Medical Image Analysis 80 (2022) 102516

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Learning disentangled representations in the imaging domain

Xiao Liu[a,1,*], Pedro Sanchez[a,1], Spyridon Thermos[a,1], Alison Q. O'Neil[a,b], Sotirios A. Tsaftaris[a,c]

[a] School of Engineering, The University of Edinburgh, Edinburgh EH9 3FG, UK
[b] Canon Medical Research Europe, Edinburgh EH6 5NP, UK
[c] The Alan Turing Institute, London NW1 2DB, UK

ARTICLE INFO

Article history:
Received 6 November 2021
Revised 5 April 2022
Accepted 10 June 2022
Available online 17 June 2022

Keywords:
Disentangled representation
Content-style
Applications
Tutorial
Medical imaging
Computer vision

ABSTRACT

Disentangled representation learning has been proposed as an approach to learning general representations even in the absence of, or with limited, supervision. A good general representation can be fine-tuned for new target tasks using modest amounts of data, or used directly in unseen domains achieving remarkable performance in the corresponding task. This alleviation of the data and annotation requirements offers tantalising prospects for applications in computer vision and healthcare. In this tutorial paper, we motivate the need for disentangled representations, revisit key concepts, and describe practical building blocks and criteria for learning such representations. We survey applications in medical imaging emphasising choices made in exemplar key works, and then discuss links to computer vision applications. We conclude by presenting limitations, challenges, and opportunities.

1. Introduction

Imagine the need to develop a method to localise the ventricles in Magnetic Resonance Imaging (MRI) and Computerised Tomography (CT) scans of the brain in patients. This method must be robust to any changes in the imaging process, scanner, and noise, as well as to anatomical and pathological variation. The current deep (supervised) learning paradigm indicates that we *must* present to the system as many examples as possible to instill robustness by learning what is unnecessary, or nuisance (Achille and Soatto, 2017), *e.g.* the patient being placed at a rotated angle in the scanner, as opposed to what matters, *i.e.* the location of the ventricle. However, collecting and annotating enough data to cover such real-world variation is an unrealistically time-consuming and costly solution.

Surprisingly, we may not always need annotated data or carefully crafted data augmentations to achieve this. With disentangled representation learning (DRL), one learns to encode the underlying factors of variation into separate latent variables (Bengio et al., 2013; Higgins et al., 2018), which ultimately capture sensitive and useful information for the task at hand and also understand the underlying causal relations amongst the variables. We choose to introduce the reader to DRL by presenting 3 indicative examples of disentangled factors in Fig. 1, which affect the colour, scale, and rotation of the rendered object in the corresponding scene. By adopting DRL, one can design deep models that will be robust to representations from unseen domains, a result that cannot always be achieved through data augmentation.

Herein,[2] we aim to unite understanding of disentangled representation learning and its applications in medical imaging. Our goals are to:

1. Expose why (in)variance matters in learning.
2. Understand the impact of causal relations in the context of disentangled representations.
3. Enforce that disentanglement requires at least one of: inductive biases, priors, or supervision.
4. Expose building blocks for encouraging disentanglement.
5. Survey medical image analysis applications.
6. Draw inspiring lessons from computer vision applications.
7. Identify limitations, and discuss opportunities and remaining challenges.

* Corresponding author.
E-mail address: Xiao.Liu@ed.ac.uk (X. Liu).
[1] Equal contribution.

[2] This paper follows a tutorial style, but also surveys a considerable (more than 260 citations) number of works. It aims to be concise and not fully survey the field (in 2021 alone, 1700 papers have appeared on arXiv on disentangled representations). This paper is inspired by two tutorials at MICCAI on disentangled representations (https://vios.science/tutorials/dream2021) and benefits greatly from feedback received from participants in these tutorials.

https://doi.org/10.1016/j.media.2022.102516

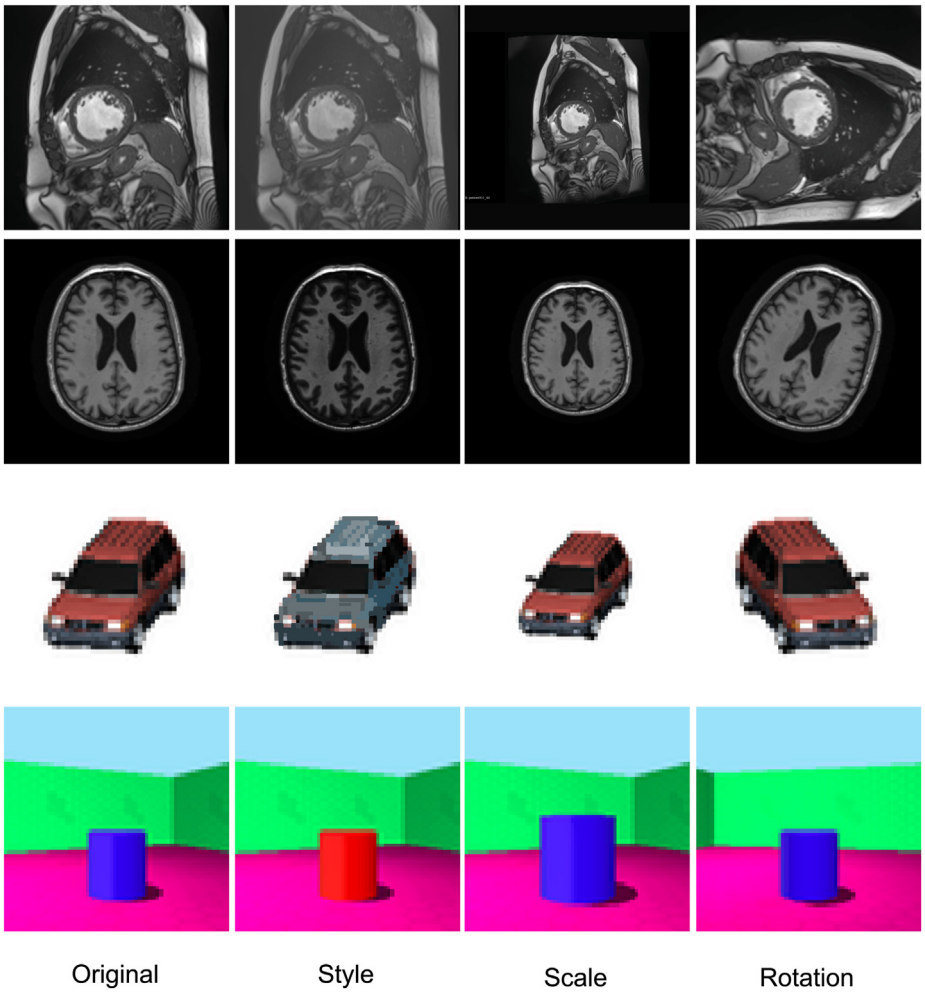

Fig. 1. Examples of factors of variations: style, scale, and rotation in the context of cardiac scans (Bernard et al., 2018), brain scans (Petersen, 2010), cars (Reed et al., 2015), and 3D shapes (Burgess and Kim, 2018).

To define disentanglement we first revisit key concepts in learning representations. We then provide an overview of key generative frameworks forming the basis of many subsequent models; building blocks of disentanglement; and evaluation metrics. We discuss exemplar models designed to address applications of disentanglement in the medical imaging and computer vision domains. We conclude by discussing opportunities and challenges. This paper is also accompanied by a repository offering links to the implementations of key methods and to existing metrics: https://github.com/vios-s/disentanglement_tutorial.

2. Revisiting key concepts in representation learning

Notation. We use x, $\mathbf{x}$ and X to denote scalars, vectors, and higher-dimensional tensors respectively, drawn from the domain $\mathcal{X}$ of corresponding dimensions. We use X_i to refer to a datum of the above tensors (of any dimension) for presentation simplicity where tensor dimensionality is implied by the context. We will assume we have access to a dataset containing samples of X_i, where $i \in [1, N]$, N denoting the number of samples. We use $\mathcal{X}$ to denote the observed variables of the input domain, $\mathcal{Z}$ for latent representations, $\mathcal{S}$ for real generating factors, and $\mathcal{Y}$ for the output domain. For example, if we choose to solve a classification task, then $\mathcal{Y}$ is a space of scalars y.

2.1. Model learning

We consider the task of learning a mapping between two domains (Vapnik, 1999), *i.e.* $f : \mathcal{X} \to \mathcal{Y}$. We will split f into two components, $f : E_\phi \circ D_\theta$. E_ϕ maps to an intermediate latent representation $\mathcal{Z}$ ($E_\phi : \mathcal{X} \to \mathcal{Z}$) whereas D_θ maps to the output ($D_\theta : \mathcal{Z} \to \mathcal{Y}$). We will term E_ϕ the "encoder" and D_θ the "decoder".[3] Thus, the goal of model learning is the solution of the task at hand by learning a good representation. Below, we discuss desirable properties of a good representation.

2.2. Representation learning

Finding good representations for the task at hand is fundamental in machine learning (Bengio et al., 2013; Schlkopf et al., 2021). Consider the task of detecting brain tumours by placing a bounding box $\mathbf{y}_i$ around each tumour in the image X_i. A dataset may contain brain samples with different morphologies, acquired using different protocols in different sites (hospitals), etc. Our goal is to create a representation suitable for the task. If the tumour changes location in the image, we would like the bounding box output to change location accordingly; our representation will be *equivariant* to the location of the object of interest. On the other hand, we

[3] D_θ is often referred as a classifier or a regressor, however we avoid this nomenclature here to be more general.

would like the representation to be *invariant* to acquisition-related changes.

Symmetries. Symmetries Ω are *transformations* that leave some aspects of the input intact (Cohen and Welling, 2016; Cohen, 2021; Bronstein et al., 2021). For instance, the category of an object does not change after applying shift operations to the image, therefore these operations are considered symmetries in the object recognition domain. Using the model f and symmetries Ω, we now proceed to define the equivariance and invariance properties.

Equivariance. A mapping $E_\phi : \mathcal{X} \to \mathcal{Z}$ is equivariant *w.r.t.* Ω, if there is a transformation $\omega \in \Omega$ of the input $X \in \mathcal{X}$ that affects the output $Z \in \mathcal{Z}$ in the same manner. Formally, this means that Ω-equivariance of E_ϕ is obtained when there exists a mapping $M_\omega : R^d \to R^d$ applying ω to an input such that:

$$E_\phi(M_\omega \circ X) = M_\omega \circ E_\phi(X), \ \forall \omega \in \Omega. \tag{1}$$

In practice, one chooses transformations that induce the desired equivariance and learned properties in accordance with the task at hand, thus a good understanding of the problem (also known as *domain knowledge*) is required (Lenc and Vedaldi, 2015). Classical examples where equivariance to translation, shift, and mirroring might be important, are image segmentation, pose estimation, and landmark detection tasks.

Invariance. A special case of equivariance occurs when M_g becomes the identity map. Formally, E_ϕ is invariant to transformations of Ω if:

$$E_\phi(M_\omega \circ X) = E_\phi(X), \ \forall \omega \in \Omega. \tag{2}$$

The transforms we want to adhere to are usually task-specific and as we will highlight in Section 6 typically enforced via design biases (and costs) to approximate the transformations.

2.3. Generating factors

Considering a distribution that characterises the domain $\mathcal{X}$, the *generating factors* $\mathcal{S}$ are the underlying variables that fully characterise the variation of the data –seen or expected to be seen. Recent studies (Bengio et al., 2013; Schlkopf et al., 2021) argue that representations should enable the decomposition (*i.e.* disentanglement) of the input data into separate factors. Each factor should correspond to a variable of interest in the underlying process that generated the data. For the rest of the paper we will refer to the real-world generating factors as "real" and to those learned by a model as "learned".

In the brain tumour detection example, several variables such as tumour texture/location, brain shape, acquisition protocol, image contrast, etc. may be involved. In general, the more complex the image, the more variables, and the higher the number of possible combinations. Enumerating all these combinations readily leads to a combinatorial explosion in the possible combinations that a dataset must contain to enable a model to learn (from data alone) the desired in/equi-variances. It is not realistic to identify every factor and cover every possible combination. Domain knowledge enables the elucidation of as many factors as possible and allows us to define which real factors we want to be in/equi-variant to.

2.4. Domain shifts

An *i.i.d.* data distribution is easy to consider but forms a strong and often unrealistic assumption. All non-synthetic datasets are somewhat biased due to the finite nature of the acquired data. If learning algorithms are trained with standard supervised learning (Vapnik, 1999) without additional assumptions, there is little hope that the learned function will be robust to domain shifts. A model's ability to maintain the desired behaviour across domain changes is also referred to as *out-of-distribution* generalisation (Shen et al., 2021). For the brain tumour detection example, both CT or MRI scanners acquire images, but we might know that a given hospital uses CT. In this case, modality-related factors are linked to the hospital-related variables. Therefore, understanding the data generation process and the underlying relations between variables can help to distill the important visual information, and to create mechanisms that are more generalisable. Such reasoning enables the design of principled strategies for mitigating the data bias (Castro et al., 2020). In fact, we can explicitly define the changes we want our model to be invariant or equivariant to, by modeling domain shifts such as: i) population, *i.e.* different cohorts, ii) acquisition, *i.e.* different cameras, sites or scanners, and iii) annotation shift, *i.e.* different annotators.

2.5. Disentangled representations

Disentangled representations can address some of the challenges described until now by learning representations with equi/in-variances to specific undesired variables, whilst considering the data generation process and potential domain shifts. Although a widely accepted definition of disentangled representations is yet to be defined, the main intuition is that by disentangling, we separate out the main factors of variation that are present in our data distribution (Bengio et al., 2013; Higgins et al., 2018; Caselles-Dupré et al., 2019; Locatello et al., 2019b). We characterise a factor as "disentangled" when any intervention on this factor results in a specific change in the generated data (Caselles-Dupré et al., 2019; Thomas et al., 2017).

2.5.1. Formalising disentanglement

Higgins et al. (2018) have recently presented a generic definition for disentanglement. Given a compositional world W and a set of transformations Ω (as defined in Section 2.2), they define a function $f : W \to Z$ that can induce Ω in the latent representation $Z \in \mathcal{Z}$ in an equivariant manner. The representation Z is defined as "disentangled" if there is a decomposition $Z = Z_1 \times \cdots \times Z_n$ such that a transformation ω applied on Z_i will result in an equivalent transformation in the input domain $\mathcal{X}$, leaving all other aspects controlled by $Z_{j \neq i}$ unchanged. This definition meets the desired properties of a disentangled representation as defined by several works in DRL (Bengio et al., 2013; Chen et al., 2016; Eastwood and Williams, 2018; Ridgeway and Mozer, 2018): a) modularity, *i.e.* each latent dimension should encode no more than one generative factor, and b) informativeness, *i.e.* all underlying generative factors are encoded in the representation.

A complementary view to the definition of Higgins et al. (2018) comes from the Information Bottleneck (IB) principle introduced in Tishby et al. (1999). IB allows for learning "good" representations for the task at hand, by trading-off sufficiency and complexity. Adopting IB, Achille and Soatto (2017) argue that such representations should be: i) *sufficient* for the task, meaning that we do not discard information required for the output; ii) among all sufficient representations, it should be *minimal* retaining as little information about the input as possible; and finally iii) it should be *invariant* to nuisance effects so that the final classifier will not overfit to any correlations between the dataset nuisances and the ground truth labels.

2.5.2. Identifiability

Learning disentangled representations without any type of supervision is impossible as an infinite family of models that could have generated the observed data exist Locatello et al. (2019b). Thus, *identifying* the model that generated the data without any additional information is impossible. Given an observation X_i, there is an infinite number of generative models that could have generated a sample from the same marginal distribution (Locatello et al., 2019b; Peters et al., 2017; Khemakhem et al., 2020).

This follows from prior work in non-linear independent component analysis (ICA) Hyvärinen and Pajunen (1999): even though the linear case is identifiable, the flexibility given by the non-linear case makes it non-identifiable without extra information. Recently Khemakhem et al. (2020), bridged the gap between non-linear ICA and other deep latent variable models, and showed that unsupervised disentanglement methods are, indeed, non-identifiable without additional assumptions.

2.5.3. A causal perspective

Learning disentangled representations of the real factors is not ideal if these factors are not truly independent of each other and are connected via causal relations. Causal relations are directional: the effect will change given an *intervention* (change) on the cause, but not the other way around. For example, the presence of the heart causes the presence of atria, ventricles, pericardium, etc. If an intervention removes the heart, the other structures will also disappear.

Therefore, causal representation learning can be seen as extending DRL (Suter et al., 2019) with additional constraints on the relationships between the latent variables (Schlkopf et al., 2021; Castro et al., 2020) and potential biases and domain shifts (Section 2.4). Recent advances in DRL (Higgins et al., 2017; Chen et al., 2018; Kim and Mnih, 2018; Locatello et al., 2019b) can be cast as learning the causal variables (*i.e.* the generating factors) of a problem without explicitly modeling the causal mechanisms between them. In addition, identifiability (Section 2.5.2) can be extended to causality: it is impossible to infer either the latent variables of the generative process or the relationships between them from observational data alone (Peters et al., 2017).

2.5.4. Disentanglement as inductive bias

The solution to identifiability is the use of domain knowledge, *i.e.* the *inductive bias*, instead of using explicit supervision (Locatello et al., 2019b; Peters et al., 2017; Khemakhem et al., 2020). Current representation learning already benefits from the inductive biases of Convolutional Neural Networks (CNNs) (Lecun et al., 1998) and Recurrent Neural Networks (RNNs) (Graves et al., 2013). Outside of the visual domain, language has been modeled with recurrent neural networks that capture the sequential nature of data for making predictions (LeCun et al., 2015). Recent attention and self-attention models, such as the transformer architecture (Vaswani et al., 2017), focus on learning the internal structure of the input data. These self-attention models essentially learn the best inductive biases for each sample in the data distribution.

