Unsupervised Learning of Disentangled Representation via Auto-Encoding: A Survey

Ikram Eddahmani, Chi-Hieu Pham, Thibault Napoléon, Isabelle Badoc, Jean-Rassaire Fouefack, Marwa El-Bouz

sensors

MDPI

Review

Unsupervised Learning of Disentangled Representation via Auto-Encoding: A Survey

Ikram Eddahmani [1,2], Chi-Hieu Pham [1,3,*], Thibault Napoléon [4], Isabelle Badoc [2] and Jean-Rassaire Fouefack [1] and Marwa El-Bouz [1]

1 L@bISEN, LSL Team, Yncrea Ouest, 29200 Brest, France
2 Generix Group, 75012 Paris, France
3 LaTIM, INSERM UMR1101, University of Brest, 29200 Brest, France
4 L@bISEN, VISION-AD Team, Yncrea Ouest, 29200 Brest, France
* Correspondence: chi-hieu.pham@isen-ouest.yncrea.fr

Abstract: In recent years, the rapid development of deep learning approaches has paved the way to explore the underlying factors that explain the data. In particular, several methods have been proposed to learn to identify and disentangle these underlying explanatory factors in order to improve the learning process and model generalization. However, extracting this representation with little or no supervision remains a key challenge in machine learning. In this paper, we provide a theoretical outlook on recent advances in the field of unsupervised representation learning with a focus on auto-encoding-based approaches and on the most well-known supervised disentanglement metrics. We cover the current state-of-the-art methods for learning disentangled representation in an unsupervised manner while pointing out the connection between each method and its added value on disentanglement. Further, we discuss how to quantify disentanglement and present an in-depth analysis of associated metrics. We conclude by carrying out a comparative evaluation of these metrics according to three criteria, (i) modularity, (ii) compactness and (iii) informativeness. Finally, we show that only the Mutual Information Gap score (MIG) meets all three criteria.

Keywords: representation learning; disentanglement; auto-encoder; generative models; neural networks; metrics

Citation: Eddahmani, I.; Pham, C.-H.; Napoléon, T.; Badoc, I.; Fouefack, J.-R.; El-Bouz, M. Unsupervised Learning of Disentangled Representation via Auto-Encoding: A Survey. *Sensors* **2023**, *23*, 2362. https://doi.org/10.3390/s23042362

Academic Editor: Paolo Gastaldo

Received: 14 December 2022
Revised: 11 February 2023
Accepted: 16 February 2023
Published: 20 February 2023

1. Introduction

Data representation is a crucial and long-standing issue in machine learning, as it has a significant impact on model performance [1]. For that reason, much of the actual efforts in the machine learning community have been toward representation learning [2–5].

Representation learning refers to finding a low-dimensional representation that captures true underlying factors of variation that explain the data [6]. A series of traditional statistical approaches have been reported for the estimation of such a low-dimensional representation, such as Principal Component Analysis (PCA) [7,8], Independent Component Analysis (ICA) [9,10] or Single Value Decomposition (SVD) [11]. They aim to identify the underlying factors of variation in the data. However, in practice, the data to be encoded may be very large in dimension and contain factors that cannot be captured with these linear methods.

Disentangled representation learning has emerged as an effective way of finding a low-dimensional space of complex data while addressing the problem of identifying the independent factors of variation. Following the definition of Bengio et al. [3]: "a disentangled representation is a representation where a change in one latent variable corresponds to a change in one generative factor, while being relatively invariant to changes in other factors". As an example, a model trained on a set of face images can capture different generative factors, i.e., pose, gender, skin color or smile, and encode them into independent latent variables in the representation space. Each latent variable is sensitive to

a change in only one generative factor. Let z_k be the latent factor controlling the facial pose, then varying z_k while fixing other factors would generate images of different facial poses but with the same other generative factors (gender, skin, color, smile). The same goes for latent variables controlling other generative factors. We illustrate the notation of generative factors and latent factors in Figure 1.

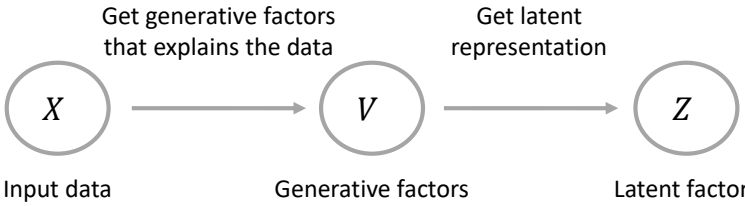

Figure 1. An illustration of the notation used in this paper. For $X = \{x_1, x_2, \ldots, x_N\}$, a set of N observations. Disentangled representation learning is expected to identify the distinct generative factors $V = \{v_1, v_2 \ldots, v_n\}$ that explain these observations and encode them with independent latent variables $Z = \{z_1, \ldots, z_n\}$ in latent space.

These representations have been useful for model generalization by discovering the causal variables in data and capturing its compositional structure [12–15], allowing the learning systems to understand real-world observations as humans do [16], and, therefore, representation learning could generalize to unseen scenarios. For example, a model trained to generate an image of a green square and blue triangle, because of the generalizability property, the model can also generate a blue square and green triangle [17]. Following this motivation, they have been of interest to downstream tasks, such as supervised learning, compression and data augmentation. In supervised learning, it can be used as an input feature when building classifiers or other predictors to improve predictive performance [18], reduce sample complexity [19] and offer interpretability [20]. For compression [21], disentangled representations are compact and low-dimensional, thus minimizing the cost associated with storing underlying factors of variation in data. Further, they can be used to generate novel examples not found in the original dataset [21]. Such feature learning also supports a variety of other applications, such as super-resolution [22], multimodal application [23–27], medical imaging [28,29], video prediction [30–34], natural language processing [35–37], transfer learning and zero-shot learning [38].

A large range of state-of-the-art methods for learning unsupervised representations is based on variational auto-encoders (VAE) [39]. Variational auto-encoders have been shown to be useful for learning high-dimensional data and inferring latent variables. However, they often fail to capture a disentangled representation of the data [20]. In order to overcome these drawbacks, several variants of VAE have been proposed [40–44] with the idea that they could allow better disentanglement [45,46]. Another line of work in this field is based on Generative Adversarial Networks (GAN) [47,48]. Numerous variants of GAN have been proposed and demonstrated the ability to learn a disentangled representation [49–53] and were reported to have comparable performance to VAE-based methods [53]. A relative field with disentangled representation learning is Self-Supervised Learning (SSL) [54–56]. Self-supervised learning provides a way for learning representation from unlabeled data. Recent efforts have been made toward using self-supervised algorithms in order to learn a disentangle representation [57–60]. However, recent studies have reported that the existing SSL methods often struggle to learn disentangled representations of the data [60].

In this paper, we aim to provide a systematic and comprehensive survey of VAE-based approaches. We attempt to shed light on some VAE variants that are considered state-of-the-art of disentangled representations and to provide an analysis of the common idea behind all these approaches. Further, we conduct an extensive review of disentanglement metrics, where we explore what makes one metric better than another. Finally, we discuss limitations and future directions in this field of research. We sum up the recent work focusing on disentanglement and the metric proposed alongside each method in Table 1.

Table 1. Overview of disentanglement methods based on auto-encoding and the metric reported with each method.

Methods	Metrics
β-VAE [20]	Z-diff Score
Factor-VAE [42]	Z-min Variance Score
β-TCVAE [41]	Mutual Information Gap (MIG)
DIP-VAE [43]	Attribute Predictability Score (SAP)
InfoMax-VAE [44]	-

To the authors' best knowledge, [61] is the only study focusing on disentanglement representation methods from a practical point of view. In [61], the authors introduce a library to train and evaluate disentangled representations. However, a detailed description of the methods is missing. On the other hand, no other study performs this type of review focused on grouping disentanglement methods as well as metrics and presents a theoretical analysis and a detailed description of each method and their added value. The purpose of this survey is to provide researchers interested in this broad field with a comprehensive overview of state-of-the-art approaches and establish a guideline to choose a suitable approach given an objective.

The rest of this paper is organized into six sections. In Section 2, we describe the main concepts that are necessary to understand the methods considered in this work. In Section 3, we review unsupervised disentanglement methods based on the auto-encoder baseline. In Section 4, we review the most well-known metrics to evaluate disentanglement. In Section 5, we present a detailed discussion of disentanglement methods and a comparison between metrics. Finally, we conclude the work and discuss its future scope in Section 6.

