https://doi.org/10.1109/CVPR52688.2022.01097
[62] R. Harkness, A. F. Frangi, K. Zucker, and N. Ravikumar, "Learning disentangled representations for explainable chest X-ray classification using Dirichlet VAEs", *arXiv*:2302.02979, 2023. https://doi.org/10.48550/arXiv.2303.02979

Ivars Namatēvs received his Mg. sc. ing. degree from Riga Technical University and his MBA degree from Riga Business School. He is currently finalising his PhD thesis at Riga Technical University. He works as a Researcher at the Institute of Electronics and Computer Science. His research interests include explainable AI, disentangled representation learning and diffusion modelling and their application in medical imaging.
E-mail: ivars.namatevs@edi.lv
ORCID iD: https://orcid.org/0000-0002-5988-5558

Kaspars Sudars holds a PhD in Computer Science from the Faculty of Computing at the University of Latvia. He is also a co-author of 26 scientific publications indexed in SCOPUS, with a primary emphasis on research and development in the fields of signal processing, deep learning, and computer vision. Currently, his research interests span topics including explainable AI, semantic image segmentation, and object detection in images. Beyond his academic work, K. Sudars is a co-founder of WeedBot, a startup company committed to providing AI-based weeding solutions for delicate crops.
E-mail: sudars@edi.lv
ORCID iD: https://orcid.org/0000-0002-9110-9065

Artūrs Ņikuļins received his Mg. sc. ing. degree from Riga Technical University. He is currently studying computer science at Riga Technical University as part of the Doctoral study programme. He works as a Programming Engineer at the Institute of Electronics and Computer Science. His research interests include computer vision in the fields of agriculture, medicine and smart energy utilisation.
E-mail: arturs.nikulins@edi.lv
ORCID iD: https://orcid.org/0000-0002-3356-4764

Anda Slaidiņa received her PhD in Medicine form Rīga Stradiņš University in 2010. She is a Leading Researcher at the Department of Prosthetic Dentistry of the Faculty of Dentistry at Rīga Stradiņš University. Her current research interests include osteoporosis, cone beam computed tomography; artificial intelligence, dental education. She is a member of the Latvian Association of Dentists and the International College of Prosthodontists.
E-mail: anda.slaidina@rsu.lv
ORCID iD: https://orcid.org/0000-0001-8353-4187

Laura Neimane received her PhD from Rīga Stradiņš University (RSU) in 2013. She is an Associate Professor at the Faculty of Dentistry at RSU in Riga, Latvia, and the Head of the Department of Diagnostic Radiology at the Institute of Stomatology at RSU. She is also a member of several associations: Latvian Association of Dentists, Latvia Radiologist association, European Radiologist association, European Association of Dentomaxillofacial Radiology, International Association of Dentomaxillofacial Radiology and IDAR.
E-mail: laura.neimane@rsu.lv
ORCID iD: https://orcid.org/0000-0002-5411-5810

Oskars Radziņš obtained an MEng degree in Biomedical Engineering from the University of Glasgow in 2017. Currently, he is pursuing a PhD in Computer Science at Riga Technical University. He currently works in Riga, Latvia, at Rīga Stradiņš University, Institute of Stomatology and the Baltic Biomaterials Centre of Excellence as a Medical Engineer. While his work is focused on surgical planning, use of 3D printing and personalized implant development, his research interests are related primarily to orthognathic surgery and clinical application of artificial intelligence.
E-mail: oskars.radzins@rsu.lv
ORCID iD: https://orcid.org/0000-0002-2443-9582