Overall, disentanglement priors add structure to the learned representations to better correspond to the underlying generation process. It is this useful bias that makes the utilised models identifiable. One of the goals of this paper is to highlight the various inductive biases used.

3. Frameworks enforcing disentanglement

We now briefly review fundamental generative models that typically are used to learn disentangled tensor spaces.

3.1. Variational autoencoders

Standard Auto-Encoders (AEs) or Variational Auto-Encoders (VAEs) (Kingma and Welling, 2014; Rezende et al., 2014) decompose factors via image reconstruction (Cheung et al., 2015; N et al., 2017). A typical VAE, depicted in Fig. 2(a), discovers and disentangles factors of variation by forcing independence between different dimensions of $\mathbf{z}$, while reconstructing the input X. Inter-factor independence is achieved by minimising the Total Correlation (TC) objective imposed on the inferred latent vector (Watanabe, 1960).

A widely-used VAE that encourages disentanglement is the β-VAE (Higgins et al., 2017). Its main objective is the maximisation of the Evidence Lower-Bound Optimisation (ELBO), $(\mathcal{L}(\theta, \phi; X, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|X)}[\log p_\theta(X \mid \mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z} \mid X)||p_\mathbf{z}))$, to balance (via $\beta > 1$) the reconstruction error *versus* adherence to the approximate posterior $q_{\mathbf{z}|X}$ from the latent prior $p_\mathbf{z}$. Note that $p_\mathbf{z}$ is usually a normal distribution with identity covariance matrix $\mathcal{N}(0, \mathbf{I})$. The diagonal covariance forces an orthogonal factorisation of the latent space, similarly to a PCA, which reasonably explains the disentanglement capabilities of VAEs (Rolinek et al., 2019; Burgess et al., 2018; Rolinek et al., 2019). A $\beta > 1$ encourages disentanglement by forcing $q(\mathbf{z} \mid X)$ to carry less information about the reconstruction by increasing the weight of the KL divergence term (Burgess et al., 2018) and consequently, increasing independence between the factors of $\mathbf{z}$. Adding more terms such as TC as exploited by several VAE-based models (Chen et al., 2018; Kim and Mnih, 2018; Esmaeili et al., 2019) further restricts the redundancy.

3.2. Generative adversarial networks

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), see Fig. 2(b), typically employ a generator G and a discriminator D in an adversarial game. G generates an image by sampling from an isotropic Gaussian distribution, while D is given the synthetic image and a real one (X), and tries to identify which input is real/fake.

Recent advances in GAN design and training have led to high-fidelity image generation (Karras et al., 2019; Brock et al., 2019; Liu et al., 2020a). GANs can learn disentangled representations by adding regularisation terms (Chen et al., 2016), by creating an architectural prior (Karras et al., 2019), or even by a post-hoc decomposition of the learned manifold after training (Shen and Zhou, 2021).

A milestone approach in regularisation is InfoGAN (Chen et al., 2016) which encourages the disentanglement between two groups of latent variables: a) $\mathbf{z}$ which encodes unstructured noise; and b) $\mathbf{c}$ which captures structured features of the data distribution. They approach this by maximising the mutual information (MI) lower bound between $\mathbf{c}$ and the generated data. Cluster-GAN (Mukherjee et al., 2019) extends the InfoGAN setting (adopting only the discrete version of $\mathbf{c}$) by employing an inverse-mapping network to project the generated data back to the latent space. This process is supervised by a clustering loss that operates as a regulariser.

Architectural priors were introduced by Karras et al. (2019, 2020). A mapping network transforms the latent variable $\mathbf{z}$ into intermediate variables that control the style at each convolutional layer of the generator (G). Interestingly, this enables feature manipulation at different levels of granularity, *e.g.* from shape down to texture. This hierarchical structure constitutes arguably a strong prior for disentanglement (Nie et al., 2020; Peebles et al., 2020).

3.3. Normalising flows

Differently from the non-invertible VAEs and GANs, the Normalising Flows (NFs) are a family of invertible probabilistic models that can compute the exact –and not the approximated as in VAEs– likelihood (Dinh et al., 2015; Rezende and Mohamed, 2015; Kingma and Dhariwal, 2018; Papamakarios et al., 2021). The NF framework derives from the change of variables formula in probability distributions (Dinh et al., 2015). Considering a variable $X = \delta^{-1}(\mathbf{z})$, where $\mathbf{z} \sim p(\mathbf{z})$ is sampled from a *prior* distribution $p(\mathbf{z})$, the posterior of X can be obtained by: $p(X) = p(\mathbf{z}) \left| \det \frac{\partial \delta^{-1}}{\partial \mathbf{z}} \right|^{-1}$. The *flow* $\delta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is constrained to be a *diffeomorphism, i.e.* differentiable and invertible transformations, with a differentiable inverse

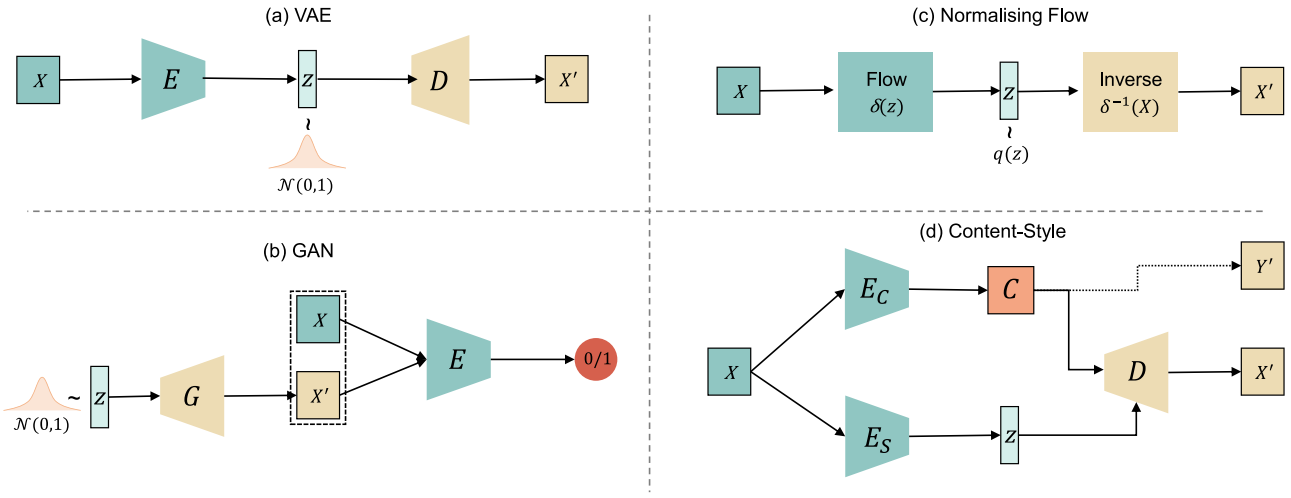

Fig. 2. Fundamental architectures for disentanglement: a) VAE, b) GAN, c) Normalising Flows, d) Content-Style disentanglement. X and X' are the input and reconstructed images. $\mathbf{z}, C$ are the latent representations, where C represents a tensor latent variable (*e.g.* image content) and $\mathbf{z}$ represents a vector latent variable. The dashed line in (d) denotes the use of C for learning a representation Y' for a parallel equivariant task (*e.g.* semantic segmentation). Finally, $\mathcal{N}$ denotes the normal distribution with zero mean and unit variance, whilst $q(\mathbf{z})$ can be any prior distribution.

(δ^{-1}). When multiple flows (δ_i^{-1}) are combined in a chain, they can approximate arbitrarily complex densities for $p(X)$. As Fig. 2(c) illustrates, δ can encode an image X into a latent space $\mathbf{z}$ or using the inverse flow δ^{-1} to create a generative model by decoding a sample $\mathbf{z} \sim p(\mathbf{z})$ into image space. NFs have been recently adapted to encode disentangled representations (Esser et al., 2020; Sankar et al., 2021) by reinforcing similarity between latent spaces $\mathbf{z}$ of pairs of images with similar generating factors. We refer the reader to Kobyzev et al. (2020) for a comprehensive review.

3.4. Content-style disentanglement

The aforementioned models typically decompose factors into a single vector representation. However, a recent trend in disentanglement focuses on the decomposition of the input image into different latent variables that encode different properties, such as geometry *vs.* style. This form of disentanglement is the so-called Content-Style Disentanglement (CSD) (Gatys et al., 2016), where an image is decomposed into domain-invariant "content" and domain-specific "style" representations (Gabbay and Hoshen, 2020; Ruta et al., 2021). Most works in CSD encode content in spatial (tensor) representations to preserve the spatial correlations and exploit them for a spatially equivariant task, such as Image-to-Image (I2I) translation (Huang et al., 2018; Lee et al., 2018) and semantic segmentation (Chartsias et al., 2019). The corresponding style, *i.e.* the information that controls the image appearance such as colour and intensity, is encoded in a vector. An abstract visualisation of a CSD model is depicted in Fig. 2(d). Note that decomposing content from style is not a trivial process, and encoding content as a high-dimensional representation is not enough. Recent work introduces several design (in terms of the model architecture) and learning (in terms of loss functions) biases to achieve this separation. We denote these inductive biases as "building blocks" and discuss them in the following section.

4. Disentanglement building blocks

We now describe common layers and modules that are used at various levels of the model design to encourage disentanglement. We associate these so-called building blocks with different high-level parts of the aforementioned AEs and generative models. We note that typically several of these are combined. In principle we would like to have the minimal set required to solve the task, noting thought that at times these blocks can compete.

4.1. Encoding modules

The following are commonly used at various levels of the encoder(s) in popular architectures as bottlenecks. We use representation bottlenecks as a way of reducing the amount of information in the data which will force the network to encode mainly useful concepts.

Instance Normalisation. Instance Normalisation (IN), originally proposed in (Ulyanov et al., 2017) for style removal, is commonly used after each convolutional layer of the content encoder to suppress style-related information. In fact, IN removes any contrast-related information from each instance (data sample), encouraging content-related features to be propagated to the following layers. An indicative example is the content encoder of Huang and Belongie (2017), where IN replaces all batch normalisation layers (Ioffe and Szegedy, 2015).

Average Pooling. Contrary to IN, average –or global– pooling is commonly used to suppress the content information in the style encoder (Huang and Belongie, 2017). By averaging values and flattening a spatial feature into a vector, this operator removes any spatial correlation and encodes the global mean statistics (*i.e.* image style).

Parsimony. For CSD models that require semantic and parsimonious content for parallel spatially equivariant tasks, there is a need for discretisation of the encoded continuous information. Such discretisation also can help to remove style-related information. The Gumbel Softmax operator is a differentiable solution to this problem. This operator mimics the reparametrisation trick performed in VAEs by sampling from a standard Gumbel distribution and using the Softmax as an approximation of the "argmax" step that is usually coupled with one-hot operators for discretisation. Another tool that can further restrict the amount of information in a latent space is known as Vector Quantisation (VQ) (Van Den Oord et al., 2017). VQ uses a dictionary of learnable entries to restrict the latent features to discrete set of values.

4.2. Entanglement modules

Effective recombination or entanglement of the content and style representations in a decoder is vital. The following ap-

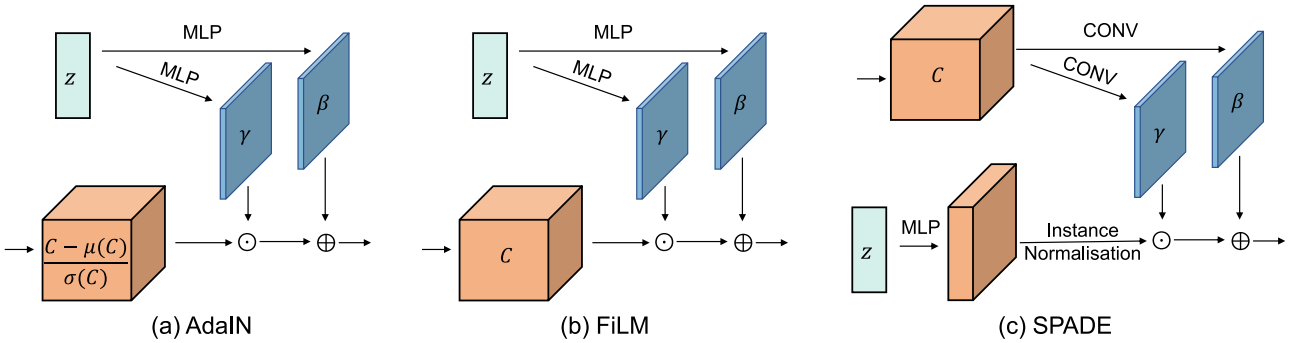

Fig. 3. Disentanglement building blocks that combine content C with style $\mathbf{z}$: a) AdaIN, b) FiLM, and c) SPADE. $\odot$ and $\oplus$ denote element-wise multiplication and addition, respectively. MLP and CONV denote multilayer perceptron and convolutional layers.

proaches or layers are commonly used for this purpose at various levels of the decoder in popular CSD architectures.

Concatenation. Simple concatenation allows the content and style to be more flexible in capturing the desired information (Lee et al., 2018; Esser et al., 2018). However, this may limit the controllability of learning the content and style as the representations may not capture desired information *e.g.* style representation capturing the shape information.

Adaptive Instance Normalisation. The Adaptive Instance Normalisation (AdaIN) layer (Huang and Belongie, 2017) is commonly used at multiple decoder levels to recombine the content and style representations. As depicted in Fig. 3(a), each AdaIN layer performs the following operation:

$$\text{AdaIN} = \gamma \frac{C_j - \mu(C_j)}{\sigma(C_j)} + \beta_j, \tag{3}$$

where each feature map C_j is first normalised separately, and then is scaled and shifted based on γ and β, which are parameters of an affine transformation of the style representation (adaptive mean and standard deviation).

Feature-wise Linear Modulation. As shown in Fig. 3(b), Feature-wise Linear Modulation (FiLM) (Perez et al., 2018) is similar to AdaIN. FiLM was initially proposed as a conditioning method for visual reasoning (the task of answering image-related questions). Using FiLM, each channel of the networks intermediate features C_j is modulated based on γ_j and β_j as follows: $FiLM(C_j|\gamma_j, \beta_j) = \gamma_j \cdot C_j + \beta_j$, where element-wise multiplication $(\cdot)$ and addition are both broadcast over the spatial dimensions. It is used in Chartsias et al. (2019) to combine the content and style in the decoder, where γ and β parameterise the affine transformation of style vectors.

Spatially-Adaptive Denormalisation. An alternative approach for combining content with style is the use of multiple Spatially-Adaptive Denormalisation (SPADE) (Park et al., 2019) layers. As depicted in Fig. 3(c), a SPADE block receives the content channels and projects them onto an embedding space using two convolutional layers to produce the modulation parameters (tensors) γ and β. These parameters are then used to scale (γ) and shift (β) the normalised activations of the style representation.

4.3. Encouraging disentanglement in the latent space

The following operations and priors can be applied on a latent space to encourage disentanglement.

Gaussian prior. Encouraging the distribution of the encoded (vector) latent representation to match a Gaussian is a common prior. As reported in Section 3.1, such prior encourages the unsupervised disentanglement of the factors of variation and enables sampling for generating new images.

Task priors. As discussed in Section 3.4, content representation can be used for a downstream equivariant task, *e.g.* semantic segmentation. Task losses, such as the segmentation loss, also contribute at learning a disentangled content representation (Chartsias et al., 2019). Other task-based priors, *e.g.* the number of human body parts (Lorenz et al., 2019), can be leveraged to encourage certain properties for the content.

Gradient reversal layer. The Gradient Reversal Layer (GRL) was introduced in (Ganin and Lempitsky, 2015) for domain adaptation, where the gradient is reversed to prevent the model from predicting undesired results. GRL is effective in learning domain-specific style representations (Gonzalez-Garcia et al., 2018). Specifically, when using the style from one domain to generate images with style from another domains, the gradient is reversed to prevent this from happening.

Latent projection. Motivated by the findings of Rolinek et al. (2019), which suggest that VAE encoders cannot model the arbitrary rotations of the representation space, Zhao et al. (2021b) propose the projection of the latent space onto the direction with more information about a generating factor. Latent projection allows the information to be disentangled between particular orientations of the data.

Frequency Decomposition. Recent studies have investigated the use of frequency decomposition transformations to encourage CSD. For example, Liu et al. (2021a) use the fast Fourier transform to extract image amplitude and phase. Intuitively, the former reflects image style, whereas the latter corresponds to image content. Huang et al. (2021) use Discrete Cosine Transformation (DCT) to extract the domain invariant and domain specific frequency components, as an approximation of content and style factors, respectively.

Structured latent. A causal approach to representation learning solves the identifiability problem discussed in Section 2.5.2 by enforcing the latent space to be structured as a SCM. Structured latents create strong inductive biases because one might not only define the desired variables –which correspond to the generating factors– but also the relationship between them. This idea can be implemented in different settings by:

1. using conditional NFs in a VAE latent (Pawlowski et al., 2020);
2. decomposing of a VAE latent space into separate parts, where each component is further processed at different levels of the decoder (Leeb et al., 2020);
3. constraining the latent variable of a BiGAN (Dumoulin et al., 2016; Donahue et al., 2016) with Bayesian networks (Dash et al., 2020);
4. forcing the latent variables of a BiGAN-style architecture (Shen et al., 2020b) to follow a graph structure prior defined as an adjacency matrix (Shen et al., 2020a).

4.4. Learning setups for disentanglement

Popular learning setups can encourage disentanglement by harmonising the interaction between blocks.