2. Background

In this section, we provide a detailed description of several notions that will be found throughout this paper.

2.1. Auto-Encoders

An auto-encoder [1,3,62–64] is a neural network architecture that is trained to reconstruct its input [65] with the least possible amount of distortion. Their main purpose is to learn a compressed meaningful representation of the data that can be used for various applications, including clustering [66] and classification [67,68].

Here we briefly describe the auto-encoder (AE) framework:

Encoder: A neural network f that maps an input (image, tensor, curve) into a hidden representation Z capturing the significant underlying factors of the data, also called an inference model. Given a data set, $\{x_1, \ldots, x_T\}$, for each x_i, we define:

$$z_i = f(x_i), \tag{1}$$

Latent space: A low-dimensional representation of the data. The vector z is the feature-vector, also called latent code or latent dimension. z is called "latent" because it is a variable produced by the model from the input data.

Decoder: A neural network g that builds back the input from its latent vectors.

$$\widetilde{x_i} = g(z_i), \tag{2}$$

Training an auto-encoder consists of learning the functions f and g that minimize the error $\mathbb{E}$ of the reconstruction loss function $\triangle$, which measures the difference between the input and its reconstruction:

$$\arg\min_{f,g} \mathbb{E}[\triangle(x_i, g(f(x_i)))] \tag{3}$$

2.2. Variational Auto-Encoders (VAEs)

Kingma et al. [39] introduce a stochastic variational inference for an auto-encoder. Variational auto-encoders attempt to describe data generation through a probabilistic modeling perspective [69]. In VAE, inputs are encoded as a distribution over latent space instead of as single points [70]. In doing so, Kingma et al. assume a posterior distribution on the latent variable z_i for each data point x_i denoted by inference model $q_\phi(z|x)$. This inference model corresponds to the probabilistic encoder, and is parameterized by ϕ [65]. In a similar vein, they introduce a generative model $p_\theta(x|z)$, which is equivalent to a probabilistic decoder determined by the parameter θ: given a latent variable z it returns a distribution on the corresponding possible values x. Finally, they consider a prior distribution over the latent variables z_i denoted by $p_\theta(z_i)$, where $p_\theta(z)$ is a standard multivariate normal distribution $\mathcal{N}(0, I)$ and I is the identity matrix. Training a VAE consists of simultaneously learning the parameters ϕ and θ. One way to estimate these parameters is to use the maximum log-likelihood (ML), a common criterion for probabilistic models.

The marginal log-likelihood is the sum of each data points $\log p_\theta(x_1, \ldots, x_N) = \sum_{i=1}^{N} \log p_\theta(x_i)$. Each point can be rewritten as [65]:

$$\log p_\theta(x_i) = \mathrm{D}_{KL}\big(q_\phi(z|x_i)||p_\theta(z|x_i)\big) + \mathcal{L}(\theta, \phi, x_i) \tag{4}$$

The first term in Equation (4) is the Kullback–Leibler (KL) divergence, which determines the distance between the approximate posterior and the true posterior, and the second term is the variational lower bound defined in [69] as:

$$\mathcal{L}(\theta, \phi, x_i) = \mathbb{E}_{q_\phi(z|x_i)}\big[-\log q_\phi(z|x) + \log p_\theta(x, z)\big] \tag{5}$$

Since the KL divergence is non-negative, $\mathcal{L}(\theta, \phi, x_i)$ is the lower marginal log-likelihood bound, also known as the Evidence Lower Bound Objective (ELBO). This can be further expressed as:

$$\mathcal{L}_{VAE} \simeq \mathcal{L}(\theta, \phi, x_i) = \mathbb{E}_{q_\phi(z|xi)}[\log p_\theta(x_i|z)] - \mathrm{D}_{\mathrm{KL}}\big(q_\phi(z|x_i)||p_\theta(z)\big) \tag{6}$$

Relying on Equation (6), one can notice that the evidence lower bound is a sum of two terms: the first term is a negative reconstruction error that needs to be maximized in order to increase the reconstruction capability of the sample, and the second term is the KL divergence that acts as a regularizer to ensure that the approximate posterior $q_\phi(z|x)$ remains close to the prior. Finding the model parameters θ and ϕ that will maximize the marginal likelihood of the data while simultaneously minimizing the KL divergence between the approximation $q_\phi(z|x)$ and the prior $p_\theta(z)$ is equivalent to maximizing the ELBO.

It can thus be said that training a variational auto-encoder consists of maximizing the variational lower bound objective. The ELBO (Equation (6)) serves as the core of the variational auto-encoder and the methods we will discuss in the rest of this paper, so it is worth spending some thinking about how it can be optimized [71]:

$$\max_{\phi,\theta} \mathbb{E}_{p(x_i)}\Big[\mathbb{E}_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)] - D_{KL}(q_\phi(z|x_i)||p_\theta(z))\Big] \tag{7}$$

As is common in machine learning, the ELBO can be optimized with regard to all parameters (ϕ and θ) using stochastic gradient descent [72], but first, more detail about $q_\phi(z|x_i)$ is required. The usual choice is a simple factorized Gaussian encoder $q_\phi(z|x_i) \sim \mathcal{N}\big(\mu_\phi(x_i), \sigma_\phi(x_i)\big)$, where $\mu_\phi(x_i)$ and $\sigma_\phi(x_i)$ are the mean and standard deviation implemented via neural networks and $\sigma_\phi(x_i)$ is constrained to be a diagonal matrix. Under this choice, we have a KL divergence between two Gaussian distributions, which is tractable. Hence we can calculate the gradient of the last term of Equation (7). However, the first term is a bit trickier as it requires an estimation by sampling from $q_\phi(z|x_i)$. Kingma et al. [39,72] propose to estimate the marginal likelihood lower bound of the full dataset using mini-batches of M data-points and then average the gradient over these mini-

batches. However, stochastic gradient descent via back-propagation cannot handle stochastic variables within the network. To solve this problem and generalize back-propagation through random sampling, they propose another way to generate samples from $q_\phi(z|x_i)$; this solution is called the "reparameterization trick" [39,73]. The key behind this trick is to define z as a deterministic function $z = g(\phi, x_{i,\xi})$, then sample from the posterior $q_\phi(z|x_i)$ using $z = \mu(x_i) + \sigma(x_i) * \xi$. Here, ξ is an auxiliary variable with independent marginal $p(\xi) \sim \mathcal{N}(0, I)$, and $g_\phi(.)$ is a vector-valued function parameterized by ϕ. In doing so, we are keeping the stochasticity of the variables, but also, we have a deterministic function of inputs that will work for stochastic gradient descent. The architecture of a variational auto-encoder (VAE) is shown in Figure 2.

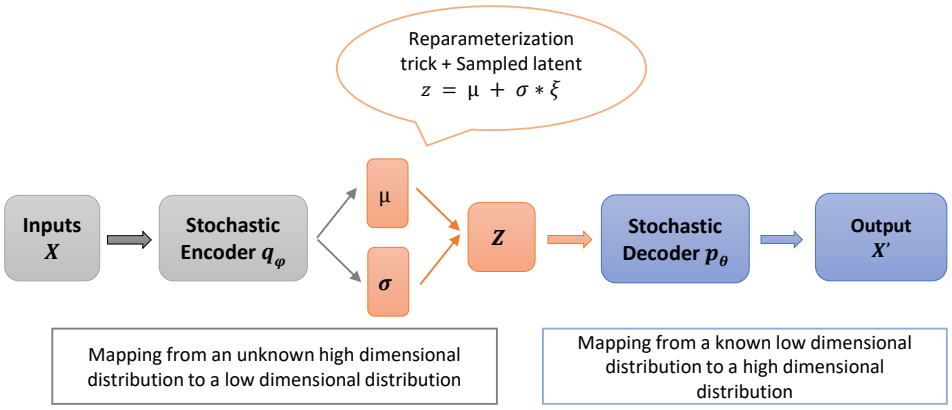

Figure 2. The structure of the variational auto-encoder (VAE). The stochastic encoder $q_\phi(z|x_i)$, also called the inference model, learns stochastic mappings between an observed X-space (input data) and a latent Z-space (hidden representation). The generative model $p_\theta(z|x_i)$, a stochastic decoder, reconstructs the data given the hidden representation.

For ease of reference, we sum up in Table 2 the main terms and corresponding mathematical symbols used in this work.

Table 2. Overview of the main terms and corresponding mathematical expression.

Term	Mathematical Expression
Prior	$p_\theta(z)$
Generative Model (Decoder)	$p_\theta(z\|x_i)$
Inference Model (Encoder)	$q_\phi(z\|x_i)$
Data Log-Likelihood	$\log p_\theta(x_i)$
Kullback–Leibler Divergence	$D_{KL}(q_\phi(z\|x_i)\|\|p_\theta(z))$
Evidence Lower Bound (ELBO)	$\mathbb{E}_{q_\phi(z\|xi)}[\log p_\theta(x_i\|z)] - D_{KL}(q_\phi(z\|x_i)\|\|p_\theta(z))$

2.3. Reconstruction Error

In order to make the optimization of the evidence lower bound objective (Equation (7)), the posterior $q_\phi(z|x_i)$ is pushed to match the unit Gaussian prior $p_\theta(z) \sim \mathcal{N}(0, I)$. Since the posterior $q_\phi(z|x_i)$ and the prior $p_\theta(z)$ are factorized (i.e., have diagonal covariance matrix) and the samples from $q_\phi(z|x_i)$ are generated using the reparameterization trick, learning a representation of the data depending only on $q_\phi(z|x_i)$ may result in a meaningless representation where only a limited number of latent variables are exploited for data reconstruction. In doing so, the amount of information that can be transmitted through the latent channels is reduced. Thus, this results in high reconstruction errors and low reconstruction fidelity [74].

2.4. Mutual Information Theory

Mutual information is a fundamental quantity for measuring dependency between random variables [75]. Let (X, Z) be a couple of random variables. The mutual information (MI) between X and Z, denoted as $I(X;Z)$, is:

$$I(X;Z) = D_{KL}(P_{XZ}||P_X \otimes P_Z) \tag{8}$$

where D_{KL} is the Kullback–Leibler (KL) divergence between the joint distribution and the product of the marginals.

3. Methods

The major challenge behind representation learning is how we can choose the model that leads to better disentanglement and thus can be useful for later downstream tasks. In this section, we present an overview of the state-of-the-art frameworks in representation learning based on auto-encoding (see Table 1).

3.1. β-Variational Auto-Encoder

Higgins et al. [20] introduce a variant of variational auto-encoders [39], β-VAE, a deep generative (unsupervised) algorithm for learning disentangled representations. The authors propose to modify the evidence lower bound objective (ELBO) by up weighting the KL divergence term in Equation (7) in order to learn a disentangled representation:

$$\max_{\phi,\theta} \mathbb{E}_{p(x_i)}\Big[\mathbb{E}_{q_\phi(z|x_i)}[\log p_\theta(x_i|z)] - \beta D_{KL}\big(q_\phi(z|x_i)||p_\theta(z)\big)\Big] \tag{9}$$

where β is an adjustable hyper-parameter higher than 1. It can be noted that β-VAE with $\beta = 1$ is equivalent to the original VAE framework [39]. Re-writing the ELBO for β-VAE:

$$\mathcal{L}_{\beta\text{-VAE}} = \mathbb{E}_{q_\phi(z|x_i)}\big[\log p_\theta(x_i|z)] - \beta D_{KL}\big(q_\phi(z|x_i)||p_\theta(z)\big)\big] \tag{10}$$

Such a penalization causes the posterior $q_\phi(z|x)$ to better match the factorized prior $p_\theta(z)$, which is associated with the need to maximize the log-likelihood of data x and push the model to learn a disentangled representation of the data.

β-VAE was performed using a number of benchmarks with known ground truth factors, such as CelebA [76] (202,599 color images of celebrity faces), Chairs [77] (86,366 color images of chairs), Faces [78] (239,840 gray-scale images of 3D faces) and 2D Shape [79] (737,280 synthetic images of 2D shapes such as heart, oval and square). The size of the images in the four datasets is 64×64 pixels.

3.2. InfoMax-Variational Auto-Encoder

As we can see from Equation (10), learning representation depending only on $q_\phi(z|x)$ may result in meaningless representations and, therefore, a poor reconstruction quality. To avoid collapsed representations, Rezaabad et al. [44] report a simple yet practical method to build a meaningful representation. They propose to extend the evidence lower bound (ELBO) with a regularizer term that maximizes the mutual information between the data and the latent representation. In so doing, the model is pushed to maximize the information about the data (input) stored in the inferred representation (latent representation) [44]. This solution is referred to as InfoMax-VAE. They end up with the following ELBO:

$$\max_{\phi,\theta} \mathbb{E}_{q(x)}\big[\mathcal{L}_{\beta\text{-VAE}}\big] + \alpha I_{q_\phi}(x,z) \tag{11}$$

where β and $\alpha \geq 0$ are regularization coefficients for the KL divergence and mutual information. Varying α controls the amount of information stored in the latent representation, also known as information preference.

Following [80], the mutual information is estimated by the average KL divergence between the joint and associated marginals: $(I_{q_\phi}(x,z) = KL\big(q_\phi(x;z)||q(x) \otimes q_\phi(z)\big)$. However, the KL divergence is hard to compute in general due to the intractable posterior, so Rezaabad et al. argue [44] that since KL divergence comes from a large class of different divergence, we can replace this term with another variational f-divergence $D_{f(t)}$ where t represents all possible functions. Thus, they arrive at the following equation:

$$\max_{\phi,\theta} \mathbb{E}_{q(x)}\left[\mathcal{L}_{\beta\text{-VAE}}\right] + \alpha D_f\big(q_\phi(x,z)||q(x)q_\phi(z)\big) \tag{12}$$

Specifically, they choose $f(t)$ to be $t \log t$ (more details about this choice can be found in [44]), arriving to the final ELBO for InfoMax-VAE:

$$\mathcal{L}_{\text{InfoMax-VAE}} = \mathcal{L}_{\beta\text{-VAE}} + \alpha\Big(\mathbb{E}_{q_\phi(x,z)}[t(x,z)] - \mathbb{E}_{q(x)q_\phi(z)}[\exp(t(x,z)-1)]\Big) \tag{13}$$

This leaves the task of evaluating $\mathbb{E}_{q_\phi(x,z)}$ and $\mathbb{E}_{q(x)q(z)}$. They propose a simple yet practical way to do so: first of all, and thanks to the reparameterization trick [72], they draw samples from $q(x)$, $(x_i, z_i) \sim q_\phi(x,z) = q_\phi(z|x)q(x)$. Afterward, to get samples from the marginal $q_\phi(z)$, they choose a random data point x_j, followed by sampling from $z \sim q_\phi(z|x_j)$.

The InfoMax-VAE was performed using the CelebA dataset and has been shown to be capable of learning meaningful and disentangled representation and outperforms β-VAE.

3.3. Factor Variational Auto-Encoder

Kim and Mnih [42] adopt another decomposition of the ELBO, specifically, they decompose the KL term in Equation (6) as Hoffman and Johnson propose in [81,82]:

$$\mathbb{E}_{p_{data}(x)}[D_{KL}(q(z|x)||p(z))] = I(x;z) + D_{KL}(q(z)||p(z)) \tag{14}$$

where $I(x;z)$ is the mutual information between x and z and $q(z) = E_{p_{data(x)}}[q(z|x)] = \frac{1}{N}\sum_{i=1}^{N} q(z|x_i)$ is the latent distribution for all data. Penalizing $D_{KL}[q(z)||p(z)]$ encourages $q(z)$ to match the factorized prior $p(z)$ and, therefore, encourages disentanglement. Further, penalizing $I(x;z)$ reduces the information about the data x kept in the latent space z, which might result in less accurate reconstructions for high values of β.