Cycle-consistency. Cycle-consistency (Zhu et al., 2017a; Almahairi et al., 2018; Hiasa et al., 2018; Zhang et al., 2019) is a technique for regularising image translation settings. In particular, it can be useful for reinforcing correspondence between input and generated images (Xia et al., 2019b; 2020), or to improve stability and reconstruction fidelity in unsupervised and semi-supervised settings (Li et al., 2017).

Latent regression. There is a gentle balance to be made in the complexity of these blocks: too complex and with lots of parameter capacity may lead to information captured within their parameters that can lead to this information not captured in the latent variables. Latent regression has been employed to force the reconstructed image to contain information encoded into this representation (Huang et al., 2018). In particular, considering an input image X, the representation $\mathbf{z}$ and the reconstructed image X', we wish to extract a new latent representation $\mathbf{z}'$ from encoding X', which will be as similar as possible to $\mathbf{z}$. In other words, we need to minimise the distance between $\mathbf{z}$ and $\mathbf{z}'$ using a latent regression loss.

5. Metrics for disentanglement

To understand disentanglement and design models that improve it, we need to be able to quantify how disentangled is (are) the encoded representation(s). Below, we briefly report the most popular disentanglement metrics, splitting them into 2 categories: i) disentanglement of factors in a single vector latent variables, and ii) disentanglement between two latent variables of the same or different dimensionality.

Single vector-based latent variable. This category consists of both qualitative and quantitative methods for measuring how disentangled a representation is.

Qualitatively, we can evaluate disentanglement by traversing a single latent dimension that alters the reconstructed image by a single aspect (*e.g.* increase image intensity). In practice, these traversals are linear interpolations which are used to perform "walks" in non-linear data manifolds and to interpret the variation controlled by each factor (Jahanian et al., 2020; Cherepkov et al., 2021). Latent traversals do not require ground truth information about the factors. Duan et al. (2020) propose a way to quantify latent traversals in a post-hoc fashion, using the unsupervised disentanglement ranking metric to select the most disentangled version of the trained model. Quantitatively, there has been considerable effort to create metrics to evaluate vector representations. Since there are different proxies for disentanglement, popular metrics focus on measuring different aspects. For example, Higgins et al. (2017) propose the first metric to quantify disentanglement when the ground truth factors of a data set are available. In fact, they evaluate disentanglement using the prediction accuracy of a linear classifier that is trained as follows: they first choose a factor k and generate data with this factor fixed, but all others varying randomly. After obtaining the representations of the generated data, they take the absolute value of the pairwise differences of these representations. Then, the mean of these statistics across the pairs gives one training input for the classifier, and the fixed factor index k is the corresponding training output. Subsequently, Kim and Mnih (2018) adopt the metric of Higgins et al. (2017), but construct the training set of the linear classifier by considering the empirical variance of normalised representations rather than the pairwise differences. Chen et al. (2018) argue that given a factor of variation, the first two dimensions of the latent vector should have the highest MI. They measure the gap between these two dimensions using the introduced mutual information gap metric. Ridgeway and Mozer (2018) propose to measure the modularity of latent representations by measuring the MI between factors, ensuring that each vector dimension encodes at most one factor of variation. Eastwood and Williams (2018) first train an encoder on a synthetic dataset with predefined factors of variation z, and encode a representation c for each data sample. Then, they train a regressor to predict each factor z given a c representation. Based on the prediction accuracy, they measure the disentanglement, completeness, and informativeness of each representation. Finally, Kumar et al. (2018) propose the separated attribute predictability score to first compute the prediction errors of the two most predictive latent dimensions for each factor, and then use the average error difference as a disentanglement metric. A more comprehensive review of metrics for vector-based disentanglement can be found in Zaidi et al. (2020).

Two latent variables. The aforementioned metrics are not applicable in CSD as they rely on either having ground truth for the factors or assuming that the latent manifold is solely vector-based. To evaluate CSD one should consider more than one latent variable and a possible difference in dimensionality, *e.g.* spatial content (tensor) and vector style. To the best of our knowledge, the only work that focuses on CSD metrics is that of Liu et al. (2021c). In this work, the authors consider the properties of uncorrelation and informativeness, and propose to combine the empirical distance correlation (Székely et al., 2007) and a metric termed information over bias, to measure the degree of disentanglement *between* content and style representations. Two other methods for measuring the uncorrelation-independence between variables of different dimensionality are the kernel-target alignment (Cristianini et al., 2002) and the Hilbert-Schmidt independence criterion (Gretton et al., 2005). However, both methods require predefined kernels.

6. Applications of disentanglement in medical imaging

We now survey medical image analysis papers using disentangled representations in a diverse set of challenging tasks.

Strategy. We use the search expression "(disentanglement OR disentangled OR disentangle) AND medical AND (image OR imaging)" over the title and abstract of papers on Google Scholar. 132 papers were found (March 2022) and these were filtered by removing duplicates (with pre-prints) and papers which do not disentangle representations from neural networks.[4] After filtering, there were 68 papers utilising disentangled representation in medical imaging.

We categorise these papers based on the investigated application, as shown in the visual summary of Fig. 4. In Tab. 1, we summarise the survey by highlighting, for each paper, its applications, the general framework for disentanglement, the generating factors being disentangled, the organ over which the method is applied, the imaging modality and we verify if code is available or not.[5] In the following, for each application, we briefly survey all the relevant papers. We pick one exemplar per application to describe the architecture, training setup, and tips and tricks in their implementation. The choice of the exemplar models is based on how popular the model is, if the code is publicly available, if the model is representative or the first work integrating disentanglement in specific applications and if the model has been extensively evaluated with public datasets.

[4] We excluded works that used the word "disentanglement" in context unrelated to representation learning.

[5] All URLs of official implementations for the highlighted methods can be found in the associated repository https://github.com/vios-s/disentanglement_tutorial.

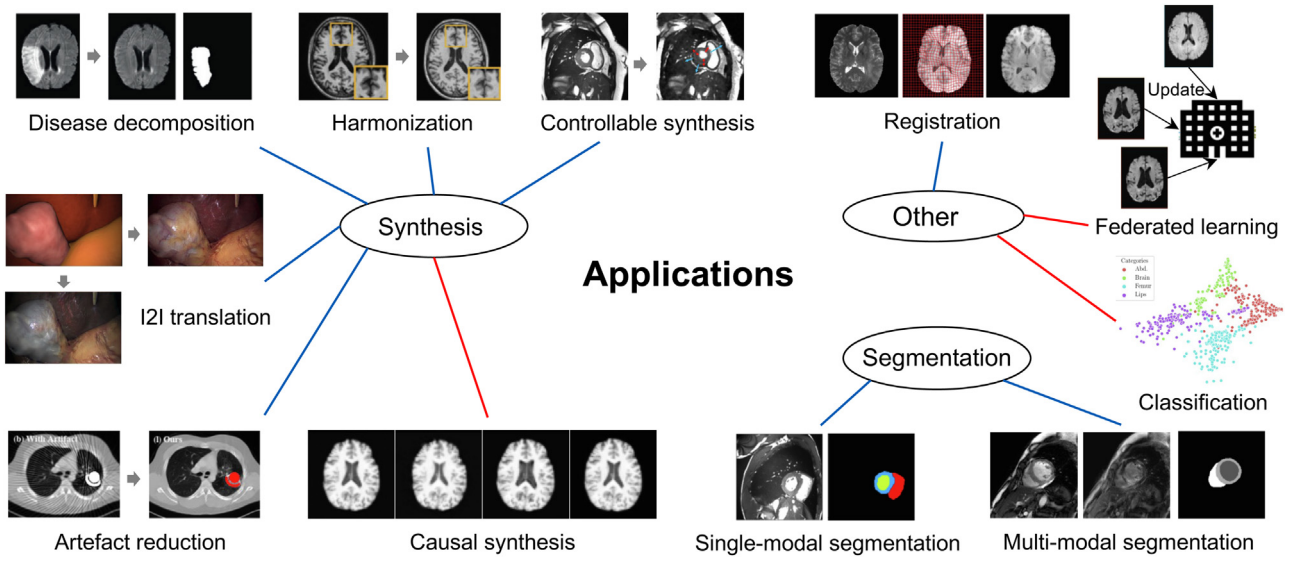

Fig. 4. Visual summary of disentangled representation learning applications in medical imaging. The visual examples are taken from the exemplar models marked with * in Tab. 1, cited and detailed in Section 6. Red connections indicate vector-based disentanglement, while blue connections indicate tensor/vector-based one (CSD). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.1. Synthesis

Many medical imaging procedures are expensive to perform, invasive, and uncomfortable for the patient. For this reason, datasets from certain modalities can be small and imbalanced. Besides, some images are impossible to acquire: doctors might wish to have an image of patient when the patient was healthy in order to perform a comparative diagnosis (these hypothetical image estimations are also called counterfactuals). Otherwise, a training dataset might be acquired in one hospital but deployed in another hospital. To address these problems, medical image synthesis is considered for augmenting and balancing these datasets. We will now review a few works which utilise disentanglement for synthesising medical images in other to mitigate these issues in different tasks.

6.1.1. Disease decomposition

Disease decomposition (Xia et al., 2019a; 2020; Kobayashi et al., 2021; Tang et al., 2021; Couronné et al., 2021) aims at disentangling *normal* from *abnormal* factors in an image. Kobayashi et al. (2021) use an architecture inspired on a VAE with vector quantised latent (Section 4.1) for disentangling brain tumor from healthy brain information. Tang et al. (2021) disentangles normal from abnormal features in lung X-ray using a MUNIT-like (Huang et al., 2018) architecture. Another view is to use temporal information to disentangle disease progression information from changes due to phenotypic differences across subjects (Couronné et al., 2021).

We select an image synthesis work of Xia et al. (2020) because it is extensively validated in public datasets. Xia et al. (2020) explore disentanglement in the context of synthesising pseudo-healthy brain MRI from patients with tumors or ischemic stroke lesions. The pathology information is disentangled from anatomical features as a segmentation mask.

Architecture. The model consists of a generator (G) extracting a healthy image from a pathological one, and a segmentor (S) predicting the remaining pathological information as a mask. Additionally, a decoder module R, responsible for reconstructing the input, receives and combines the extracted healthy image and mask enforcing consistency between input and reconstruction. Each module consists of a U-net like architecture with residual blocks and sigmoid output activation, whilst two discriminators follow for the healthy images and the masks.

Training setup. Two discriminators enforce the (generated) healthy images and the masks to be realistic ($\mathcal{L}_{GAN_1}$ and $\mathcal{L}_{GAN_2}$). When ground truth masks are available, a DICE score is used to learn the segmentations ($\mathcal{L}_{seg}$). Otherwise, adversarial training allows semi-supervised learning of the segmentation with unpaired masks. A cycle-consistency loss ($\mathcal{L}_{CC_1}$) ensures that the subjects keep the "identity". The goal is to solely change the pathological aspect while preserving patient identity. Therefore, an additional ($\mathcal{L}_{CC_2}$) cycle consistency objective is introduced to prevent the generator from making unnecessary changes (*e.g.* generation of pathologies for healthy images). The final loss is a combination of the above losses:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{GAN_1} + \lambda_2 \mathcal{L}_{GAN_2} + \lambda_3 \mathcal{L}_{CC_1} + \lambda_4 \mathcal{L}_{CC_2} + \lambda_5 \mathcal{L}_{seg} \quad (4)$$

The model is evaluated on the following datasets: ISLES (Maier et al., 2017), BraTS, and Cam-CAN (Taylor et al., 2017) brain datasets.

Tips & Tricks. The main bias introduced in Xia et al. (2020) is the use of an auxiliary network for guiding synthesis in a cycle consistency setting. In addition, they observed that, in practice, using a Wasserstein loss coupled with gradient penalty (Gulrajani et al., 2017) is more beneficial compared to the Least Squares discriminator loss (Mao et al., 2017).

6.1.2. Image-to-image translation

Translating one image representation into another, which differs in a specific factor (*e.g.* style) but maintains others, is termed I2I translation. Translation can be useful in medical imaging when, for instance, one modality is very costly, invasive or even harmful to acquire. In this case, one might choose to acquire an image with a cheaper and/or safer method (with similar content) and subsequently translate the image to the desired domain. MUNIT (Huang et al., 2018) is the first to incorporate a CSD paradigm into I2I translation and it has become a widely adopted benchmark. Several others have taken the base architecture from Huang et al. (2018) and extended it several tasks in medical imaging (Li et al., 2019a; Pfeiffer et al., 2019). Li et al. (2019a) translate Fluorescein Fundus (FF) images, which are non-invasive and safe, into Fluorescein Fundus Angiography (FFA), which is the preferred

Table 1
Summary of the surveyed medical imaging applications. * denotes the examplar models that we detail in Section 6.

Application		Model	Framework	Factors	Organ	Modality	Code
Synthesis	Disease decomposition	Xia et al. (2020) *	GAN	Normal, Abnormal	Brain	MRI	√
		Kobayashi et al. (2021)	VAE + GAN	Normal, Abnormal	Brain	MRI	√
		(Couronné et al., 2021)	VAE	Disease Progression Inter-patient variability	Brain	MRI	√
		Tang et al. (2021)	VAE + GAN	Normal, Abnormal	Lungs	X-ray	✗
	I2I translation	Li et al. (2019a)	CSD	Content, Style	Eye	Fluorescein Fundus	✗
		Li et al. (2019b)	GAN	Lung, Bones, Other	Chest	X-ray, CT	√
		Pfeiffer et al. (2019) *	CSD	Content, Style	Liver, Abdomen	Laparoscopic Images	√
		Fei et al. (2021)	CSD	Content, Style	Brain	MRI	✗
	Artefact reduction	Liao et al. (2020) *	GAN	Content, Artifact	Many	CBCT, CT	√
		Huang et al. (2020)	GAN	Content, Noise	Eye	OCT	✗
		Niu et al. (2021)	GAN	Content, Artifact	Many l	CT	✗
		Tang et al. (2022)	GAN	Content, Artifact	Brain	MRI	✗
	Harmonisation	Wang et al. (2021c)	VAE + NF	Sex, Age, Site, Image	Brain	MRI	✗
		Dewey et al. (2020)	CSD	Content, Style	Brain	MRI	✗
		Li et al. (2021)	CSD	Content, Style	Brain	MRI	✗
		Zuo et al. (2021b) *	CSD	Content, Style	Brain	MRI	√
		Zuo et al. (2021a)	CSD	Content, Style	Brain	MRI	√
	Controllable	Liu et al. (2018)	CSD	Segmentation Meshes, Residue	Lung	CT	✗
		Liu et al. (2020b)	GAN	Content, Style	Chest	MRI, CT	√
		Kelkar and Anastasio (2022)	GAN	Content, Style	Brain	MRI	✗
		Thermos et al. (2021) *	CSD + GAN	Content, Style	Heart	MRI	√
		Hochberg et al. (2021)	GAN	Hierarchical Style	Liver	CT	✗
		Havaei et al. (2021)	CSD + GAN	Content, Style	Skin, Lung	Cermatoscopic Images, CT	✗
	Causal	Pawlowski et al. (2020) *	VAE + NF	Sex, Age, Brain Volume, Image	Brain	MRI	√
		Reinhold et al. (2021)	VAE + NF	Lesion, Ventricle and Brain volume, Slice, Image, Disability Score, Sex	Brain	MRI	√
Segmentation	Single-modal	Chartsias et al. (2019) *	CSD	Content, Style	Heart, Abdomen, Spinal Cord	MRI, CT	√
		Valvano et al. (2019)	CSD	Content, Style, Temporal	Heart	MRI	√
		Jiang et al. (2020)	CSD	Content, Pathology, Modality	Heart	MRI, LGE	√
		Liu et al. (2020c)	CSD	Content, Style	Heart	MRI	√
		Shin et al. (2021)	Auto-Encoder	Intensity, Non-intensity	Abdomen	CT	√
		Liu et al. (2021b)	CSD	Content, Style, Domain	Heart, Spinal Cord	MRI	√
		Memmel et al. (2021)	VAE + GAN	Content, Domain	Brain	MRI	√
		Kalkhof et al. (2022)	CSD	Content, Domain	Bowel	CT	✗
	Multi-modal	Yang et al. (2019)	CSD	Content, Style	Liver	MRI, CT	✗
		Chen et al. (2019a)	CSD	Content, Style	Brain	MRI	✗
		Chen et al. (2019b) *	CSD	Content, Style	Heart	MRI, LGE	✗
		Jiang and Veeraraghavan (2020)	CSD	Content, Style, Domain	Abdomen	MRI, CT	✗
		Lyu et al. (2020)	CSD + GAN	Content, Style	Spine	CBCT, CT	✗
		Yang et al. (2020)	CSD	Content, Style	Liver	MRI, CT	✗

(continued on next page)

Table 1 (*continued*)

Application	Model	Framework	Factors	Organ	Modality	Code
	Xie et al. (2020)	Auto-Encoder, GAN	Content	Eye, Colon	Colonoscopic Images, Fundus Images	✗
	Wang and Zheng (2021)	CSD + GAN	Content, Style	Heart	MRI	✗
	Ouyang et al. (2021)	CSD	Content, Style	Brain	MRI	√
	Chartsias et al. (2021)	CSD	Content, Style	Heart, Abdomen	MRI, LGE	√
	Pei et al. (2021)	CSD	Content, Style	Heart	MRI, CT	√
	Chen et al. (2021a)	CSD	Content, Style	Heart	MRI, CT	√
	Chen et al. (2021b)	CSD	Content, Style	Brain, Heart	MRI, CT	✗
	Ning et al. (2021)	CSD	Content, Style, Domain	Heart, Abdomen	MRI, CT	✗
	Wang and Zheng (2022)	CSD	Content, Style	Heart	MRI, CT	✗
	Liu et al. (2022)	CSD	Content, Style	Nuclei	Microscopy Images	✗
Classification	Ben-Cohen et al. (2019)	Classifier + GAN	Class Specific, Agnostic	Liver	CT	✗
	Gyawali et al. (2019)	VAE	Class Specific, Agnostic	Chest	X-ray	√
	Zhao et al. (2019)	VAE + Regressor	Age-specific, Agnostic	Brain	MRI	√
	Puyol-Antón et al. (2020)	VAE + Classifier	Different Biomarkers	Heart	MRI	√
	Meng et al. (2021) *	Classifier + Clustering	Categorical, Domain	Fetal	Fetal Ultrasound	✗
	Bass et al. (2020, 2021)	CSD	Task Specific, Agnostic	Brain	MRI	√
	Zou et al. (2021)	VAE	Sex, Rest	Hip	CT	✗
	Zhao et al. (2021a)	VAE + Regressor	Age-specific, Site-specific, Agnostic	Bowel	Endoscopic Images	✗
	Harada et al. (2021)	GAN	Location-dependent, Ulcerative Colitis-dependent	Brain	MRI	√
	Wang et al. (2021a)	Auto-Encoder	Disease-related, Disease-irrelevant	Breast	X-ray	✗
	Gravina et al. (2021)	Auto-Encoder + GAN	Shading, Albedo, Deformation Field	Breast	MRI	✗
	Yang et al. (2021b)	VAE	Time-invariant, Time-variant	Brain	MRI	✗
	Zhou et al. (2022)	CSD + VAE	Structure, Texture	Lung	Chest Radiograph	√
Registration	Qin et al. (2019) *	CSD + GAN	Content, Style	Lung, Brain	MRI, CT	✗
	Chartsias et al. (2021)	CSD	Content, Style	Heart, Abdomen	MRI, LGE	√
	Maillard et al. (2022)	CSD	Content, Style	Brain	MRI	✗
Federated learning	Li et al. (2020)	CSD + GAN	Content, Style	Heart	CBCT, CT	√
	Bercea et al. (2021b) *	Auto-Encoder	Content, Style	Brain	MRI	✗

modality for diagnosis but it is invasive and has potential side effects. Pfeiffer et al. (2019) use an I2I architecture for translating synthetic images from a simulator into realistic laparoscopic images for data augmentation. Other works (Li et al., 2019b) use a I2I network to disentangle different anatomy factors such lungs and bones in Chest X-ray. Fei et al. (2021) disentangle content from a modality for synthesising brain MR images of different sequences.