Based on this observation, Kim and Mnih [42] argue that it may not be necessary or desirable to penalize the mutual information between x and z in order to have a better disentanglement. However, they propose to add an additional term to the VAE objective (Equation (6)) that penalizes the dependence of variables within the latent space [46]:

$$E_{p(x)}[E_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z))] - \gamma D_{KL}(q(z))||\prod_{i=1}^{d} q(z_j) \tag{15}$$

Re-writing the ELBO for Factor-VAE:

$$\mathcal{L}_{\text{Factor-VAE}} = \mathcal{L}_{\text{VAE}} - \gamma D_{KL}(q(z)||\prod_{i=1}^{d} q(z_j)) \tag{16}$$

The second term is total correlation (TC)[83], a general measure of dependence between several random variables. This term is intractable since the estimation of both $q(z)$ and $q(z_j)$ requires a pass through the entire data set. Hence Kim and Mnih [42] propose another alternative for optimizing this term using the density ratio trick [84,85]. The density

ratio trick consists of training a binary classifier/discriminator that returns the probability $d(z)$ that its input was sampled from $q(z)$ rather than from $q(z_j)$:

$$TC(z) = D_{KL}(q(z)||q(z_j)) = \mathbb{E}_{q(z)]}[\log \frac{q(z)}{q(z_j)}] \simeq E_{q(z)}[\log \frac{d(z)}{1-d(z)}] \tag{17}$$

The VAE and the discriminator are trained jointly. The VAE parameters are fine-tuned using the objective in Equation (16) with the total correlation term changed to its approximation in Equation (17).

Factor-VAE was performed using several benchmarks, such as CelebA, Chairs and Faces. Moreover, two synthetic datasets were used: 2D Shapes and 3D Shapes containing 480,000 $64 \times 64 \times 3$ RGB images of 3D shapes [79].

3.4. β-Total Correlation Variational Auto-Encoder

Concurrently to Kim and Mnih [42], Chen et al. [41] proposed another approach that surpasses both β-VAE performance and Factor-VAE complexity. β-TCVAE is a deep unsupervised approach for learning disentangled representation, a replacement of β-VAE, with no additional hyper-parameters during training.

Chen et al. [41] proposed a different decomposition of the second term in Equation (10), arriving at the following split up:

$$\begin{aligned}\mathbb{E}_{p(x)}[D_{KL}(q_\phi(z|x)||p(z))] =& D_{KL}(q(z,x)||q(z)p(x)) + D_{KL}(q(z)||\prod_j q(z_j)) \\ &+ \sum_j D_{KL}(q(z_j)||p(z_j))\end{aligned} \tag{18}$$

The first term is known as the index-code mutual information (MI), which denotes the mutual information between the data and the latent space. The second term is total correlation (TC). The last term is the dimension-wise KL, which primarily encourages the latent dimensions to better match their corresponding priors.

According to Chen et al. [41], the total correlation term in the ELBO is the one that affects disentanglement. To verify this claim, the TC-term is evaluated using a Monte-Carlo approximation [86].

A unique integer index is assigned to each training sample, and they use the following estimator given a mini-batch of samples $\{n_1, n_2, \ldots, n_M\}$:

$$\mathbb{E}_{q(z)}[\log q(z)] \simeq \frac{1}{M}\sum_{i=1}^{M}\left[\log \frac{1}{NM}\sum_{j=1}^{M} q(z(n_i)|n_j)\right] \tag{19}$$

where $q(z|n_i)$ is close to 0 for a randomly sampled component but large if z comes from component n_i.

To achieve better disentanglement, the authors up-weight each term of the ELBO individually. Re-writing the β-TCVAE objective:

$$\mathcal{L}_{\beta\text{-TCVAE}} = \mathcal{L}_{\beta\text{-VAE}} - \alpha I_q(z;n) - \beta KL(q(z)||\prod_j q(z_j)) \tag{20}$$

β-TCVAE was performed using the same datasets as Factor-VAE (CelebA, Chairs, Faces and 3D Shapes).

3.5. DIP-Variational Auto-Encoder

Another line of work has argued that pushing the posterior $q_\phi(z)$ to match a factorized prior $p(z)$ can lead to a better disentanglement. Kumar et al. [43] added a regularizer to the ELBO to encourage disentanglement during inference, therefore:

$$\mathcal{L}_{\text{VAE}} - \lambda D(q(z)||p(z)) \tag{21}$$

where λ is a hyper-parameter controlling its effect on the evidence lower bound objective, and D is an (arbitrary) divergence. In order to estimate this term, they suggest matching the moments of these distributions. In particular, they propose to penalize the ℓ_2 distance between $q_\phi(z)$ and $\mathcal{N}(0,1)$ in order to match their covariances.

Let us denote:

$$Cov_{q_\phi(z)}[z] = E_{p(x)}Cov_{q_\phi(z|x)}[z] + Cov_{p(x)}(E_{q_\phi(z|x)}[z]) \tag{22}$$

where $E_{q_\phi(z|x)}[z]$ and $Cov_{q_\phi(z|x)}[z]$ are random variables that are functions of random variable x. Since $q_\phi(z|x) \sim N\big(\mu_\phi(x), \sum\sigma(x)\big)$, Equation (22) becomes:

$$Cov_{q_\phi(z)}[z] = E_{p(x)}\left[\sum\sigma_\phi(x)\right] + Cov_{p(x)}\left[\mu_\phi(x)\right] \tag{23}$$

Kumar et al. [43] explored two options for disentangling regularizers to get this term close to the identity matrix: (i) regularizing the deviation of $Cov_{p(x)}\left[\mu_\phi(x)\right]$ from the identity matrix, which they refer to as DIP-VAE-*I*. (ii) regularizing $Cov_{q_\phi(x)}[z]$, which they denote as DIP-VAE-*II*.

Maximizing the objective of either DIP-VAE-*I* Equation (24) or DIP-VAE-*II* Equation (25) leads to better disentanglement.

$$\mathcal{L}_{\text{DIP-VAE-I}} = \mathcal{L}_{\text{VAE}} - \lambda_1 \sum_{i\neq j}\left[Cov_{p(x)}\left[\mu_{\phi(x)}\right]\right]^2_{ij} - \lambda_2 \sum_i\left(\left[Cov_{p(x)}\left[\mu_{\phi(x)}\right]\right]_{ii} - 1\right)^2 \tag{24}$$

$$\mathcal{L}_{\text{DIP-VAE-II}} = \mathcal{L}_{\text{VAE}} - \lambda_1 \sum_{i\neq j}\left[Cov_{q_\phi}[z]_{ij}\right]^2 - \lambda_2 \sum_i\left(\left[Cov_{q_\phi}[z]\right]_{ii} - 1\right)^2 \tag{25}$$

DIP-VAE was performed using three datasets: CelebA, 3D Chairs and 2D Shapes. DIP-VAE has been shown to be superior to β-VAE and capable of learning disentangled factors without having any conflict between disentanglement and quality reconstruction.

To fairly evaluate such methods, the authors have chosen the same CNN architecture. Table 3 describes the experimental details, including the encoder and decoder architectures:

Table 3. Details on the encoder and decoder architecture used to implement methods discussed in this paper depending on the chosen dataset.

Datasets	Encoder	Decoder
2D Shape 3D Shape 3D Chairs 3D Faces	Input: 64 × 64 × number of channels, Conv 32 × 4 × 4 (stride 2), 32 × 4 × 4 (stride 2), 64 × 4 × 4 (stride 2), 64 × 4 × 4 (stride 2), FC 256, ReLU activation.	Input: R^N, FC, 256 ReLU, FC, 4 × 4 × 64 ReLU, Upconv 64 × 4 × 4 (stride 2), 32 × 4 × 4 (stride 2), 32 × 4 × 4 (stride 2), 4 × 4 × number of channels (stride 2), ReLU activation. **Bernoulli Decoder**
CelebA	Input: 64 × 64 × 3, Conv 32 × 4 × 4 (stride 2), 32 × 4 × 4 (stride 2), 64 × 4 × 4 (stride 2), 64 × 4 × 4 (stride 2), FC 256, ReLU activation	R^N, FC, 256 ReLU, FC, 4 × 4 × 64 ReLU, Upconv 64 × 4 × 4 (stride 2), 32 × 4 × 4 (stride 2), 32 × 4 × 4 (stride 2), 4 × 4 × number of channels (stride 2), ReLU activation. **Gaussian Decoder**

4. Metrics

In order to evaluate the approaches described above and use them for downstream tasks, a metric of disentanglement is required. Most prior works relied on a visual inspection of the latent representation [87], but recently more rigorous metrics have been proposed.