Due to the foundational nature of MUNIT we proceed to detail their learning biases next.

Architecture. The basic assumption is that multi-domain images share common content information, but differ in style. A content encoder maps images to multi-channel feature maps, by removing style with IN layers (Ulyanov et al., 2017). A second encoder extracts global style information with fully connected layers and average pooling. Finally, style and content representations are combined in the decoder through AdaIN modules (Huang and Belongie, 2017). Disentanglement is further encouraged by a bidirectional reconstruction loss (Zhu et al., 2017b) that enables style transfer. In order to learn a smooth representation manifold, two latent regression losses are applied on content and style extracted from input images, namely a content-based latent regression loss that penalises the distance to the content extracted from reconstructed images, and a style-based latent regression that encourages the encoded style distributions to match their Gaussian prior. Finally, adversarial learning encourages realistic synthetic images.

Training setup. MUNIT achieves unsupervised multi-modal I2I translation by minimising the following loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{GAN} + \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{c-rec} + \lambda_3 \mathcal{L}_{s-rec}, \quad (5)$$

where $\mathcal{L}_{rec}$ is the image reconstruction loss, $\mathcal{L}_{c-rec}$ and $\mathcal{L}_{s-rec}$ denote the content and style reconstruction losses, and $\lambda_1 = 10$, $\lambda_2 = 1$ are the hyperparameters used by the authors. The model has been evaluated on SYNTHIA (Ros et al., 2016), Cityscapes (Cordts et al., 2016), edge-to-shoes (Zhu et al., 2016), and summer-to-winter (Zhu et al., 2017a) datasets. In the medical imaging, (Li et al., 2019a) applied a similar strategy to translation from Fluorescein Fundus (FF) towards Fluorescein Fundus Angiography (FFA) images of the eye with the Isfahan MISP (Alipour et al., 2012) dataset. Pfeiffer et al. (2019) applied a variation of MUNIT for synthesising realistic laparoscopic images from synthetic images. With multi-modal images of the same patient available, Fei et al. (2021) achieve better synthesis performance by fusing the shared content information of multi-modal images and creating the pseudo-target modality image as the pseudo-supervision of the synthesised image.

Tips & Tricks. MUNIT has shown robust and impressive performance on multiple I2I scenarios. The style representation is sampled directly from $N(\mathbf{0}, \mathbf{1})$, which means the style latent space is smoother and better for style traversal compared to the style learned by minimising KL divergence using the re-parameterisation trick (Kingma and Welling, 2014). Assuming a semantic content prior, the AdaIN layers in the decoder can be replaced with the SPADE module to achieve more controllable translation. The provided hyperparameters can be used for most datasets without the need for intensive tuning. The major drawback of MUNIT is the vague definition of content *i.e.* the domain-invariant representation, which is achieved by the bi-directional reconstruction. The content is not interpretable, and it is not trivial to measure how domain-invariant it is.

6.1.3. Artefact reduction

The presence of artefacts, noise or speckles is common in medical images due to challenges in acquisition. Metal artefacts appear in computed tomography acquisitions, for instance, when a patient carries metallic implants. One might be able to alleviate this issue by disentangling content space from artefact space (Liao et al., 2019; 2020; Niu et al., 2021; Huang et al., 2020; Tang et al., 2022), similar to I2I translation in Section 6.1.2. Huang et al. (2020), for instance, use a similar architecture for speckle noise reduction in optical coherence tomography (OCT) images. Tang et al. (2022) disentangle noise information from image information with an attention based module for image restoration. The artefact disentanglement network (ADN) (Liao et al., 2019; 2020), reduces the artefacts by disentangling content from artefacts in the latent space utilising unpaired data (*i.e.* unsupervised). Niu et al. (2021) extends ADN by further reinforcing disentanglement using regularisation in a lower dimensional manifold with ideas from differential geometry.

We now discuss ADN because their method is extensively validated on publicly available datasets.

Architecture. The architecture in ADN (Liao et al., 2020) contains two groups of encoders and decoders, one for each domain of images with ($\mathcal{I}_a$) and without ($\mathcal{I}$) artefacts. For domain $\mathcal{I}$, an encoder $E_c^{\mathcal{I}}$ outputs a content latent representation z_c with the entire image information and a decoder $D^{\mathcal{I}}$ reconstructs the artefact-free image. For domain $\mathcal{I}_a$, two encoders $E_c^{\mathcal{I}_a}$ and $E_a^{\mathcal{I}_a}$ split the images into two disentangled representations z_c and z_a, and a decoder $D^{\mathcal{I}_a}$ reconstructs the image with artefacts. The goal is to learn a transformation from $\mathcal{I}_a \rightarrow \mathcal{I}$ by using $D^{\mathcal{I}} \circ E_c^{\mathcal{I}_a}$ for a given image with artefacts. In addition, one discriminator for each domain is present for reinforcing realism in the reconstructed images.

Training setup. During training, unpaired images are used as inputs such that the images with artefacts are split into z_c and z_a and the images without artefacts are encoded into z_c. By using the decoders to reconstruct versions of the input image both with and without artefacts, they train the neural networks with the following losses:

$$\mathcal{L}_{total} = \lambda_1(\mathcal{L}_{adv}^{\mathcal{I}_a} + \mathcal{L}_{adv}^{\mathcal{I}_a}) + \lambda_2\mathcal{L}_{art} + \lambda_3\mathcal{L}_{rec} + \lambda_4\mathcal{L}_{self}, \quad (6)$$

where $\mathcal{L}_{adv}$ are adversarial losses, $\mathcal{L}_{rec}$ and $\mathcal{L}_{self}$ are image reconstruction losses computed using cycle consistency, and $\mathcal{L}_{art}$ is a loss that forces the reconstructed image to be anatomically precise.

Tips & Tricks. The artefact decoder takes two latent spaces as inputs (content + artefact representations). (Liao et al., 2020) merge these latent spaces using a variation of the feature pyramid network (FPN) (Lin et al., 2017). The artefact representations are concatenated at several levels of the artefact encoder with the decoder, using 1x1 convolutional layers to locally merge the features.

6.1.4. Harmonisation

Harmonisation refers to an I2I translation process in medical imaging that aims to reduce domain shifts between acquisitions and improve generalisation (Bashyam et al., 2020). Some works Dewey et al. (2020), Zuo et al. (2021a), Zuo et al. (2021b), Wang et al. (2021c) focus on harmonising magnetic resonance (MR) imaging from different sites. It might even be useful to homogenise acquisitions between different scanners from the same manufacturer (Li et al., 2021). MRI measurements tend to have inconsistencies between images from different sites due to differences in acquisition protocols. Dewey et al. (2020), Zuo et al. (2021a) disentangle MRI information into content and style information such that input images can be translated to a different style, simulating a different acquisition protocol. Wang et al. (2021c) use NF for modelling disentangled domain changes in the latent space of a VAE.

Next we detail the biases of the method in Zuo et al. (2021b) because it was extensively validated in publicly available datasets and show superior performance compared to previous methods.

Architecture. The architecture comprises style E_s and content E_c encoders, and a decoder D that are used for both domains. The harmonisation part of their models utilises 2D slices from the 3D images. This enables to have an intra-patient prior about the style/contrast information: the style should not change within a volume. They also leverage the fact that two MRI sequences (T1 and T2 weighted) are available for each patient, so part of the training can be done in a supervised manner. They also include a discriminator D_c in the content latent space z_c to ensure that the latent is not informative about the style, as opposed to previous methods that used discriminators in image space. The style latent z_s has a probabilistic representation with mean and variance, similar to the VAE in Section 3.1.

Training setup. During training, content is disentangled from style in two dimensions: MRI sequence and site. Two images (T1 and T2 sequences) from two different patients (site A and site B) are fed into the network. Therefore, a total of 4 images are used as input. Interestingly, the style encoder E_s and the content encoder E_c have different inputs that belong to the same image. This forces the style latent space z_s to contain only information about the style and not any structural information. Naturally, this setting assumes that different slices in the same image have the same style. The final loss function is

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{adv} + \lambda_2\mathcal{L}_{z_c} + \lambda_3\mathcal{L}_{rec} + \lambda_4\mathcal{L}_{percep} + \lambda_5\mathcal{L}_{KL}, \quad (7)$$

where $\mathcal{L}_{adv}$ forces z_c to not contain information about the style, $\mathcal{L}_{z_c}$ encourages the z_c between representations of the same patient but different MRI sequences to be similar, $\mathcal{L}_{rec}$ and $\mathcal{L}_{percep}$, which is a perceptual loss (Johnson et al., 2016), forces the reconstruction to be the same as the input images.

Tips & Tricks. Two other tricks are used for avoiding style information in the content space: the Gumbel-softmax reparameteri-

sation in z_c and random swapping of channels in z_c between latent spaces of the same patient but different MRI sequences.

6.1.5. Controllable synthesis

Acquiring annotated data at scale with rare diseases or conditions remains a challenge. It would be extremely useful to have a method that controllably synthesises (Liu et al., 2018; Thermos et al., 2021; Liu et al., 2020b; Hochberg et al., 2021; Kelkar and Anastasio, 2022; Havaei et al., 2021) images that can correct such underrepresentation. Hochberg et al. (2021) uses StyleGAN (Karras et al., 2019) with an encoder for controlling style of the synthesised image. Liu et al. (2020b) and Kelkar and Anastasio (2022) inject the style from different modalities into the decoder to translate the original image into other styles based on StyleGAN (Karras et al., 2020). Havaei et al. (2021) disentangles content from style using conditional GANs and dual adversarial inference (Lao et al., 2019). Thermos et al. (2021) proposed DAAGAN to use the concept of anatomy arithmetic for such controllable generation.

We now discuss in detail DAAGAN due to being first of explicitly doing arithmetic in tensor spaces.

Architecture. DAAGAN uses a pre-trained SDNet (Chartsias et al., 2019) to disentangle the images into anatomy and modality representations. It contains a generator, a pathology classifier, and a discriminator. After extracting the spatial anatomy representations with the pre-trained SDNet, DAAGAN performs the arithmetic operation (mixing and swapping of selected channels) on the anatomy representations of different images (labeled as different pathology or health). The generator takes the mixed novel anatomy factor as input and use AdaIN to combine the anatomy and modality factors to synthesise the corresponding image. In particular, DAAGAN introduced a localised noise injection module in the generator to avoid abrupt mixing of the anatomy channels. The pathology classifier is pre-trained and used to guide the generator to synthesise images with desired pathology. DAAGAN has been evaluated on two cardiac datasets including ACDC (Bernard et al., 2018) and M&Ms (Campello et al., 2021).

Training setup. Apart from the pathology classification loss and the adversarial loss for the generator and discriminator, DAAGAN introduced two consistency losses to encourage the anatomical factors which are not related to the heart to remain unaltered after the arithmetic and noise injection steps. $\mathcal{L}_{cons}$ measures the difference of the background area of the anatomy factor before and after the noise injection and $\mathcal{L}_{bg}$ measures the difference of the background area of original image and synthesised image. To train DAAGAN, the total loss is:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{adv} + \lambda_2 \mathcal{L}_{path} + \lambda_3 \mathcal{L}_{cons} + \lambda_4 \mathcal{L}_{bg}. \quad (8)$$

Tips & Tricks. Since DAAGAN uses SDNet as the extractor of disentangled representations, it is possible to apply DAAGAN to other datasets on which SDNet works well such as CHAOs Kavur et al. (2021) for abdomen, and SCGM (Prados et al., 2017) for gray matter and spinal cord. The noise injection module in DAAGAN acts as a mixing corrector to modify the mixed anatomy factors such that the non-suitable mixing of channels from different anatomy factors can be corrected, which has a limitation on the controllability of mixing. Finally, to mix anatomy factors from different images, it is required to first register the anatomy of each image.

6.1.6. Causal synthesis

Causal image synthesis (Pawlowski et al., 2020; Reinhold et al., 2021) is a special case of conditional generation in which the conditioning architecture follows a structural causal model (SCM). SCMs consist of graphs where the nodes are generating factors and the edges are causal relationships (Peters et al., 2017). In fact, the edges represent physical mechanisms of the real world. The representation of each causal variable needs to be disentangled and their relation should be specified by the designer. This is a stronger form of bias than disentangling variables only. With causal models, one might also answer *counterfactual* queries, such as "What would have happened to an individual if variable '$\mathcal{S}_i$' had been different?". This can be seen as an intervention at individual level.

CausalGAN (Kocaoglu et al., 2018) initially introduced generative models following a causal structure, however, they were not capable of estimating counterfactuals. Pawlowski et al. (2020) design a causal model capable of performing counterfactuals with imaging data using normalising flows (Rezende and Mohamed, 2015). A causal graph for a brain MRI problem is constructed where the brain ventricular volume depends on the age, but not on the patient's sex. Reinhold et al. (2021) extend Pawlowski et al. (2020) to higher resolution images (Dolatabadi et al., 2020) and to a more complex SCM of multiple sclerosis disease. Wang et al. (2021c) use a similar setup with VAE and NFs for creating a causal model which takes into account image acquisition site information for image harmonisation.

Next we discuss in more detail the method in Pawlowski et al. (2020) due to be the first work using generative causal models in medical imaging.

Architecture. Pawlowski et al. (2020) rely on NFs (Rezende and Mohamed, 2015) (invertible neural models (Sec. 3.3) for modeling attributes such as age, sex, brain and ventricular volume and their relationships; and conditional VAEs for synthesising imaging counterfactuals. The conditionals follow a structure based on clinical knowledge about the problem. The authors enable counterfactual estimation by using invertible models. This allows the prediction of a latent representation of an observation and subsequent local intervention by changing the desired latent space.

Training setup. The networks associated with the NFs and VAEs are trained jointly with backpropagation using the ELBO as a loss function. The model has been evaluated on brain MRI scans from the UK Biobank (Sudlow et al., 2015).

Tips & Tricks. The main method used for reinforcing the structural biases is NF, which are based on neural spine flows (Durkan et al., 2019). Additionally, the authors realised that normalisation as a pre-processing step is necessary, as it prevents dependencies being learned on the variable with the largest magnitude, and helps with combining the scalar attributes with imaging information. The implementation was done using the Pyro library (Bingham et al., 2019) which can be useful for several probabilistic programming tasks.

6.2. Segmentation

The goal of deep learning based segmentation is to train a model to accurately predict the pixel-wise labels (segmentation mask) from an image input. Disentangled representation can help by separating out all the information necessary for segmentation (such as the content or shape in an image) from other information such as style.

6.2.1. Single-modal

Regarding single-modal medical image segmentation, the input images are acquired with only one modality, *e.g.* MRI images. Spatial Decomposition Network (SDNet) (Chartsias et al., 2019), decomposes 2D medical images into spatial anatomical factors (content) and non-spatial modality factors (style). When temporal information is available, temporal consistency objectives can be applied to boost the performance as in (Valvano et al., 2019). Based on SDNet, Jiang et al. (2020) additionally disentangle the pathology factor to perform semi-supervised pathology segmentation. Disentanglement methods in segmentation also provides the possi-

bility to handle the domain shifts across different domains Additionally, the variational encoding of the style representation allows for sampling and interpolation of the appearance factors, enabling the synthesis of new plausible images (Liu et al., 2020c). To learn generalisable representations, gradient-based meta-learning can be applied as a leaning strategy when giving multi-domain data (Liu et al., 2021b). Shin et al. (2021) disentangle intensity and non-intensity for domain adaptation in CT images. Kalkhof et al. (2022) also disentangles content from style information using a conditional GAN for cross-domain segmentation.