To the authors' best knowledge, [88,89] are the only studies analyzing disentanglement metrics. In this review, we choose to focus on discussing the metrics proposed alongside each method above Table 1), clarifying their strengths and shortcomings.

4.1. Z_{diff} Score

Higgins [20] introduced a disentanglement metric called Z-diff, also known as the β-VAE metric based on the following intuition: if one generative factor is fixed while randomly sampling all others, we will have a disentangled representation in which the latent variable corresponding to the fixed generative factor will vary less than the others. Applying this metric involves following these steps:

1. Randomly select a generative factor f_k.
2. Create a batch of couples vectors, p_1 and p_2, where the value of the chosen factor f_k is kept fixed and equal within the pair while the other generative factors f_{k-1} are chosen randomly. For a batch of L samples:

$$p_1 = (x_{1,1}, \ldots, x_{1,L}), p_2 = (x_{2,1}, \ldots, x_{2,L})$$

   with $x_{1,L} = x_{2,L}$
3. Map each generated pair to a pair of latent variables using the inference model $q(z|x) \sim N(\mu(x), \sigma(x))$.

$$z_{1,1} = \mu(x_{1,1}), z_{2,1} = \mu(x_{2,1})$$

4. Compute the value of the absolute linear difference between the variables related to the sample:

$$e = (|z_{1,1} - z_{2,1}|, \ldots, |z_{1,L} - z_{2,L}|)$$

5. The mean of all pair differences in a batch gives a single instance in the final training set. These steps are repeated for each generative factor in order to create a substantial training set.
6. Train a linear classifier on the generated training set to predict which generative factor has been fixed.
7. Z_{diff} score, also known as β-VAE metric, is the accuracy of the classifier.

In a perfectly disentangled representation, we would expect a zero in the dimension of the training input associated with the fixed generative factor, and the classifier would learn to map the zero-value index to the factor index.

4.2. Z_{min} Variance Score

To overcome certain weaknesses of Z_{diff} score, Kim and Mnih [42] introduced an unsupervised metric called Z-min Variance, also known as Factor-VAE metric. The intuition behind this metric is the same as the β-VAE metric with some improvements: a change in the way the latent representation is formed when a generative factor is fixed, in addition to the use of a specific type of classifier to predict which factor has been fixed. Calculating the Factor-VAE metric requires these steps:

1. Randomly choose a generative factor f_k.
2. Generate a batch of vectors, where the value of the selected factor f_k is held fixed in the batch while the other generative factors f_{k-1} are randomly selected. For a batch of L samples:

$$p_1 = (x_{1,1}, \ldots, x_{1,L})$$

3. Map each generated vector to latent code using the inference model:

$$q(z|x) \sim N(\mu(x), \sigma(x))$$

4. Normalize each variable within the latent representation using its empirical standard deviation calculated on the dataset. For a batch of L samples:

$$(z_1/s \dots z_L/s)$$

5. Calculate the empirical variance in each code of the normalized representations.

$$e = (Varz_1/s \dots Varz_L/s)$$

6. The factor index k and the latent variable index that has the lowest variance provide a training instance for the classifier. The factor index k and the index of the code dimension with the lowest variance give one training point for the classifier. These steps are repeated for each generative factor in order to create a substantial training set.
7. Train a majority vote classifier on the generated training set to predict which generative factor was fixed.
8. Z_{min} Variance score is equivalent to the classifier accuracy.

By normalizing the latent representations, the authors ensure that the argmin is insensitive to the rescaling of the representation in each latent variable. For a perfectly disentangled representation, one expects to have an empirical variance of zero in the dimension corresponding to the fixed factor.

4.3. Mutual Information Gap (MIG Score)

Chen et al. [41] introduced a new disentanglement metric based on mutual information theory. Mutual Information Gap (MIG) computes the mutual information (MI) between each generative factor x_i and latent code z_j. Higher mutual information denotes a deterministic relationship between z_j and x_j. The mutual information gap score can be estimated through the steps below:

1. Calculate the mutual information between each pair of latent variables and known generative factors.
2. Each generative factor may have high mutual information with several latent variables. Therefore, for every single factor, classify latent variables according to the amount of information they stored about this factor.
3. Calculate the difference between the top two values of mutual information for each generative factor.
4. Normalize this difference by dividing by the entropy of the corresponding generative factor.
5. The Mutual Information Gap (MIG) score is equivalent to the average of these differences [41]:

$$MIG(x,z) = \frac{1}{K} \sum_k \frac{1}{H(x_k)} \left(I(z_{j^{(k)}}; x_k) - \max_{j \neq j^{(k)}} I(z_j; x_k) \right) \tag{26}$$

where $j^{(k)} = \arg\max_j I(z_j, x_k)$, and K is the known generative factors.

4.4. Attribute Predictability Score (SAP)

In parallel to the MIG score, Kumar et al. [43] provided a metric of disentanglement also based on Mutual Information. Applying this metric requires these steps:

1. For each generative factor, compute the R2 score of linear regression (for continuous factors) or classification score (balanced accuracy for categorical factors) of predicting a j-th generative factor using only a i-th variable in the latent representation.
2. Compute the difference between the top two most-predictive latent codes.

3. The mean of those differences is the Attribute Predictability Score (SAP) [43].

$$SAP(x,z) = \frac{1}{K}\sum_{k}\left(I_{i_K,K} - \max_{j \neq i_k} I_{j,k}\right) \tag{27}$$

where $i_k = \arg\max_i I_{i,k}$, and K is the number of known generative factors.

5. Discussion

In this section, we discuss the methods and metrics presented in this review and highlight associated opportunities and open challenges.

5.1. Methods

The computational methods aim to use variational encoding along with different ELBO decompositions to learn the disentangled representation of the data. The shared point between each of these methods is either up-weighting the VAE objective, Equation (10), or adding some regularizers to the VAE objective that act to match the approximate posterior $q_\phi(z)$ to the factorized prior $p_\theta(z)$. Figure 3 illustrates this idea:

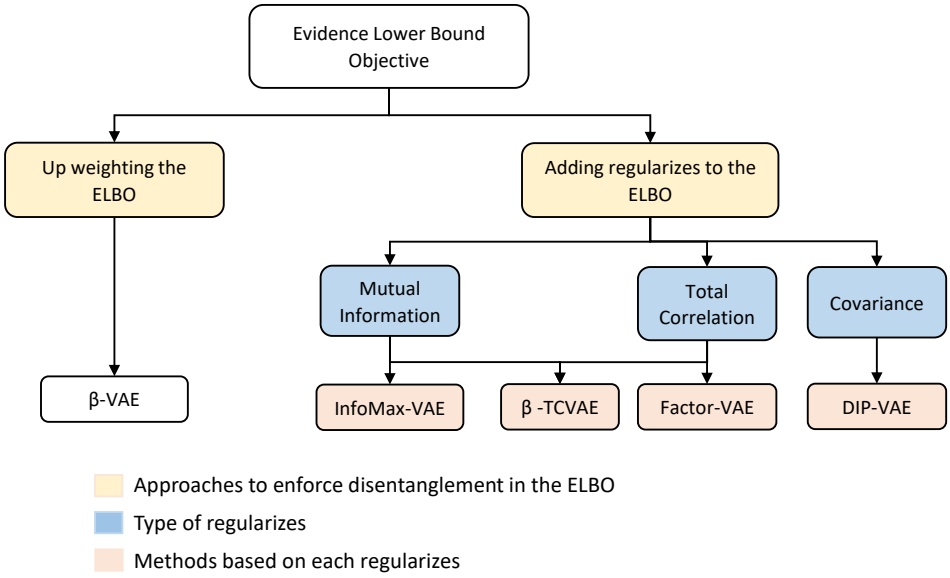

Figure 3. Schematic overview of different choices of augmenting the evidence lower bound of VAE. To improve disentanglement, most approaches focus on regularizing the original VAE objective by (i) up-weighting the ELBO with an adjustable hyperparameter β, resulting in a β-VAE approach. (ii) Adding different terms to the ELBO, such as mutual information, total correlation or covariance, resulting in InfoMax-VAE, β-TCVAE, Factor-VAE and DIP-VAE approaches, respectively.

By merely up-weighting the ELBO of VAE, the posterior $q_\phi(z|x)$ is pushed to correspond to the factorized prior $p_\theta(z)$, which results in a better disentanglement in comparison to the variational auto-encoder. β-VAE has shown acceptable performance. However, this penalization increases the tension between maximizing the data likelihood and disentanglement. As a result, there is a compromise between the accuracy of the reconstruction and the quality of disengagement within the latent representations learned. A higher value of β allows the achievement of better disentanglement but restricts latent channel information capacity and, therefore, a loss of information as it crosses this limited capacity latent z. The loss of high-frequency details about the data leads to poor reconstruction quality.

InfoMax-VAE outperforms β-VAE by constraining the latent representation so that the quantity of information kept about the observed data is maximized. In doing so, InfoMax-VAE is capable of obtaining high disentangling performance while maintaining a better reconstruction quality.

Factor-VAE ensures independence in the latent space by penalizing the total correlation term in the ELBO. A higher value of γ leads to a lower total correlation and, therefore, encourages independence in the code distribution. On the other hand, by not penalizing the mutual information, Factor-VAE keeps the information stored in z about x. Thus the model preserves high-frequency details about the data. By doing so, Factor-VAE improves upon β-VAE and has been reported to achieve a better balance between disentanglement and reconstruction quality. However, this model mostly remains difficult to train since it calls for an auxiliary discriminator and an internal optimization loop. On the other hand, the addition of a hyper-parameter during training may affect the model stability.

In β-TCVAE, the authors confirm the importance of total correlation for learning disentangled representation, and they propose an improvement in β-VAE and Factor-VAE. After breaking down the ELBO to a KL divergence term, mutual information and total correlation, they claim that adjusting β produces the best results. Thus they set $\alpha = \gamma = 1$, arriving at the same objective as Factor-VAE but with a simple way to estimate the total correlation without any additional hyper-parameter for more stable training. β-TCVAE has been capable of capturing independent factors in data distribution without having any degradation in the quality of reconstruction, thus surpassing β-VAE. However, we have the same disentanglement performance as Factor-VAE but in a simple manner to compute the total correlation.

DIP-VAE attempts to improve the performance of disentanglement by encouraging independence during inference. Having a disentangled prior that can be the basis for a generative disentangled model is the key idea behind DIP-VAE. DIP-VAE pushes the aggregated posterior $q_\phi(z)$ to correspond to a factorized prior $p(z)$ by matching the moments of the two distributions. By doing so, DIP-VAE is capable of learning disentangled representation without introducing any trade-off between disentangling latent variables and maximizing the data likelihood. As a result, DIP-VAE has a better reconstruction quality contrary to β-VAE and with similar performance to Factor-VAE and β-TCVAE with learning disentangled representation without introducing a compromise between disentangling latent variables and the plausibility of the observed data.

In Table 4, we summarize the different choices of the regularizers applied for each method.

Table 4. Summary of different ELBO decompositions alongside the regularization applied. For each method, the learning objective is given by $\mathcal{L}_{common}$ + *regularizer*.

Method	$\mathcal{L}_{common}$	Regularizers
VAE	$\mathcal{L}_{\text{VAE}}$	–
β-VAE	$\mathcal{L}_{\beta\text{-VAE}}$	–
InfoMax-VAE	$\mathcal{L}_{\beta\text{-VAE}}$	$\alpha I_{q_\phi}(x,z)$
Factor-VAE	$\mathcal{L}_{\text{VAE}}$	$\text{TC}\big(q_\phi(z)\big)$
β-TCVAE	$\mathcal{L}_{\beta\text{-VAE}}$	$\alpha I_{q_\phi}(x,z) + \text{TC}\big(q_\phi(z)\big)$
DIP-VAE	$\mathcal{L}_{\text{VAE}}$	$\lambda \left\| \text{cov}_{q_\phi(z)}[z] - \text{I} \right\|_F^2$

5.2. Metrics

It is currently unclear what exactly makes a disentangled metric better than another, but before analyzing metrics, we propose three criteria that constitute a disentangled representation, and we seek to analyze to what extent the metrics respect these criteria.

5.2.1. Properties of a Disentangled Representation

Modularity: changes in one factor have no impact on other factors [90]. The same analogy is in the representation space, and the factors are also independent. This property is also known as disentanglement in [87].

Compactness: the extent to which each generative factor is entered by one latent variable [88]. In other words, varying an underlying factor should have as small as possible effect on the latent space. Ideally, each generative factor is associated with only one latent code. In [87], the author refers to these criteria as completeness.

Informativeness: the amount of information shared between latent variables and generative factors [87]. In other words, the value of a given factor can be precisely determined from the code. This criterion is also known as explicitness in [90].

5.2.2. Comparison

Previous attempts to quantify disentangling have considered different aspects of modularity, compactness and informativeness criteria.

Higgins [20] argues that a good representation is one where the modularity of the latent representation holds. To make sure this property holds, he assumes that generative factors are independent and quantify the independence of the inferred latent variables using a simple classifier. However, this metric has several weaknesses. The Z_{diff} score is based on the classifier used to achieve the score. Nevertheless, the classifier could be sensitive to several hyper-parameters, such as the choice of the optimizer, the initialization of the weights and the epochs. Furthermore, there may be a perfectly disentangled representation where a generative factor corresponds to several latent variables instead of one variable. Thus Z_{diff} does not satisfy the compactness property. For example, if we fix this generative factor, the corresponding latent codes will have a variation of 0. Therefore, the classifier fails to distinguish between the latent codes and returns an accuracy of less than 1. Finally, the β-VAE metric does not require any assumptions about the factor-code relationship and therefore does not satisfy informativeness criteria.

Z_{min} Variance addresses several issues of the β-VAE metric. For instance, a majority vote classifier that predicts the fixed generative factor according to the variation of the latent variables actually allows having fewer additional hyper-parameters to optimize and, therefore, having a more reliable final score. On the other hand, and similar to the β-VAE metric, Z_{min} Variance depends on data with an independent factor. Furthermore, $Zmin$ variance satisfies the compactness property by fixing the number of subsets of data and generating a training set that covers all possible combinations of factor-latent codes while maintaining a fixed factor. However, it does not require any assumptions about the factor-code relations. Thus it does not fulfill the informativeness criteria.

Unlike the Z_{diff} or Z_{min} metrics, the SAP score does not require any additional classifier and, therefore, returns a more reliable final score. The SAP metric satisfies informativeness and compactness criteria by computing the score of predicting each generative factor using a single variable in the latent representation. However, it does not penalize the modularity between generative factors.

The main advantage of the Mutual Information Gap (MIG) score is that it does not require many additional hyper-parameters contrary to β-VAE and Factor-VAE. Moreover, it encourages the compactness of the representation by pushing only one latent variable to be informative about a factor. By definition, it computes the amount of information shared between latent variables and generative factors.

Table 5 summarizes the findings from our analysis.

Table 5. Summary of our findings from metric comparison. We found that most of the existing metrics can either satisfy modularity, compactness or informativeness criteria, except for the Mutual Information Gap score that captures all the three properties of a representation.

Metric	Satisfy Modularity	Satisfy Compactness	Satisfy Informativeness
Z_{diff}	Yes	No	No
Z_{min}	Yes	Yes	No
SAP	No	Yes	Yes
MIG	**Yes**	**Yes**	**Yes**

6. Conclusions and Future Directions

In this work, we conduct an extensive survey of disentangled representation learning through five state-of-the-art approaches focused on variational auto-encoders, alongside a detailed study on how to quantify disentanglement as well as conducting a comparison on the state-of-the-art of supervised disentanglement metrics.

We highlighted the underlying processes of disentangled representation learning methods driven by the development and deployment of computer vision algorithms using deep neural network approaches. In fact, each method considered herein is a variant of variational auto-encoder, and more precisely, they only vary the VAE objective, known as the Evidence Lower Bound Objective (ELBO). In particular, these methods either (i) up-weight the evidence lower bound, (ii) add a regularization to the ELBO that acts to match the approximated posterior to the factorized prior or (iii) combine a regularization and overweight the ELBO. By just up-weighting the evidence lower bound objective, one can observe a clear trade-off between disentanglement and reconstruction quality, which is the case for β-VAE. On the other hand, adding a regularizer to the evidence lower bound objective (Factor-VAE, InfoMax-VAE and DIP-VAE) or up-weighting ELBO while adding a regularizer (β-TCVAE) allows better disentanglement while preserving reconstruction quality.