We now detail SDNet as it has been widely used and extended to many medical tasks.

Architecture. SDNet uses two different encoders for factorising content into a spatial representation and style into a vector one. A decoder is responsible for reconstructing the input by combining the two latent variables, while a segmentation module is applied on the content latent space to learn to predict the segmentation mask for each cardiac part. SDNet learns the content which is represented as multi-channel binary maps of the same resolution as the input. This is obtained with a softmax and a thresholding function. To encourage the style encoder to encode only style-related information, the authors employ a VAE network. Then, style and content are combined to reconstruct the input image by applying a series of convolutional layers with FiLM layers (Perez et al., 2018).

Training setup. SDNet is trained by minimising the following loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{KL} + \lambda_2 \mathcal{L}_{seg} + \lambda_3 \mathcal{L}_{rec} + \lambda_4 \mathcal{L}_{z_{rec}}, \tag{9}$$

where $\mathcal{L}_{KL}$ is the KL Divergence measured between the sampled and the predicted style vectors, $\mathcal{L}_{rec}$ is the image reconstruction loss, $\mathcal{L}_{seg}$ is the segmentation loss, and $\mathcal{L}_{z_{rec}}$ is the latent regression loss between the sampled and the re-encoded style vectors. $\lambda_1 = 0.01, \lambda_2 = 10, \lambda_3 = 1,$ and $\lambda_4 = 1$ are the hyperparameters used by the authors. SDNet has been extensively evaluated on the ACDC (Bernard et al., 2018), MM-WHS (Zhuang and Shen, 2016a; Zhuang, 2013; Zhuang et al., 2010), CHAOs Kavur et al. (2021), and M&Ms (Campello et al., 2021) cardiac datasets, as well as on the SCGM (Prados et al., 2017) spinal one.

Tips & Tricks. SDNet encodes highly semantic content representation, which shows the advantage of content interpretability. Other modifications include using a SPADE module to replace FiLM, which has led to a performance improvement in several studies (Liu et al., 2021c). Gumbel-Softmax (Jang et al., 2016) can replace the naive softmax (Thermos et al., 2021) and binary thresholding.

6.2.2. Multi-modal and cross-modal

For multi-modal or cross-modal medical image segmentation, at least two modalities are required (*e.g.* CT and MRI scans). The goal is to accurately predict the segmentation mask given a specific patient, exploiting both (all) available modalities. The most popular models for this task are the Multimodal Unsupervised Image-to-image Translation (MUNIT) (Huang et al., 2018) and the concurrent and similar work of MUNIT, DRIT (Lee et al., 2018). Apart from MUNIT or DRIT, other works include the use of CT data to improve segmentation performance on cone beam computed tomography scans (Lyu et al., 2020) evaluated on CBCT and CT data (Glocker et al., 2013). Similarly, Wang and Zheng (2021) used CSD for segmentation in a cross-modality setting. SDNet has been also extended to multi-modal setting with the exploitation of aligned Cine and LGE data (Chartsias et al., 2021). Xie et al. (2020), Chen et al. (2021a) propose cycle-consistency-based GANs to generate better cross-modal images for segmentation by applying mutual information constraints to preserve the image-object information in the content features.

Next we detail how the MUNIT-based models are designed for multi-modal and cross-model medical image segmentation due to the popularity of MUNIT. We refer the readers to Section 6.1.2 for the details about the architecture and training setup of MUNIT.

Architecture. In multi-modal or cross-modal medical image segmentation, there are two ways of using MUNIT. The first strategy is aligning the content spaces of data from different modalities *e.g.* CT and MRI (Yang et al., 2020; 2019; Pei et al., 2021; Chen et al., 2019a; Ouyang et al., 2021; Ning et al., 2021). Then, a segmentation network is used to predict the mask with content representations as input. Alternatively, the scenarios of imbalanced domains and domain adaptation is considered, where there is a domain with more data and annotations (*e.g.* MRI) and a domain with less data and few or no annotations (*e.g.* CT) (Chen et al., 2019b; Jiang and Veeraraghavan, 2020; Jiang et al., 2022; Liu et al., 2022; Chen et al., 2021b; Wang and Zheng, 2022). After training a MUNIT model on the two domains, the mappings between them can be obtained. During inference, the samples from CT domain are initially translated to MRI scans, which are then used to predict the segmentation mask by a segmentor trained on the MRI domain.

Training setup. MUNIT-based models have been evaluated on the following medical datasets: LiTS (Christ et al., 2017) liver, NCANDA (Zhao et al., 2021c) and BraTS (Menze et al., 2015) brain, CHAOs (Kavur et al., 2021) abdomen organs, and MS-CMRSeg (Zhuang, 2016; 2019) and MM-WHS (Zhuang and Shen, 2016a; Zhuang, 2013; Zhuang et al., 2010) cardiac datasets, respectively.

Tips & Tricks. First, following the typical MUNIT training setup presented in Section 6.1.2, a consistency loss can be further applied as a regulariser on the task output. For example, the predicted segmentation masks of the original MRI scan and the corresponding translated CT-style images must be consistent (Hoffman et al., 2018). Further, the anatomical (content) latent variables of the different modalities can be aligned or fused by applying adversarial training (Jiang and Veeraraghavan, 2020) and prior constraints as in Chartsias et al. (2021), Chen et al. (2019a), Ouyang et al. (2021). Compared to SDNet, MUNIT's disentangled content is less interpretable. MUNIT needs to be trained using a bi-directional setup. This means that it cannot be used for single-modal datasets.

6.3. Classification

A classification task and domain knowledge can be used to disentangle both the task-specific representation $\mathbf{z}_c$ from the classifier and a task-agnostic representation $\mathbf{z}_a$ (Ben-Cohen et al., 2019; Meng et al., 2019; 2021; Gyawali et al., 2019; Zhao et al., 2019; Berenguer et al., 2020; Harada et al., 2021; Zhou et al., 2021; Yang et al., 2021b; Zhou et al., 2022). By merging and decoding the representations, the image can be reconstructed. Berenguer et al. (2020) train a conditional VAE to pretrain an encoder for the subsequent diagnosis classification task. Zhou et al. (2021, 2022) disentangle structure and texture on chest X-ray images and show that a pre-trained texture encoder can be efficiently fine-tuned for COVID-19 outcome prediction. Zhao et al. (2021b) learn a representation in which a projection is disentangled and Jung et al. (2020) use capsules, both resulting in better representations for a downstream classification task. Zhao et al. (2021a) extend Zhao et al. (2021b) to a disentangled direction for different MRI sequences. Yang et al. (2021a) disentangle time-variant and time-invariant information in longitudinal studies for improving classification based on time-invariant representation. Wang et al. (2021a) use a graph convolutional AE to disentangle disease-specific and disease-invariant features for improving disease prediction. Harada et al. (2021) disentangle the location-dependent and ulcerative colitis (UC)-dependent representations with the classification losses to achieve semi-supervised

learning method for UC classifcation. Bass et al. (2021) detect salient features for classification and regression by explicitly disentangling task-specific and task-agnostic information using ICAM (Bass et al., 2020). Cheng et al. (2021) use disentanglement for clustering patients with characteristic phenotypes in order to understand disease progression. Zou et al. (2021) uses a VAE over meshes for disentangling sex information from hip bones. Puyol-Antón et al. (2020) uses a VAE for disentangling different biomarkers from segmentation masks which are connected to a classifier for interpretability.

We use as exemplar the Mutual Information-based Disentangled Neural Networks (MIDNet) (Meng et al., 2021), which was initially developed for ultrasound fetal imaging. Whilst building on earlier work (Meng et al., 2019), this approach leverages components that can be easily adapted to other applications and offers a multi-task framework to disentangled task-specific representations. Note that the main goal of disentanglement is to find representations that are invariant to different tasks and domains.

Architecture. The neural network is composed by two encoders E_c and E_a, a classifier C that takes $\mathbf{z}_c$ as input and output the desired class, and a decoder to reconstruct the images by re-entangling the latent spaces $\mathbf{z}_c$ and $\mathbf{z}_a$.

Training setup. In addition to classification $\mathcal{L}_{cls}$ and reconstruction $\mathcal{L}_{rec}$ losses $\mathbf{z}_c$ and $\mathbf{z}_a$ are disentangled using Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018) ($\mathcal{L}_{MI}$). Domain invariance in $\mathbf{z}_c$ is further reinforced via a clustering loss $\mathcal{L}_{clus}$ that encourages samples from different domains, but with the same label, to have similar task-specific representations. Finally, the network is trained in a semi-supervised way ($\mathcal{L}_{SSL}$) using the MixMatch method (Berthelot et al., 2019) and an alignment loss for improving generalisation in a domain adaptation setting. The total loss is:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{cls} + \lambda_3\mathcal{L}_{MI} + \lambda_4\mathcal{L}_{clus} + \lambda_5\mathcal{L}_{ssl}. \quad (10)$$

The method is evaluated using fetal ultrasound datasets from the iFIND project dataset[6].

Tips & Tricks. The main inductive bias in Meng et al. (2021) is the definition of class specific and class agnostic representations. In addition, the authors use MINE (Belghazi et al., 2018), which is a learned loss function, for disentangling the two vectors. Other differentiable metrics such as the Hilbert-Schmidt Independence Criterion (Gretton et al., 2005) could also be used as done in Liu et al. (2021b).

6.4. Registration

Image registration, defined as the alignment of the content of two images based on a transformation, constitutes an important pre-processing step in medical image analysis. This transformation can be parameterised by either an affine matrix (rigid) or by a displacement field (non-rigid). A major challenge is to define a cost function for multi-modal cases; for example, comparing MRI and ultrasound scans using pixel-level metrics is not effective since the intensity, view and artefacts are different. To address this problem, Chartsias et al. (2021) use the disentangled anatomical factors to register the cine- with the LGE-MRI scans of the same patient. The work of Qin et al. (2019) addresses it by leveraging CSD (Section 3.4). Registration of images with pathologies to atlases can also be problematic, therefore, Han et al. (2020) disentangles the disease features from the normal features, as in Section 6.1.1, and generates a deformation field based on the healthy features. We also refer the readers to a recent method (Maillard et al., 2022) that introduces a deep residual learning implementation of metamorphosis model to handle pathological medical images.

We detail the work of Qin et al. (2019) which uses a CNN to estimate the displacement field from the content space of two images (from different modalities). We choose to discuss this work as it is the first paper to leverage disentanglement for the task of medical image registration.

Architecture. Initially, a system based on the DRIT (Lee et al., 2018) architecture is used for image-to-image translation between two images of different domains. Then, they used the content E_c and style E_s encoders for each domain $\mathcal{X}_i$. Secondly, the content representations from two images are fed into a registration network G_{reg} that outputs deformation fields mapping one image to the other.

Training setup. A first training of the CSD is done as defined in Lee et al. (2018). Then, the registration network is learned by computing a bidirectional loss function based on the content latent space of the deformed images plus a regularisation loss over the latent space. The method is evaluated on lung CT scans from the COPDGene (Bakas et al., 2017) dataset and brain MRI from the BraTS corpus.

Tips & Tricks. The main inductive bias used by this model is the fact that registration depends only on the image content; the style can be ignored. In addition, preservation of topological information is an important constraint in medical image registration. The authors use a Huber loss over the gradients of the deformation field for this purpose, reinforcing smoothness of the deformation field.

6.5. Federated learning

DRL has just started to be leveraged for maintaining privacy by becoming invariant to private features (Marx et al., 2019; Aloufi et al., 2020; Bercea et al., 2021b). Considering that privacy issues in machine learning have attracted significant attention (Liu and Tsaftaris, 2020; Su et al., 2021; Jegorova et al., 2021; Hartley and Tsaftaris, 2022), we believe that there is a new, emerging domain for learning privacy-preserved disentangled representations. As for every new domain, it will be challenging to connect and exploit existing concepts, such as differential privacy (Dwork, 2006) and federated learning (Rieke et al., 2020; Li et al., 2020; Liu et al., 2021a; Bercea et al., 2021a; 2021b), with the disentanglement paradigm. Federated learning allows the model to be trained collaboratively by multiple local parties without exchanging or sharing their local data. In this case, the data distributed in local parties are better protected as they are not exposed to external authorities. Among the previous work in federated learning, methods with disentanglement (Liu et al., 2021a; Bercea et al., 2021a; 2021b) have shown improved performance on Cardiac-CT datasets including CT20 (Xu et al., 2019), CT34LC and CT34MC (Zhuang and Shen, 2016b) and several brain MRI datasets.

In particular, the work of Bercea et al. (2021b), termed FedDis, learns disentangled representations in the federated learning setting to detect brain anomalies by only sharing the disentangled shape information between clients, while the disentangled appearance information is kept locally. We detail FedDis as it has been extensively evaluated on many public datasets.

Architecture. FedDis does not have strong design biases to enforce disentanglement, which is mainly achieved by learning biases. FedDis has two auto-encoders to reconstruct the input image and encode the appearance and shape information.

Training setup. Apart from the reconstruction loss $\mathcal{L}_{rec}$ for the auto-encoders, FedDis introduces two losses as learning biases to enforce the auto-encoders to separately encode the shape and appearance information. The shape consistency loss $\mathcal{L}_{SCL}$ penalises the difference between the encoded shape embeddings of the original image and the Gamma-shifted image. The latent orthogonality loss $\mathcal{L}_{LOL}$ pushes away the distributions of the shape and appear-

[6] http://www.ifindproject.com/

ance embeddings. The overall loss is:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{SCL} + \lambda_3 \mathcal{L}_{LOL}. \quad (11)$$

FedDis has been extensively evaluated on several public brain MRI datasets including MSISBI (Ghosh et al., 2019), OASIS (LaMontagne et al., 2019), MSLUB (Lesjak et al., 2018), ADNI (Rieke et al., 2020) and BraTS (Menze et al., 2015).

Tips & Tricks. After training with healthy subjects, the shape auto-encoder is used to detect the brain anomalies. It assumes that the model cannot properly reconstruct the anomalies (*e.g.* tumors) as the anomalies are not seen during training. Hence, the anomalies are the parts where the reconstruction error is high. Although the reported results showcase the effectiveness of FedDis on detecting the anomalies, it may not work well when the shape encoder generalises well to reconstruct some tumors. Hence, some extra constraints to avoid such scenarios could be helpful to improve the robustness and performance of the model.

7. What can we learn from computer vision?

We are now well aware that learning disentangled representations requires supervision or design and learning biases. Using task prior knowledge to incorporate proper biases to learn the desired disentangled representations is key for disentanglement in both domains. Medical applications can use, for instance, building blocks (Section 4) originally designed for computer vision tasks. One can also draw inspiration from how prior knowledge on the vision tasks has motivated the specific biases used. Below, with some exemplar computer vision tasks, we discuss the connections between disentanglement in the computer vision and medical domains.

7.1. Image-to-image translation

Image-to-image (I2I) translation aims to translate one image into another without changing the shape, *i.e.* content, which differ in a specific characteristic (*e.g.* style). A representative model is MUNIT (Huang and Belongie, 2017) for which we provide details in Section 6.

Connections to medical. Image-to-image translation in computer vision motivated many medical applications as we detailed in Section 6. In fact, several medical models are directly built based on MUNIT such as the ones in medical I2I translation (Li et al., 2019a), multi-modal and cross-modal segmentation (Yang et al., 2020), and registration (Qin et al., 2019). The parallels here of domain-invariant spatial content and the domain-specific style representation, relate to separating anatomy and modality representations in the corresponding medical applications. A major difference though is that typically in medical image translation we are particularly sensitive to maintaining identity when changing style. Several vision works show examples of day to night where content has changed slightly in the background. Such change will not be desired in medical tasks.

7.2. Facial attribute transfer

This task concerns the generation of a synthetic face that contains the target attribute, but without altering the subject identity (*e.g.* adding bangs to a subjects forehead). Most methods that focus on facial attribute transfer struggle with: a) transferring more than one attribute at a time, b) generating images based on exemplars, and c) achieving high-fidelity results. The first model to address the aforementioned challenges is ELEGANT (Xiao et al., 2018), which encodes disentangled attribute representations of two exemplars in a vector latent space and performs attribute swapping. Apart from ELEGANT, Lin et al. (2021) propose a GAN model with a domain classifier to learn to transfer attributes between multiple domains. He et al. (2019) present a GAN that conditions the face generation of opposite samples (e.g. smile, no smile) using one-hot attribute vectors. Zhou et al. (2017) exploit cycle consistency to transfer attributes, with the limitation that the attributes should have approximately the same spatial location.

Connections to medical. When transferring facial attributes, the subject identity should be preserved and only some attributes transferred. This transferral is desirable in several medical applications such as brain aging Xia et al. (2019b) and controllable synthesis Thermos et al. (2021), where the synthesised brain or heart images should contain the identity information of the original images but with different ages or pathology. ELEGANT preserves the identity information by only modifying the local part of the image. The medical models similarly modify the local anatomy parts but also apply the identity or consistency losses to the remaining parts of the image. We should note that most face models rely on pre-trained or pre-extracted strong priors to identify facial features. Such strong priors are rarely available in medical imaging.

7.3. Pose estimation

For pose estimation, the human body constitutes a strong content prior that can be exploited to encode body structure in a spatial and semantic latent space, to be used for equivariant tasks that require body joint position. Lorenz et al. (2019) propose to apply the equivariance and invariance losses to learn the equivariant (content) and invariant (style) representations and use this type of disentanglement for this challenging articulated body pose estimation task. Esser et al. (2018) adopt the disentanglement of the human body pose from the corresponding appearance (style) information in the context of a dual-encoder VAE setting, where they use the body-related factors for human appearance transfer and synthesis (Esser et al., 2019).