Further, this study performs a comprehensive analysis and a fair comparison of the most well-known supervised disentanglement metrics. It considered three criteria that make one disentangled representation better than another (i) modularity, (ii) compactness and (iii) informativeness, and then compared metrics with respect to each criterion. We found that it is difficult to satisfy all three criteria at the same time. Most of the existing metrics meet one or two out of the three criteria. Only the mutual information gap score is robust to these criteria and able to give a general measure of the disentanglement quality.

However, some limitations remain unresolved and could be addressed in further work in order to improve the relevance of disentangled representation learning approaches in future work. (i) Despite the empirical success, existing disentangled representation learning approaches tend to ignore the latent variables and produce unrealistic, blurry samples with a significant reconstruction error when applied to complex datasets. There are several papers that discuss the issue of latent variable collapse [91–94], but more analysis on this issue is needed. (ii) Most of the approaches are based on VAE or GAN, more research on other potential models, e.g., diffusion model [95], would allow new ways for disentangled representation learning. (iii) Although disentangled representation learning has achieved several successes in generalization to unseen scenarios, state-of-the-art approaches for learning such representations have so far only been evaluated on small synthetic datasets. It will be interesting to explore the ability to generalize on complex real-world datasets. (iv) Finally, recent works discussed how to handle visual attacks and anomaly detection using dimensionality reduction, such as the SVD algorithm for neural networks and GAN [96–98]. Inspired by these works, it will be interesting to explore the application of disentangled representation learning while preventing visual attacks.

Moving forward, this survey is not only useful to provide insights for researchers that are currently working in the related area but can also be used as a basis for the implementation of new approaches. Our current goal is that we can build on these methods to develop an approach capable of identifying the underlying generative factors on more challenging datasets for a real-world application in an industrial environment.

Author Contributions: Conceptualization, I.E., C.-H.P., T.N. and M.E.-B.; methodology, I.E., C.-H.P, T.N. and M.E.-B.; writing—original draft preparation, I.E.; writing—review and editing, C.-H.P., T.N., M.E.-B. and J.-R.F.; supervision, M.E.-B. and I.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bengio, Y. Learning deep architectures for AI. *Found. Trends Mach. Learn.* **2009**, *2*, 1–127. [CrossRef]
2. Higgins, I.; Amos, D.; Pfau, D.; Racaniere, S.; Matthey, L.; Rezende, D.; Lerchner, A. Towards a definition of disentangled representations. *arXiv* **2018**, arXiv:1812.02230.
3. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798–1828. [CrossRef] [PubMed]
4. Szabó, A.; Hu, Q.; Portenier, T.; Zwicker, M.; Favaro, P. Challenges in disentangling independent factors of variation. *arXiv* **2017**, arXiv:1711.02245.
5. Suter, R.; Miladinovic, D.; Schölkopf, B.; Bauer, S. Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness. In Proceedings of the 36th International Conference on Machine Learning, Beach, CA, USA, 10–15 June 2019; Volume 97, pp. 6056–6065.
6. Le-Khac, P.; Healy, G.; Smeaton, A. Contrastive Representation Learning: A Framework and Review. *IEEE Access* **2020**, *8*, 193907–193934. [CrossRef]
7. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
8. Báscones, D.; González, C.; Mozos, D. Hyperspectral Image Compression Using Vector Quantization, PCA and JPEG2000. *Remote Sens.* **2018**, *10*, 907. [CrossRef]
9. Stone, J. Independent component analysis: An introduction. *Trends Cogn. Sci.* **2002**, *6*, 59–64. [CrossRef]
10. Naik, G.; Kumar, D. An overview of independent component analysis and its applications. *Informatica* **2011**, *35*, 63–82.
11. Henry, E.; Hofrichter, J. Singular value decomposition: Application to analysis of experimental data. *Methods Enzymol.* **1992**, *210*, 129–192.
12. Montero, M.; Ludwig, C.; Costa, R.; Malhotra, G.; Bowers, J. The Role of Disentanglement in Generalisation. Available online: https://openreview.net/forum?id=qbH974jKUVy (accessed on 15 February 2023).
13. Shen, Z.; Liu, J.; He, Y.; Zhang, X.; Xu, R.; Yu, H.; Cui, P. Towards out-of-distribution generalization: A survey. *arXiv* **2021**, arXiv:2108.13624.
14. Duan, S.; Matthey, L.; Saraiva, A.; Watters, N.; Burgess, C.; Lerchner, A.; Higgins, I. Unsupervised model selection for variational disentangled representation learning. *arXiv* **2019**, arXiv:1905.12614.
15. Zheng, H.; Lapata, M. Real-World Compositional Generalization with Disentangled Sequence-to-Sequence Learning. *arXiv* **2022**, arXiv:2212.05982.
16. Dittadi, A.; Träuble, F.; Locatello, F.; Wüthrich, M.; Agrawal, V.; Winther, O.; Bauer, S.; Schölkopf, B. On the transfer of disentangled representations in realistic settings. *arXiv* **2020**, arXiv:2010.14407.
17. Montero, M.; Bowers, J.; Costa, R.; Ludwig, C.; Malhotra, G. Lost in Latent Space: Disentangled Models and the Challenge of Combinatorial Generalisation. *arXiv* **2022**, arXiv:2204.02283.
18. Locatello, F.; Tschannen, M.; Bauer, S.; Rätsch, G.; Schölkopf, B.; Bachem, O. Disentangling factors of variation using few labels. *arXiv* **2019**, arXiv:1905.01258.
19. Schölkopf, B.; Janzing, D.; Peters, J.; Sgouritsa, E.; Zhang, K.; Mooij, J. On causal and anticausal learning. *arXiv* **2012**, arXiv:1206.6471.
20. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In Proceedings of the International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016.
21. Ridgeway, K. A survey of inductive biases for factorial representation-learning. *arXiv* **2016**, arXiv:1612.05299.
22. Wang, Q.; Zhou, H.; Li, G.; Guo, J. Single Image Super-Resolution Method Based on an Improved Adversarial Generation. *Appl. Sci.* **2022**, *12*, 6067. [CrossRef]
23. Revell, G. Madeleine: Poetry and Art of an Artificial Intelligence. *Arts* **2022**, *11*, 83. [CrossRef]
24. Tsai, Y.; Liang, P.; Zadeh, A.; Morency, L.; Salakhutdinov, R. Learning factorized multimodal representations. *arXiv* **2018**, arXiv:1806.06176.
25. Hsu, W.; Glass, J. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv* **2018**, arXiv:1805.11264.
26. Xu, Z.; Lin, T.; Tang, H.; Li, F.; He, D.; Sebe, N.; Timofte, R.; Van Gool, L.; Ding, E. Predict, prevent, and evaluate: Disentangled text-driven image manipulation empowered by pre-trained vision-language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 18229–18238.
27. Zou, W.; Ding, J.; Wang, C. Utilizing BERT Intermediate Layers for Multimodal Sentiment Analysis. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 11–15 July 2022; pp. 1–6.
28. Liu, X.; Sanchez, P.; Thermos, S.; O'Neil, A.; Tsaftaris, S. Learning disentangled representations in the imaging domain. *Med. Image Anal.* **2022**, *80*, 102516. [CrossRef] [PubMed]
29. Chartsias, A.; Joyce, T.; Papanastasiou, G.; Semple, S.; Williams, M.; Newby, D.; Dharmakumar, R.; Tsaftaris, S. Disentangled representation learning in cardiac image analysis. *Med. Image Anal.* **2019**, *58*, 101535. [CrossRef]