Connections to medical. Similar to the human body, human organs *e.g.* brain and heart, have strong anatomical structure priors, which can be similarly used for learning disentangled representations with equivariance and invariance properties. For example, similar to the invariance loss in Lorenz et al. (2019), Bercea et al. (2021b) applies the shape consistency loss to encourage the shape embeddings of brain MRI images to be invariant to Gamma shifts. However, it is not always possible to assume such strong structural priors as diseases or abnormalities exist.

8. Limitations, opportunities and open challenges

In this section, we identify three key limitations of existing DRL methods and discuss ideas and research directions for improvement. We also present opportunities as well as various challenges to be addressed by the community.

8.1. New strategies for learning disentangled representations

Limitation. Learning disentangled representations requires complex architectures and objective functions. As we saw in Section 6, most approaches employ several loss functions and modules and, hence multiple hyperparameters. While flexibility is desirable, tuning complex systems can be difficult and it creates a barrier for further adoption of the disentanglement paradigm by the broader research community. Methods that require less hyperparameter tuning or techniques for automating this process or less complex approaches will be welcomed. Below, we discuss three possible strategies to learn disentangled representations in a simpler fashion.

Integrating self-supervised and contrastive learning. Fundamentally speaking most disentanglement approaches we reviewed

here use a reconstruction approach. This may not be necessary. Recently, contrastive learning (He et al., 2020; Chen et al., 2020a; 2020b; Grill et al., 2020; Zbontar et al., 2021) has shown impressive performance for self-supervised representation learning. In particular, patch-wise contrastive learning (Park et al., 2020) has been successfully used as an auxiliary loss function for reinforcing disentanglement (Zhou et al., 2021; Tomar et al., 2021). Additionally, Mitrovic et al. (2021) and Kügelgen et al. (2021) developed an understanding of contrastive learning from a causal perspective and argue that it can be interpreted as CSD where the representation is focusing on learning only the content, whilst developing style invariance. Methods such as MOCO (He et al., 2020), SimCLR (Chen et al., 2020a; 2020b), BYOL (Grill et al., 2020), and the Barlow Twins (Zbontar et al., 2021) achieve this through augmentation and regularisation. Wang et al. (2021d) use contrastive learning for disentangling group invariant representations. Ren et al. (2021) propose to discover the disentangled representations with contrasting learning at the post-hoc stage. Zimmermann et al. (2021) have taking it a step further to suggest that contrastive learning under certain assumptions can indeed invert the data generating process. While it is possible to learn representations that are robust (invariant) to specific interventions, it remains challenging to design augmentations and regularisations which are invariant to general interventions.

Intervention as a prior. Caselles-Dupré et al. (2019) suggest that a symmetry-based understanding of disentanglement can only be achieved upon interaction with an environment. To illustrate this point, Suter et al. (2019) propose a disentanglement metric based on interventional robustness. Moreover, statistical independence between latent variables might not hold for real-life settings where the generating factors are correlated (Dittadi et al., 2021; Träuble et al., 2021a). With this intuition, Besserve et al. (2020) provide a causal understanding of disentanglement in generative models based on interventions and counterfactuals. Leeb et al. (2021) propose a strategy for probing the latent space of VAEs by applying interventions. Their method allows quantification of the consistency of the representation with a chosen prior as well as finding holes in the latent manifold. These works pave a new path for using interventions as a prior for DRL.

Compositionality as a prior. As reported in Section 3.4, in current CSD models the content is vaguely defined as domain-invariant (Huang and Belongie, 2017), task-equivariant (Lorenz et al., 2019) or even simply as spatial and binary (Chartsias et al., 2019). These definitions usually point to the task-driven model designs for learning the desired content, which are tailored to specific datasets or tasks. Enforcing compositionality could be the solution for learning generalisable and robust content representations in vision. This intuition is based on the compositional nature of the human cognition, which is robust for recognising new concepts by composing individual components (Stone et al., 2017). Considering compositionality within disentanglement could be a fruitful direction.

8.2. Disentanglement with additional properties

Limitation. Generalisation on unseen data is the holy grail even in medical applications (Sermesant et al., 2021). Although disentangled representations should be general, Recent studies (Montero et al., 2020; Schott et al., 2021) found that disentanglement does not guarantee, for instance, combinatorial generalisation (understand and produce novel combinations of familiar elements). Another important limitation is learning disentangled representation from correlated data (Träuble et al., 2021b). As detailed in Section 2.4, real data is not *i.i.d.* and bias exist due to domain shifts. In these cases, it has been shown that factorization-based inductive biases as described in Section 3.1 are not enough to learn the true generating factors. These biases can have significant implications for domain generalisation and fairness (biased towards sensitive attributes).

Domain generalisable disentangled representations. Domain generalisation is a setting which considers that no information from the target domain is available and that a model trained on multiple source domains needs to generalise well to the unseen target domain (Li et al., 2018; Wang et al., 2021b). To address this, Meng et al. (2021) use task-specific representations and feature clustering to achieve domain invariance, and Liu et al. (2021b) use meta-learning to explicitly improve domain invariance in disentangled representations. Concepts of causal representation learning (Schlkopf et al., 2021) (Section 2.5.3) can help when defining and becoming robust to domain shifts when there are data biases (Arjovsky et al., 2019; Krueger et al., 2021). Recent work (Wang et al., 2021d) disentangles group-invariant representations in a self-supervised setting using ideas from causal invariance (Arjovsky et al., 2019). Learning robust and generalisable representations, however, remains an open problem.

Fair disentangled representations. Fairness is an important concept in machine learning whenever an algorithm tends to be biased towards sensitive attributes such as race or gender (Puyol-Antón et al., 2021a; 2021b). Therefore, a fair model should be invariant to sensitive attributes. Developing fair algorithms is tightly related to domain generalisation as detailed in Creager et al. (2021) and disentanglement provides a useful framework for dealing with these issues (Locatello et al., 2019a; Creager et al., 2019; Sarhan et al., 2020; Xianjing et al., 2021).

8.3. Robust measurements of disentanglement

Limitation. As analysed by Locatello et al. (2019b), most of the metrics reported in Section 5 require ground truth for each latent factor or do not perform consistently for different tasks and datasets. Additionally, as experiments of Träuble et al. (2021b) show, most existing metrics struggle when measuring the disentanglement of models trained with data that include correlated factors of variation.

Metrics for real data. Although a recent method considers to measure the disentanglement of hierarchically structured representations (Dang-Nhu, 2021), robust disentanglement metrics that work well with real-world data (with any form and structure of generative factors) is still an open challenge. On the other hand, CSD disentanglement has attracted significant attention, with the exception of the metrics proposed by Liu et al. (2021c), the development of metrics that work with latents of diverse dimensionality is still an open problem. Such metrics can be further exploited to improve disentanglement itself in an iterative manner, as Estermann et al. (2020) have done.

9. Conclusion

Overall, disentangled representation learning is a tool for introducing inductive biases (expert knowledge) into deep learning settings in order to simulate real-life scenarios with non-*i.i.d.* data. In this article, we have reviewed methods for implicitly or explicitly forcing representations to be invariant or equivariant to specific changes in the input data. We have emphasised building blocks for introducing disentanglement into a diverse set of tasks. In summary, disentanglement can be achieved with modifications in the model architecture (*e.g.* MUNIT, StyleGAN) and/or regularisation constraints (*e.g.* β-VAE). We highlight that disentanglement can be especially useful in low data regimes where biases are more relevant. By detailing limitations, opportunities and open challenges we hope to inspire the community to continue to investigate this extremely important area for learning better data representations.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

This work was supported by the Royal Academy of Engineering and Canon Medical Research Europe, and partially supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. S.A. Tsaftaris acknowledges the support of Canon Medical and the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme (grant RCSRF1819 8 25). We thank the participants of the DREAM tutorials for feedback. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

References

Achille, A., Soatto, S., 2017. Emergence of Invariance and Disentanglement in Deep Representations. Journal of Machine Learning Research 19, 1–34.
Alipour, S.H.M., Rabbani, H., Akhlaghi, M.R., 2012. Diabetic retinopathy grading by digital curvelet transform. Computational and Mathematical Methods in Medicine.
Almahairi, A., Rajeshwar, S., Sordoni, A., Bachman, P., Courville, A., 2018. Augmented CycleGAN: Learning many-to-many mappings from unpaired data. In: Proc. International Conference on Machine Learning (ICML), pp. 195–204.
Aloufi, R., Haddadi, H., Boyle, D., 2020. Privacy-preserving voice analysis via disentangled representations. In: Proc. ACM SIGSAC Conference on Cloud Computing Security Workshop, pp. 1–14.
Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D., 2019. Invariant risk minimization. preprint arXiv:1907.02893.
Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C., 2017. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific data 4 (1), 1–13.
Bashyam, V.M., Doshi, J., Erus, G., Srinivasan, D., Abdulkadir, A., Habes, M., Fan, Y., Masters, C.L., Maruff, P., Zhuo, C., et al., 2020. Medical image harmonization using deep learning based canonical mapping: Toward robust and generalizable learning in imaging. preprint arXiv:2010.05355.
Bass, C., da Silva, M., Sudre, C., Tudosiu, P.-D., Smith, S., Robinson, E., 2020. Icam: Interpretable classification via disentangled representations and feature attribution mapping. Advances in Neural Information Processing Systems 33, 7697–7709.
Bass, C., da Silva, M., Sudre, C.H., Williams, L.Z.J., Tudosiu, P.-D., Alfaro-Almagro, F., Fitzgibbon, S.P., Glasser, M., Smith, S.M., Robinson, E.C., 2021. ICAM-reg: Interpretable classification and regression with feature attribution for mapping neurological phenotypes in individual scans. Medical Imaging with Deep Learning.
Belghazi, M.I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., Hjelm, D., 2018. Mutual information neural estimation. In: Proc. International Conference on Machine Learning (ICML), pp. 531–540.
Ben-Cohen, A., Mechrez, R., Yedidia, N., Greenspan, H., 2019. Improving CNN training using disentanglement for liver lesion classification in CT. In: Proc. IEEE Engineering in Medicine and Biology Society (EMBC), pp. 886–889.
Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8), 1798–1828.
Bercea, C., Wiestler, B., Rueckert, D., Albarqouni, S., 2021. Federated disentangled representation learning for unsupervised brain anomaly. detection.
Bercea, C.I., Wiestler, B., Rueckert, D., Albarqouni, S., 2021. FedDis: Disentangled federated learning for unsupervised brain pathology segmentation. arXiv:2103.03705. [Online]. Available: https://arxiv.org/abs/2103.03705.
Berenguer, A.D., et al., 2020. Explainable-by-design semi-supervised representation learning for COVID-19 diagnosis from CT imaging. arXiv:2011.11719. [Online]. Available: https://arxiv.org/abs/2011.11719.
Bernard, O., et al., 2018. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Transactions on Medical Imaging 37 (11), 2514–2525.
Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., Raffel, C.A., 2019. Mixmatch: A holistic approach to semi-supervised learning. In: Proc. of Advances in Neural Information Processing Systems (NeurIPS).
Besserve, M., Mehrjou, A., Sun, R., Schlkopf, B., 2020. Counterfactuals uncover the modular structure of deep generative models. International Conference on Learning Representations (ICLR).
Bingham, E., Chen, J.P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P., Horsfall, P., Goodman, N.D., 2019. Pyro: Deep universal probabilistic programming. The Journal of Machine Learning Research (JMLR) 20 (1), 973–978.
Brock, A., Donahue, J., Simonyan, K., 2019. Large scale GAN training for high fidelity natural image synthesis. International Conference on Learning Representations Workshop (ICLR).
Bronstein, M.M., Bruna, J., Cohen, T., Velikovi, P., 2021. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv:2104.13478. [Online]. Available: https://arxiv.org/abs/2104.13478.
Burgess, C., Kim, H., 2018. 3D shapes dataset. [Online]. Available: https://github.com/deepmind/3d-shapes.
Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A., 2018. Understanding disentangling in β-VAE. arXiv:1804.03599. [Online]. Available: https://arxiv.org/abs/1804.03599.
Campello, V.M., et al., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge. IEEE Transactions on Medical Imaging (Early Access).
Caselles-Dupré, H., Garcia Ortiz, M., Filliat, D., 2019. Symmetry-based disentangled representation learning requires interaction with environments. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 4606–4615.
Castro, D.C., Walker, I., Glocker, B., 2020. Causality matters in medical imaging. Nature Communications 11 (1), 1–10.
Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2019. Disentangled representation learning in cardiac image analysis. Medical Image Analysis 58, 101535.
Chartsias, A., Papanastasiou, G., Wang, C., Semple, S., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A., 2021. Disentangle, align and fuse for multimodal and semi-supervised image segmentation. IEEE Transactions on Medical Imaging 40 (3), 781–792.
Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.-A., 2019. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 447–456.
Chen, C., Ouyang, C., Tarroni, G., Schlemper, J., Qiu, H., Bai, W., Rueckert, D., 2019. Unsupervised multi-modal style transfer for cardiac MR segmentation. In: Proc. International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM), pp. 209–219.
Chen, J., Zhang, Z., Xie, X., Li, Y., Xu, T., Ma, K., Zheng, Y., 2021. Beyond mutual information: Generative adversarial network for domain adaptation using information bottleneck constraint. IEEE Transactions on Medical Imaging.
Chen, R.T.Q., Li, X., Grosse, R.B., Duvenaud, D.K., 2018. Isolating sources of disentanglement in variational autoencoders. In: Proc. Advances in Neural Information Processing Systems (NeurIPS).
Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. In: Proc. International Conference on Machine Learning (ICML), pp. 1597–1607.
Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E., 2020. Big self-supervised models are strong semi-supervised learners. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 22243–22255.
Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P., 2016. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 2180–2188.
Chen, X., Lian, C., Wang, L., Deng, H., Kuang, T., Fung, S.H., Gateno, J., Shen, D., Xia, J.J., Yap, P.-T., 2021. Diverse data augmentation for learning image segmentation with cross-modality annotations. Medical Image Analysis 71, 102060.
Cheng, C., Kennedy, J., Seymour, C., Weiss, J.C., 2021. Disentangled hyperspherical clustering for sepsis phenotyping. Artificial Intelligence in Medicine. Springer.
Cherepkov, A., Voynov, A., Babenko, A., 2021. Navigating the GAN parameter space for semantic image editing. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
Cheung, B., Livezey, J.A., Bansal, A.K., Olshausen, B.A., 2015. Discovering hidden factors of variation in deep networks. International Conference on Learning Representations Workshop (ICLRW).
Christ, P., Ettlinger, F., Grün, F., Lipkova, J., Kaissis, G., 2017. Lits-liver tumor segmentation challenge. ISBI and MICCAI.
Cohen, T., 2021. Equivariant Convolutional Networks. University of Amsterdam.
Cohen, T., Welling, M., 2016. Group equivariant convolutional networks. In: Proc. International Conference on Machine Learning (ICML), pp. 2990–2999.
Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B., 2016. The Cityscapes dataset for semantic urban scene understanding. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223.
Couronné, R., Vernhet, P., Durrleman, S., 2021. Longitudinal self-supervision to disentangle inter-patient variability from disease progression. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 231–241.
Creager, E., Jacobsen, J.-H., Zemel, R., 2021. Environment inference for invariant learning. In: Proc. International Conference on Machine Learning. PMLR, pp. 2189–2200.
Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., Zemel, R., 2019. Flexibly fair representation learning by disentanglement. In: Proc. International Conference on Machine Learning (ICML). PMLR, pp. 1436–1445.
Cristianini, N., Shawe-Taylor, J., Elisseeff, A., Kandola, J.S., 2002. On kernel-target alignment. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 367–373.
Dang-Nhu, R., 2021. Evaluating disentanglement of structured representations. In: Proc. International Conference on Learning Representations (ICLR).
Dash, S., Balasubramanian, V.N., Sharma, A., 2020. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals. arXiv:2009.08270. [Online]. Available: https://arxiv.org/abs/2009.08270.

Dewey, B.E., et al., 2020. A disentangled latent space for cross-site mri harmonization. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 720–729.
Dinh, L., Krueger, D., Bengio, Y., 2015. Nice: Non-linear independent components estimation. In: International Conference on Learning Representations Workshops (ICLRW).
Dittadi, A., Träuble, F., Locatello, F., Wüthrich, M., Agrawal, V., Winther, O., Bauer, S., Schölkopf, B., 2021. On the transfer of disentangled representations in realistic settings. In: Proc. International Conference on Learning Representations (ICLR).
Dolatabadi, H.M., Erfani, S., Leckie, C., 2020. Invertible generative modeling using linear rational splines. In: Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 4236–4246.
Donahue, J., Krähenbühl, P., Darrell, T., 2016. Adversarial feature learning. International Conference on Learning Representations (ICLR).
Duan, S., Matthey, L., Saraiva, A., Watters, N., Burgess, C., Lerchner, A., Higgins, I., 2020. Unsupervised model selection for variational disentangled representation learning. International Conference on Learning Representations (ICLR).
Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovski, M., Mastropietro, O., Courville, A., 2016. Adversarially learned inference. International Conference on Learning Representations (ICLR).
Durkan, C., Bekasov, A., Murray, I., Papamakarios, G., 2019. Neural spline flows. In: Proc. Advances in Neural Information Processing Systems (NeurIPS).
Dwork, C., 2006. Differential privacy. In: International Colloquium on Automata, Languages, and Programming. Springer, pp. 1–12.
Eastwood, C., Williams, C.K.I., 2018. A framework for the quantitative evaluation of disentangled representations. International Conference on Learning Representations (ICLR).
Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D.H., Dy, J., van de Meent, J., 2019. Structured disentangled representations. In: Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 2525–2534.
Esser, P., Haux, J., Ommer, B., 2019. Unsupervised robust disentangling of latent characteristics for image synthesis. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2699–2709.
Esser, P., Rombach, R., Ommer, B., 2020. A disentangling invertible interpretation network for explaining latent representations. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9223–9232.
Esser, P., Sutter, E., Ommer, B., 2018. A variational U-Net for conditional appearance and shape generation. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8857–8866.
Estermann, B., Marks, M., Yanik, M.F., 2020. Robust disentanglement of a few factors at a time. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 13387–13398.
Fei, Y., Zhan, B., Hong, M., Wu, X., Zhou, J., Wang, Y., 2021. Deep learning-based multi-modal computing with feature disentanglement for MRI image synthesis. Medical Physics 48 (7), 3778–3789.
Gabbay, A., Hoshen, Y., 2020. Demystifying inter-class disentanglement. International Conference on Learning Representations (ICLR).
Ganin, Y., Lempitsky, V., 2015. Unsupervised domain adaptation by backpropagation. In: Proc. International Conference on Machine Learning, pp. 1180–1189.
Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2414–2423.
Ghosh, S., Das, N., Das, I., Maulik, U., 2019. Understanding deep learning techniques for image segmentation. ACM Computing Surveys (CSUR) 52 (4), 1–35.
Glocker, B., Zikic, D., Konukoglu, E., Haynor, D.R., Criminisi, A., 2013. Vertebrae localization in pathological spine ct via dense classification from sparse annotations. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 262–270.
Gonzalez-Garcia, A., Van De Weijer, J., Bengio, Y., 2018. Image-to-image translation for cross-domain disentanglement. arXiv:1805.09730. [Online]. Available: https://arxiv.org/abs/1805.09730.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Proc. Advances in Neural Information Processing Systems (NeurIPS).
Graves, A., Mohamed, A.-r., Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In: Proc. of IEEE international conference on acoustics, speech and signal processing, pp. 6645–6649.
Gravina, M., Marrone, S., Sansone, M., Sansone, C., 2021. DAE-CNN: Exploiting and disentangling contrast agent effects for breast lesions classification in DCE-MRI. Pattern Recognition Letters 145, 67–73.
Gretton, A., Bousquet, O., Smola, A., Schölkopf, B., 2005. Measuring statistical dependence with hilbert-schmidt norms. In: Proc. of International Conference on algorithmic learning theory, pp. 63–77.
Grill, J.-B., et al., 2020. Bootstrap your own latent: A new approach to self-supervised learning. arXiv:2006.07733..
Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A., 2017. Improved training of wasserstein GANs. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 5769–5779.
Gyawali, P.K., Li, Z., Ghimire, S., Wang, L., 2019. Semi-supervised learning by disentangling and self-ensembling over stochastic latent space. In: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 766–774.
Han, X., Shen, Z., Xu, Z., Bakas, S., Akbari, H., Bilello, M., Davatzikos, C., Niethammer, M., 2020. A deep network for joint registration and reconstruction of images with pathologies. In: Machine Learning in Medical Imaging: 11th International Workshop (MLMI), pp. 342–352.
Harada, S., Bise, R., Hayashi, H., Tanaka, K., Uchida, S., 2021. Order-guided disentangled representation learning for ulcerative colitis classification with limited labels. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 471–480.
Hartley, J., Tsaftaris, S.A., 2022. Measuring unintended memorisation of unique private features in neural networks. preprint arXiv:2202.08099.
Havaei, M., Mao, X., Wang, Y., Lao, Q., 2021. Conditional generation of medical images via disentangled adversarial inference. Medical Image Analysis 72, 102106.
He, K., Fan, H., Wu, Y., Xie, S., Girshick, R., 2020. Momentum contrast for unsupervised visual representation learning. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9726–9735.
He, Z., Zuo, W., Kan, M., Shan, S., Chen, X., 2019. AttGAN: Facial attribute editing by only changing what you want. IEEE Transactions on Image Processing 28 (11), 5464–5478.
Hiasa, Y., Otake, Y., Takao, M., Matsuoka, T., Takashima, K., Carass, A., Prince, J.L., Sugano, N., Sato, Y., 2018. Cross-modality image synthesis from unpaired data using CycleGAN. In: Proc. International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), pp. 31–41.
Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., Lerchner, A., 2018. Towards a definition of disentangled representations. arXiv:1812.02230. [Online]. Available: https://arxiv.org/abs/1812.02230.
Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A., 2017. β-VAE: Learning basic visual concepts with a constrained variational framework. International Conference on Learning Representations (ICLR).
Hochberg, D.C., Giryes, R., Greenspan, H., 2021. Style encoding for class-specific image generation. Medical Imaging 2021: Image Processing. International Society for Optics and Photonics,SPIE. 834–840
Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., Darrell, T., 2018. CyCADA: Cycle-consistent adversarial domain adaptation. In: Proc. International Conference on Machine Learning (ICML), pp. 1989–1998.
Huang, J., Guan, D., Xiao, A., Lu, S., 2021. FSDR: Frequency space domain randomization for domain generalization. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6891–6902.
Huang, X., Belongie, S., 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1501–1510.
Huang, X., Liu, M.-Y., Belongie, S., Kautz, J., 2018. Multimodal unsupervised image–to-image translation. In: Proc. European Conference on Computer Vision (ECCV), pp. 172–189.
Huang, Y., Xia, W., Lu, Z., Liu, Y., Zhou, J., Fang, L., Zhang, Y., 2020. Disentanglement network for unsupervised speckle reduction of optical coherence tomography images. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 675–684.
Hyvärinen, A., Pajunen, P., 1999. Nonlinear independent component analysis: Existence and uniqueness results. Neural networks 12 (3), 429–439.
Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proc. International Conference on Machine Learning (ICML), pp. 448–456.
Jahanian, A., Chai, L., Isola, P., 2020. On the "steerability" of generative adversarial networks. International Conference on Learning Representations Workshop (ICLR).
Jang, E., Gu, S., Poole, B., 2016. Categorical reparameterization with Gumbel-softmax. arXiv:1611.01144. [Online]. Available: https://arxiv.org/abs/1611.01144.
Jegorova, M., Kaul, C., Mayor, C., O'Neil, A.Q., Weir, A., Murray-Smith, R., Tsaftaris, S.A., 2021. Survey: Leakage and privacy at inference time. arXiv:2107.01614. [Online]. Available: https://arxiv.org/abs/2107.01614.
Jiang, H., Chartsias, A., Zhang, X., Papanastasiou, G., Semple, S., Dweck, M., Semple, D., Dharmakumar, R., Tsaftaris, S.A., 2020. Semi-supervised pathology segmentation with disentangled representations. In: Proc. MICCAI Workshop on Domain Adaptation and Representation Transfer (DART). Springer, pp. 62–72.
Jiang, J., Veeraraghavan, H., 2020. Unified cross-modality feature disentangler for unsupervised multi-domain mri abdomen organs segmentation. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 347–358.
Jiang, K., Quan, L., Gong, T., 2022. Disentangled representation and cross-modality image translation based unsupervised domain adaptation method for abdominal organ segmentation. International Journal of Computer Assisted Radiology and Surgery 1–13.
Johnson, J., Alahi, A., Li, F.-F., 2016. Perceptual losses for real-time style transfer and super-resolution. In: Proc. European Conference on Computer Vision (ECCV), pp. 694–711.
Jung, D., Lee, J., Yi, J., Yoon, S., 2020. iCaps: An interpretable classifier via disentangled capsule networks. arXiv:2008.08756. [Online]. Available: https://arxiv.org/abs/2008.08756.
Kalkhof, J., González, C., Mukhopadhyay, A., 2022. Disentanglement enables cross–domain hippocampus segmentation. preprint arXiv:2201.05650.
Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410.
Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T., 2020. Analyzing and improving the image quality of StyleGAN. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8107–8116.
Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.-H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al., 2021. CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis 69, 101950.

Kelkar, V.A., Anastasio, M.A., 2022. Prior image-based medical image reconstruction using a style-based generative adversarial network. preprint arXiv:2202.08936.
Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A., 2020. Variational autoencoders and nonlinear ICA: A unifying framework. In: Proc. International Conference on Artificial Intelligence and Statistics (AISTATS), Vol. 108, pp. 2207–2217.
Kim, H., Mnih, A., 2018. Disentangling by factorising. In: Proc. International Conference on Machine Learning (ICML), pp. 2649–2658.
Kingma, D.P., Dhariwal, P., 2018. Glow: Generative flow with invertible 1x1 convolutions. In: Proc. Advances in Neural Information Processing Systems (NeurIPS).
Kingma, D.P., Welling, M., 2014. Auto-encoding variational bayes. International Conference on Learning Representations (ICLR).
Kobayashi, K., Hataya, R., Kurose, Y., Miyake, M., Takahashi, M., Nakagawa, A., Harada, T., Hamamoto, R., 2021. Decomposing normal and abnormal features of medical images for content-based image retrieval of glioma imaging. Medical Image Analysis 74, 102227.
Kobyzev, I., Prince, S., Brubaker, M., 2020. Normalizing flows: An introduction and review of current methods. IEEE Transactions on Pattern Analysis and Machine Intelligence (Early Access).
Kocaoglu, M., Snyder, C., Dimakis, A.G., Vishwanath, S., 2018. CausalGAN: Learning causal implicit generative models with adversarial training. International Conference on Learning Representations (ICLR).
Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Priol, R.L., Courville, A., 2021. Out-of-distribution generalization via risk. extrapolation.
Kügelgen, J.V., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in Neural Information Processing Systems.
Kumar, A., Sattigeri, P., Balakrishnan, A., 2018. Variational inference of disentangled latent concepts from unlabeled observations. International Conference on Learning Representations (ICLR).
LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., et al., 2019. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. MedRxiv.
Lao, Q., Havaei, M., Pesaranghader, A., Dutil, F., Jorio, L.D., Fevens, T., 2019. Dual adversarial inference for text-to-image synthesis. In: Proc. IEEE/CVF International Conference on Computer Vision (CVPR), pp. 7567–7576.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86 (11), 2278–2324.
Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M.K., Yang, M.-H., 2018. Diverse image–to-image translation via disentangled representations. In: Proc. European Conference on Computer Vision (ECCV), pp. 36–52.
Leeb, F., Annadani, Y., Bauer, S., Schölkopf, B., 2020. Structured representation learning using structural autoencoders and hybridization. arXiv:2006.07796. [Online]. Available: https://arxiv.org/abs/2006.07796.
Leeb, F., Bauer, S., Schölkopf, B., 2021. Interventional assays for the latent space of autoencoders. preprint arXiv:2106.16091.
Lenc, K., Vedaldi, A., 2015. Understanding image representations by measuring their equivariance and equivalence. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 991–999.
Lesjak, Ž., Galimzianova, A., Koren, A., Lukin, M., Pernuš, F., Likar, B., Špiclin, Ž., 2018. A novel public MR image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus. Neuroinformatics 16 (1), 51–63.
Li, C., Liu, H., Chen, C., Pu, Y., Chen, L., Henao, R., Carin, L., 2017. ALICE: Towards understanding adversarial learning for joint distribution matching. In: Proc. Advances in Neural Information Processing Systems (NeurIPS).
Li, D., Kar, A., Ravikumar, N., Frangi, A.F., Fidler, S., 2020. Federated simulation for medical imaging. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 159–168.
Li, D., Yang, Y., Song, Y.-Z., Hospedales, T.M., 2018. Learning to generalize: Meta-learning for domain generalization. In: Proc. AAAI Conference on Artificial Intelligence (AAAI).
Li, H., Gopal, S., Sekuboyina, A., Zhang, J., Niu, C., Pirkl, C., Kirschke, J., Wiestler, B., Menze, B., 2021. Unpaired MR image homogenisation by disentangled representations and its uncertainty. Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Perinatal Imaging, Placental and Preterm Image Analysis. Springer International Publishing.
Li, K., Yu, L., Wang, S., Heng, P.-A., 2019. Unsupervised retina image synthesis via disentangled representation learning. In: Proc. International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), pp. 32–41.
Li, Z., Li, H., Han, H., Shi, G., Wang, J., Zhou, S.K., 2019. Encoding CT anatomy knowledge for unpaired chest X-ray image decomposition. In: Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 275–283.
Liao, H., Lin, W.-A., Zhou, S.K., Luo, J., 2019. Artifact disentanglement network for unsupervised metal artifact reduction. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).
Liao, H., Lin, W.-A., Zhou, S.K., Luo, J., 2020. ADN: Artifact disentanglement network for unsupervised metal artifact reduction. IEEE Transactions on Medical Imaging 39 (3), 634–643.
Lin, J., Chen, Z., Xia, Y., Liu, S., Qin, T., Luo, J., 2021. Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation. IEEE Transactions on Pattern Analysis and Machine Intelligence 43 (4), 1254–1266.
Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection. In: Proc. IEEE/CVF International Conference on Computer Vision (CVPR), pp. 2117–2125.
Liu, B., Zhu, Y., Fu, Z., de Melo, G., Elgammal, A., 2020. OOGAN: Disentangling GAN with one-hot sampling and orthogonal regularization. In: Proc. AAAI Conference on Artificial Intelligence (AAAI), pp. 4836–4843.
Liu, Q., Chen, C., Qin, J., Dou, Q., Heng, P.-A., 2021. FedDG: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1013–1023.
Liu, S., Dowling, J.A., Engstrom, C., Greer, P.B., Crozier, S., Chandra, S.S., 2020. Manipulating medical image translation with manifold. disentanglement.
Liu, S., Gibson, E., Grbic, S., Xu, Z., Setio, A.A.A., Yang, J., Georgescu, B., Comaniciu, D., 2018. Decompose to manipulate: Manipulable object synthesis in 3D medical images with structured image decomposition. arXiv:1812.01737. [Online].. Available: https://arxiv.org/abs/1812.01737
Liu, X., Thermos, S., Chartsias, A., O'Neil, A., Tsaftaris, S.A., 2020. Disentangled representations for domain-generalized cardiac segmentation. In: Proc. International Workshop on Statistical Atlases and Computational Models of the Heart (STACOM), pp. 187–195.
Liu, X., Thermos, S., O'Neil, A., Tsaftaris, S.A., 2021. Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).
Liu, X., Thermos, S., Valvano, G., Chartsias, A., O'Neil, A., Tsaftaris, S.A., 2021. Measuring the biases and effectiveness of content-style disentanglement. In: Proc. British Machine Vision Conference (BMVC).
Liu, X., Tsaftaris, S.A., 2020. Have you forgotten? a method to assess if machine learning models have forgotten data. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 95–105.
Liu, Y., Wagner, S.J., Peng, T., 2022. Multi-modality microscopy image style augmentation for nuclei segmentation. Journal of Imaging 8 (3), 71.
Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., Bachem, O., 2019. On the fairness of disentangled representations. Advances in Neural Information Processing Systems (NeurIPS) 32.
Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O., 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In: Proc. International Conference on Machine Learning (ICML), pp. 4114–4124.
Lorenz, D., Bereska, L., Milbich, T., Ommer, B., 2019. Unsupervised part-based disentangling of object shape and appearance. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10955–10964.
Lyu, Y., Liao, H., Zhu, H., Zhou, S.K., 2020. Joint unsupervised learning for the vertebra segmentation, artifact reduction and modality translation of CBCT images. arXiv:2001.00339. [Online]. Available: https://arxiv.org/abs/2001.00339.
Maier, O., et al., 2017. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis 35, 250–269.
Maillard, M., François, A., Glaunès, J., Bloch, I., Gori, P., 2022. A deep residual learning implementation of metamorphosis. preprint arXiv:2202.00676.
Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Smolley, S.P., 2017. Least squares generative adversarial networks. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2813–2821.
Marx, C., Phillips, R., Friedler, S., Scheidegger, C., Venkatasubramanian, S., 2019. Disentangling influence: Using disentangled representations to audit model predictions. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 4496–4506.
Memmel, M., Gonzalez, C., Mukhopadhyay, A., 2021. Adversarial continual learning for multi-domain hippocampal segmentation. In: Proc. Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health (DART). Springer, pp. 35–45.
Meng, Q., Matthew, J., Zimmer, V.A., Gomez, A., Lloyd, D.F.A., Rueckert, D., Kainz, B., 2021. Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging. IEEE Transactions on Medical Imaging 40 (2), 722–734.
Meng, Q., Pawlowski, N., Rueckert, D., Kainz, B., 2019. Representation disentanglement for multi-task learning with application to fetal ultrasound. In: Proc. Smart Ultrasound Imaging and Perinatal, Preterm and Paediatric Image Analysis (PIPPI), pp. 47–55.
Menze, B.H., et al., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging 34 (10), 1993–2024.
Mitrovic, J., McWilliams, B., Walker, J.C., Buesing, L.H., Blundell, C., 2021. Representation learning via invariant causal mechanisms. International Conference on Learning Representations (ICLR).
Montero, M.L., Ludwig, C.J., Costa, R.P., Malhotra, G., Bowers, J., 2020. The role of disentanglement in generalisation. International Conference on Learning Representations (ICLR).
Mukherjee, S., Asnani, H., Lin, E., Kannan, S., 2019. ClusterGAN: Latent space clustering in generative adversarial networks. In: Proc. AAAI Conference on Artificial Intelligence (AAAI), pp. 4610–4617.
N, S., Paige, B., van de Meent, J.-W., Desmaison, A., Goodman, N., Kohli, P., Wood, F., Torr, P., 2017. Learning disentangled representations with semi-supervised deep generative models. In: Proc. Advances in Neural Information Processing Systems (NeurIPS).
Nie, W., Karras, T., Garg, A., Debnath, S., Patney, A., Patel, A., Anandkumar, A., 2020. Semi-supervised StyleGAN for disentanglement learning. In: Proc. International Conference on Machine Learning (ICML), pp. 7360–7369.

Ning, M., Bian, C., Wei, D., Yu, S., Yuan, C., Wang, Y., Guo, Y., Ma, K., Zheng, Y., 2021. A new bidirectional unsupervised domain adaptation segmentation framework. In: Proc. International Conference on Information Processing in Medical Imaging (IPMI). Springer, pp. 492–503.

Niu, C., Cong, W., Fan, F.-L., Shan, H., Li, M., Liang, J., Wang, G., 2021. Low-dimensional manifold constrained disentanglement network for metal artifact reduction. IEEE Transactions on Radiation and Plasma Medical Sciences doi:10.1109/TRPMS.2021.3122071. 1–1

Ouyang, J., Adeli, E., Pohl, K.M., Zhao, Q., Zaharchuk, G., 2021. Representation disentanglement for multi-modal brain MRI analysis. In: Proc. International Conference on Information Processing in Medical Imaging (IPMI). Springer, pp. 321–333.

Papamakarios, G., Nalisnick, E., Rezende, D.J., Mohamed, S., Lakshminarayanan, B., 2021. Normalizing flows for probabilistic modeling and inference. Journal of Machine Learning Research 22, 1–64.

Park, T., Efros, A.A., Zhang, R., Zhu, J.-Y., 2020. Contrastive learning for unpaired image-to-image translation. In: Proc. European Conference on Computer Vision (ECCV), pp. 319–345.

Park, T., Liu, M.-Y., Wang, T.-C., Zhu, J.-Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2337–2346.

Pawlowski, N., Castro, D.C., Glocker, B., 2020. Deep structural causal models for tractable counterfactual inference. In: Proc. Advances in Neural Information Processing Systems (NeurIPS).

Peebles, W., Peebles, J., Zhu, J.-Y., Efros, A.A., Torralba, A., 2020. The hessian penalty: A weak prior for unsupervised disentanglement. In: Proc. European Conference on Computer Vision (ECCV), pp. 581–597.

Pei, C., Wu, F., Huang, L., Zhuang, X., 2021. Disentangle domain features for cross–modality cardiac image segmentation. Medical Image Analysis 71, 102078.

Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.C., 2018. FiLM: Visual reasoning with a general conditioning layer. In: Proc. AAAI Conference on Artificial Intelligence (AAAI), pp. 3942–3951.

Peters, J., Janzing, D., Schölkopf, B., 2017. Elements of causal inference: foundations and learning algorithms. The MIT Press.

Petersen, R.C.o., 2010. Alzheimer's disease neuroimaging initiative (ADNI). Neurology 74 (3), 201–209.

Pfeiffer, M., et al., 2019. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In: Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 119–127.

Prados, F., Ashburner, J., Blaiotta, C., Brosch, T., Carballido-Gamio, J., Cardoso, M.J., Conrad, B.N., Datta, E., Dávid, G., De Leener, B., et al., 2017. Spinal cord grey matter segmentation challenge. Neuroimage 152, 312–329.

Puyol-Antón, E., Chen, C., Clough, J.R., Ruijsink, B., Sidhu, B.S., Gould, J., Porter, B., Elliott, M., Mehta, V., Rueckert, D., Rinaldi, C.A., King, A.P., 2020. Interpretable deep models for cardiac resynchronisation therapy response prediction. In: Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI). Springer.

Puyol-Antón, E., Ruijsink, B., Harana, J.M., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., Chowienczyk, P., King, A.P., 2021. Fairness in cardiac magnetic resonance imaging: Assessing sex and racial bias in deep learning-based segmentation. medRxiv.

Puyol-Antón, E., Ruijsink, B., Piechnik, S.K., Neubauer, S., Petersen, S.E., Razavi, R., King, A.P., 2021. Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 413–423.

Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., Kamen, A., 2019. Unsupervised deformable registration for multi-modal images via disentangled representations. In: Proc. International Conference on Information Processing in Medical Imaging (IPMI), pp. 249–261.

Reed, S.E., Zhang, Y., Zhang, Y., Lee, H., 2015. Deep visual analogy-making. In: Proc. Advances in Neural Information Processing Systems (NeurIPS).

Reinhold, J.C., Carass, A., Prince, J.L., 2021. A structural causal model for MR images of multiple sclerosis. In: Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 782–792.

Ren, X., Yang, T., Wang, Y., Zeng, W., 2021. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In: Proc. International Conference on Learning Representations (ICLR).

Rezende, D., Mohamed, S., 2015. Variational inference with normalizing flows. In: Proc. International Conference on Machine Learning (ICML), pp. 1530–1538.

Rezende, D.J., Mohamed, S., Wierstra, D., 2014. Stochastic backpropagation and approximate inference in deep generative models. In: Proc. International Conference on Machine Learning (ICML), pp. 1278–1286.

Ridgeway, K., Mozer, M.C., 2018. Learning deep disentangled embeddings with the F-statistic loss. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 185–194.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al., 2020. The future of digital health with federated learning. NPJ Digital Medicine 3 (1), 1–7.

Rolinek, M., Zietlow, D., Martius, G., 2019. Variational autoencoders pursue PCA directions (by accident). In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12406–12415.

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M., 2016. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3234–3243.

Ruta, D., Motiian, S., Faieta, B., Lin, Z., Jin, H., Filipkowski, A., Gilbert, A., Collomosse, J., 2021. ALADIN: All layer adaptive instance normalization for fine-grained style similarity. arXiv:2103.09776. [Online]. Available: https://arxiv.org/abs/2103.09776.

Sankar, A., Keicher, M., Eisawy, R., Parida, A., Pfister, F., Kim, S.T., Navab, N., 2021. GLOWin: A flow-based invertible generative framework for learning disentangled feature representations in medical images. arXiv:2103.10868 Available: https://arxiv.org/abs/2103.10868.

Sarhan, M.H., Navab, N., Eslami, A., Albarqouni, S., 2020. Fairness by learning orthogonal disentangled representations. In: Proc. European Conference on Computer Vision (ECCV). Springer, pp. 746–761.

Schott, L., von Kügelgen, J., Träuble, F., Gehler, P., Russell, C., Bethge, M., Schölkopf, B., Locatello, F., Brendel, W., 2021. Visual representation learning does not generalize strongly within the same domain. In: Proc. International Conference on Learning Representations (ICLR).

Schlkopf, B., Locatello, F., Bauer, S., Ke, N.R., Kalchbrenner, N., Goyal, A., Bengio, Y., 2021. Toward causal representation learning. Proceedings of the IEEE 109 (5).

Sermesant, M., Delingette, H., Cochet, H., Jaïs, P., Ayache, N., 2021. Applications of artificial intelligence in cardiovascular imaging. Nature Reviews Cardiology 1–10.

Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., Zhang, T., 2020. Disentangled generative causal representation learning. preprint arXiv:2010.02637.

Shen, X., Zhang, T., Chen, K., 2020. Bidirectional generative modeling using adversarial gradient estimation. preprint arXiv:2002.09161.

Shen, Y., Zhou, B., 2021. Closed-form factorization of latent semantics in GANs. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1532–1540.

Shen, Z., Liu, J., He, Y., Zhang, X., Xu, R., Yu, H., Cui, P., 2021. Towards out-of-distribution generalization: A survey. preprint arXiv:2108.13624.

Shin, S.Y., Lee, S., Summers, R.M., 2021. Unsupervised domain adaptation for small bowel segmentation using disentangled representation. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 282–292.

Stone, A., Wang, H., Stark, M., Liu, Y., Scott Phoenix, D., George, D., 2017. Teaching compositionality to CNNs. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5058–5067.

Su, R., Liu, X., Tsaftaris, S.A., 2021. Why patient data cannot be easily. forgotten?.

Sudlow, C., et al., 2015. Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. Plos med 12 (3), e1001779.

Suter, R., Miladinovic, D., Schölkopf, B., Bauer, S., 2019. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In: Proc. International Conference on Machine Learning (ICML), pp. 6056–6065.

Székely, G.J., Rizzo, M.L., Bakirov, N.K., et al., 2007. Measuring and testing dependence by correlation of distances. The Annals of Statistics 35 (6), 2769–2794.

Tang, F., Zhu, X., Hu, J., Tie, J., Zhou, J., Fu, Y., 2022. Generative adversarial unsupervised image restoration in hybrid degradation. scenes.

Tang, Y., Tang, Y., Zhu, Y., Xiao, J., Summers, R.M., 2021. A disentangled generative model for disease decomposition in chest X-rays via normal image synthesis. Medical Image Analysis 67, 101839.

Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam–CAN, Henson, R.N., 2017. The Cambridge centre for ageing and neuroscience (Cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. NeuroImage 144, 262–269.

Thermos, S., Liu, X., O'Neil, A., Tsaftaris, S.A., 2021. Controllable cardiac synthesis via disentangled anatomy arithmetic. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Thomas, V., Pondard, J., Bengio, E., Sarfati, M., Beaudoin, P., Meurs, M.-J., Pineau, J., Precup, D., Bengio, Y., 2017. Independently controllable features. arXiv:1708.01289. [Online]. Available: https://arxiv.org/abs/1708.01289.

Tishby, N., Pereira, F.C., Bialek, W., 1999. The information bottleneck method. In: Proc. Annual Allerton Conference on Communication, Control and Computing.

Tomar, D., Zhang, L., Portenier, T., Goksel, O., 2021. Content-preserving unpaired translation from simulated to realistic ultrasound images. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 659–669.

Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., Schölkopf, B., Bauer, S., 2021. On disentangled representations learned from correlated data. In: Proc. International Conference on Machine Learning (ICML). PMLR, pp. 10401–10412.

Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., Schölkopf, B., Bauer, S., 2021. On disentangled representations learned from correlated data. In: Proc. International Conference on Machine Learning (ICML), pp. 10401–10412.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4105–4113.

Valvano, G., Chartsias, A., Leo, A., Tsaftaris, S.A., 2019. Temporal consistency objectives regularize the learning of disentangled representations. In: Proc. MICCAI Workshop on Domain Adaptation and Representation Transfer (DART), pp. 11–19.

Van Den Oord, A., Vinyals, O., et al., 2017. Neural discrete representation learning. Advances in neural information processing systems 30.

Vapnik, V.N., 1999. An overview of statistical learning theory. IEEE Transactions on Neural Networks 10 (5), 988–999.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, u., Polosukhin, I., 2017. Attention is all you need. In: Proc. Advances in Neural Information Processing Systems (NeurIPS).
Wang, C., Sun, X., Zhang, F., Yu, Y., Wang, Y., 2021. Dae-gcn: Identifying disease-related features for disease prediction. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 43–52.
Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., 2021. Generalizing to unseen domains: A survey on domain. generalization.
Wang, R., Chaudhari, P., Davatzikos, C., 2021. Harmonization with flow-based causal inference. In: Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 181–190.
Wang, R., Zheng, G., 2021. Unsupervised cross-modality cardiac image segmentation via disentangled representation learning and consistency regularization. In: Machine Learning in Medical Imaging. Springer, pp. 517–526.
Wang, R., Zheng, G., 2022. CyCMIS: Cycle-consistent cross-domain medical image segmentation via diverse image augmentation. Medical Image Analysis 76, 102328.
Wang, T., Yue, Z., Huang, J., Sun, Q., Zhang, H., 2021. Self-supervised learning disentangled group representation as feature. In: Proc. Advances in Neural Information Processing Systems (NeurIPS).
Watanabe, S., 1960. Information theoretical analysis of multivariate correlation. IBM Journal of research and development 4 (1), 66–82.
Xia, T., Chartsias, A., Tsaftaris, S.A., 2019. Adversarial pseudo healthy synthesis needs pathology factorization. In: Proc. International Conference on Medical Imaging with Deep Learning (MIDL), pp. 512–526.
Xia, T., Chartsias, A., Tsaftaris, S.A., 2020. Pseudo-healthy synthesis with pathology disentanglement and adversarial learning. Medical Image Analysis 64, 101719.
Xia, T., Chartsias, A., Tsaftaris, S.A., Initiative, A.D.N., et al., 2019. Consistent brain ageing synthesis. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 750–758.
Xianjing, L., E, B.E., J, N.W., B, W.E., V, R.G., Li, 2021. Projection-wise disentangling for fair and interpretable representation learning: Application to 3d facial shape analysis, pp. 814–823.
Xiao, T., Hong, J., Ma, J., 2018. ELEGANT: Exchanging latent encodings with GAN for transferring multiple face attributes. In: Proc. European Conference on Computer Vision (ECCV), pp. 172–187.
Xie, X., Chen, J., Li, Y., Shen, L., Ma, K., Zheng, Y., 2020. MI2GAN: Generative adversarial network for medical image domain adaptation using mutual information constraint. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 516–525.
Xu, X., Wang, T., Shi, Y., Yuan, H., Jia, Q., Huang, M., Zhuang, J., 2019. Whole heart and great vessel segmentation in congenital heart disease using deep neural networks and graph matching. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 477–485.
Yang, F., Meng, R., Cho, H., Wu, G., Kim, W.H., 2021. Disentangled sequential graph autoencoder for preclinical Alzheimer's disease characterizations fromADNIstudy. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 362–372.
Yang, F., Meng, R., Cho, H., Wu, G., Kim, W.H., 2021. Disentangled sequential graph autoencoder for preclinical Alzheimer's disease characterizations from ADNI study. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer.
Yang, J., Dvornek, N.C., Zhang, F., Chapiro, J., Lin, M., Duncan, J.S., 2019. Unsupervised domain adaptation via disentangled representations: Application to cross–modality liver segmentation. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 255–263.
Yang, J., Li, X., Pak, D., Dvornek, N.C., Chapiro, J., Lin, M., Duncan, J.S., 2020. Cross–modality segmentation by self-supervised semantic alignment in disentangled content space. In: Proc. MICCAI Workshop on Domain Adaptation and Representation Transfer (DART), pp. 52–61.
Zaidi, J., Boilard, J., Gagnon, G., Carbonneau, M.-A., 2020. Measuring disentanglement: A review of metrics. arXiv:2012.09276. [Online]. Available: https://arxiv.org/abs/2012.09276.
Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S., 2021. Barlow twins: Self-supervised learning via redundancy reduction. arXiv:2103.03230. [Online]. Available: https://arxiv.org/abs/2103.03230.
Zhang, R., Pfister, T., Li, J., 2019. Harmonic unpaired image-to-image translation. International Conference on Learning Representations (ICLR).
Zhao, Q., Adeli, E., Honnorat, N., Leng, T., Pohl, K.M., 2019. Variational autoEncoder for regression: Application to brain aging analysis. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 823–831.
Zhao, Q., Adeli, E., Pohl, K.M., 2021. Longitudinal correlation analysis for decoding multi-modal brain development. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 400–409.
Zhao, Q., Liu, Z., Adeli, E., Pohl, K.M., 2021. Longitudinal self-supervised learning. Medical Image Analysis 71, 102051.
Zhao, Q., Sullivan, E.V., Honnorat, N., Adeli, E., Podhajsky, S., De Bellis, M.D., Voyvodic, J., Nooner, K.B., Baker, F.C., Colrain, I.M., et al., 2021. Association of heavy drinking with deviant fiber tract development in frontal brain systems in adolescents. JAMA psychiatry 78 (4), 407–415.
Zhou, L., Bae, J., Liu, H., Singh, G., Green, J., Gupta, A., Samaras, D., Prasanna, P., 2022. Lung swapping autoencoder: Learning a disentangled structure-texture representation of chest radiographs. preprint arXiv:2201.07344.
Zhou, L., Bae, J., Liu, H., Singh, G., Green, J., Samaras, D., Prasanna, P., 2021. Chest radiograph disentanglement for COVID-19 outcome prediction. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, pp. 345–355.
Zhou, S., Xiao, T., Yang, Y., Feng, D., He, Q., He, W., 2017. GeneGAN: Learning object transfiguration and attribute subspace from unpaired data. In: Proc. British Machine Vision Conference (BMVC).
Zhu, J.-Y., Krähenbühl, P., Shechtman, E., Efros, A.A., 2016. Generative visual manipulation on the natural image manifold. In: Proc. European Conference on Computer Vision (ECCV), pp. 597–613.
Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2223–2232.
Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E., 2017. Toward multimodal image-to-image translation. In: Proc. Advances in Neural Information Processing Systems (NeurIPS), pp. 465–476.
Zhuang, X., 2013. Challenges and methodologies of fully automatic whole heart segmentation: A review. Journal of Healthcare Engineering 4 (3), 371–407.
Zhuang, X., 2016. Multivariate mixture model for cardiac segmentation from multi--sequence mri. In: Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 581–588.
Zhuang, X., 2019. Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (12), 2933–2946.
Zhuang, X., Rhode, K.S., Razavi, R.S., Hawkes, D.J., Ourselin, S., 2010. A registration-based propagation framework for automatic whole heart segmentation of cardiac MRI. IEEE Transactions on Medical Imaging 29 (9), 1612–1625.
Zhuang, X., Shen, J., 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. Medical Image Analysis 31, 77–87.
Zhuang, X., Shen, J., 2016. Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Medical image analysis 31, 77–87.
Zimmermann, R.S., Sharma, Y., Schneider, S., Bethge, M., Brendel, W., 2021. Contrastive learning inverts the data generating process. In: Meila, M., Zhang, T. (Eds.), Proceedings of the 38th International Conference on Machine Learning. PMLR, pp. 12979–12990.
Zou, K., Faisan, S., Heitz, F., Epain, M., Croisille, P., Fanton, L., Valette, S., 2021. Disentangled representations: Towards interpretation of sex determination from hip bone. arXiv.
Zuo, L., Dewey, B.E., Carass, A., Liu, Y., He, Y., Calabresi, P.A., Prince, J.L., 2021. Information-based disentangled representation learning for unsupervised mr harmonization. Lecture Notes in Computer Science 12729 LNCS, 346–359.
Zuo, L., Dewey, B.E., Liu, Y., He, Y., Newsome, S.D., Mowry, E.M., Resnick, S.M., Prince, J.L., Carass, A., 2021. Unsupervised mr harmonization by learning disentangled representations using information bottleneck theory. NeuroImage 243.