30. Hsieh, J.; Liu, B.; Huang, D.; Fei-Fei, L.; Niebles, J. Learning to decompose and disentangle representations for video prediction. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 515–524.
31. Denton, E.L. Unsupervised learning of disentangled representations from video. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4417–4426.
32. Comas, A.; Zhang, C.; Feric, Z.; Camps, O.; Yu, R. Learning disentangled representations of videos with missing data. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 3625–3635.
33. Guen, V.; Thome, N. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11474–11484.
34. Fan, K.; Joung, C.; Baek, S. Sequence-to-Sequence Video Prediction by Learning Hierarchical Representations. *Appl. Sci.* **2020**, *10*, 8288. [CrossRef]
35. Zou, Y.; Liu, H.; Gui, T.; Wang, J.; Zhang, Q.; Tang, M.; Li, H.; Wang, D. Divide and Conquer: Text Semantic Matching with Disentangled Keywords and Intents. *arXiv* **2022**, arXiv:2203.02898.
36. Dougrez-Lewis, J.; Liakata, M.; Kochkina, E.; He, Y. Learning disentangled latent topics for twitter rumour veracity classification. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 3902–3908.
37. Zhu, Q.; Zhang, W.; Liu, T.; Wang, W. Neural stylistic response generation with disentangled latent variables. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Bangkok, Thailand, 1–6 August 2020; pp. 4391–4401.
38. Lake, B.; Ullman, T.; Tenenbaum, J.; Gershman, S. Building machines that learn and think like people. *Behav. Brain Sci.* **2017**, *40*, e253. [CrossRef]
39. Kingma, D.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
40. Burgess, C.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; Lerchner, A. Understanding disentangling in β-VAE. *arXiv* **2018**, arXiv:1804.03599.
41. Chen, R.; Li, X.; Grosse, R.; Duvenaud, D. Isolating Sources of Disentanglement in Variational Autoencoders. *arXiv* **2018**, arXiv:1802.04942.
42. Kim, H.; Mnih, A. Disentangling by factorising. *arXiv* **2018**, arXiv:1802.05983.
43. Kumar, A.; Sattigeri, P.; Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv* **2017**, arXiv:1711.00848.
44. Rezaabad, A.; Vishwanath, S. Learning representations by maximizing mutual information in variational autoencoders. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020; pp. 2729–2734.
45. Hejna, J.; Vangipuram, A.; Liu, K. Improving Latent Representations via Explicit Disentanglement. 2020. Available online: http://joeyhejna.com/files/disentanglement.pdf (accessed on 13 December 2022).
46. Locatello, F.; Bauer, S.; Lucic, M.; Rätsch, G.; Gelly, S.; Schölkopf, B.; Bachem, O. A sober look at the unsupervised learning of disentangled representations and their evaluation. *arXiv* **2020**, arXiv:2010.14766.
47. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
48. Cho, W.; Choi, Y. LMGAN: Linguistically Informed Semi-Supervised GAN with Multiple Generators. *Sensors* **2022**, *22*, 8761. [CrossRef]
49. Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; Abbeel, P. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 2172–2180.
50. Radford, A.; Metz, L.; Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv* **2015**, arXiv:1511.06434.
51. Lin, Z.; Thekumparampil, K.; Fanti, G.; Oh, S. Infogan-cr: Disentangling generative adversarial networks with contrastive regularizers. *arXiv* **2019**, arXiv:1906.06034.
52. Xiao, T.; Hong, J.; Ma, J. Dna-gan: Learning disentangled representations from multi-attribute images. *arXiv* **2017**, arXiv:1711.05415.
53. Jeon, I.; Lee, W.; Pyeon, M.; Kim, G. Ib-gan: Disentangled representation learning with information bottleneck generative adversarial networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 7926–7934.
54. Jing, L.; Tian, Y. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 4037–4058. [CrossRef] [PubMed]
55. Ericsson, L.; Gouk, H.; Loy, C.; Hospedales, T. Self-supervised representation learning: Introduction, advances, and challenges. *IEEE Signal Process. Mag.* **2022**, *39*, 42–62. [CrossRef]
56. Schiappa, M.; Rawat, Y.; Shah, M. Self-supervised learning for videos: A survey. *ACM Comput. Surv.* **2022**. [CrossRef]
57. Xie, Y.; Arildsen, T.; Tan, Z. Disentangled Speech Representation Learning Based on Factorized Hierarchical Variational Autoencoder with Self-Supervised Objective. In Proceedings of the 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), Gold Coast, Australia, 25–28 October 2021; pp. 1–6.
58. Zhang, Z.; Zhang, L.; Zheng, X.; Tian, J.; Zhou, J. Self-supervised adversarial example detection by disentangled representation. *arXiv* **2021**, arXiv:2105.03689.

59. Kaya, B.; Timofte, R. Self-supervised 2D image to 3D shape translation with disentangled representations. In Proceedings of the 2020 International Conference on 3D Vision (3DV), Fukuoka, Japan, 25–28 November 2020; pp. 1039–1048.
60. Wang, T.; Yue, Z.; Huang, J.; Sun, Q.; Zhang, H. Self-supervised learning disentangled group representation as feature. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18225–18240.
61. Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *Int. Conf. Mach. Learn.* **2019**, *97*, 4114–4124.
62. Baldi, P. Autoencoders, unsupervised learning, and deep architectures. In Proceedings of the ICML Workshop on Unsupervised And Transfer Learning, Bellevue, DC, USA, 27 June 2012; pp. 37–49.
63. Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.; Bottou, L. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.
64. Pham, C.; Ladjal, S.; Newson, A. PCA-AE: Principal Component Analysis Autoencoder for Organising the Latent Space of Generative Networks. *J. Math. Imaging Vis.* **2022**, *64*, 569–585. [CrossRef]
65. Bank, D.; Koenigstein, N.; Giryes, R. Autoencoders. *arXiv* **2020**, arXiv:2003.05991.
66. Song, C.; Liu, F.; Huang, Y.; Wang, L.; Tan, T. Auto-encoder Based Data Clustering. In Proceedings of the CIARP, Havana, Cuba, 20–23 November 2013.
67. Gogoi, M.; Begum, S. Image classification using deep autoencoders. In Proceedings of the 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Tamil Nadu, India, 14–16 December 2017; pp. 1–5.
68. Zhang, Y.; Lee, K.; Lee, H. Augmenting Supervised Neural Networks with Unsupervised Objectives for Large-scale Image Classification. In Proceedings of the 33rd International Conference on Machine Learning, York City, NY, USA, 19–24 June 2016; Volume 48, pp. 612–621.
69. Hoffman, M.; Blei, D.; Wang, C.; Paisley, J. Stochastic variational inference. *J. Mach. Learn. Res.* **2013**, *14*, 1303–1347.
70. Jha, A.; Anand, S.; Singh, M.; Veeravasarapu, V. Disentangling factors of variation with cycle-consistent variational auto-encoders. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 805–820.
71. Doersch, C. Tutorial on variational autoencoders. *arXiv* **2016**, arXiv:1606.05908.
72. Kingma, D.; Welling, M. An introduction to variational autoencoders. *arXiv* **2019**, arXiv:1906.02691.
73. Rezende, D.; Mohamed, S.; Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. *Int. Conf. Mach. Learn.* **2017**, *32*, 1278–1286.
74. Asperti, A.; Trentin, M. Balancing reconstruction error and kullback-leibler divergence in variational autoencoders. *IEEE Access* **2020**, *8*, 199440–199448. [CrossRef]
75. Hu, M.; Liu, Z.; Liu, J. Learning Unsupervised Disentangled Capsule via Mutual Information. In Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 18–23 July 2022; pp. 1–8.
76. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015; pp. 3730–3738.
77. Aubry, M.; Maturana, D.; Efros, A.; Russell, B.; Sivic, J. Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of cad models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3762–3769.
78. Paysan, P.; Knothe, R.; Amberg, B.; Romdhani, S.; Vetter, T. A 3D face model for pose and illumination invariant face recognition. In Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, Genova, Italy, 2–4 September 2009; pp. 296–301.
79. Matthey, L.; Higgins, I.; Hassabis, D.; Lerchner, A. dSprites: Disentanglement Testing Sprites Dataset. 2017. Available online: https://github.com/deepmind/dsprites-dataset/ (accessed on 13 December 2022).
80. Kullback, S.*Information Theory and Statistics*; (Courier Corporation) Dover Publications: New York, NY, USA, 1997.
81. Hoffman, M.; Johnson, M. Elbo surgery: Yet another way to carve up the variational evidence lower bound. In Proceedings of the Workshop in Advances in Approximate Bayesian Inference, NIPS, Barcelona, Spain, 9 December 2016; Volume 1.
82. Makhzani, A.; Frey, B. Pixelgan autoencoders. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1972–1982.
83. Watanabe, S. Information Theoretical Analysis of Multivariate Correlation. *IBM J. Res. Dev.* **1960**, *4*, 66–82. [CrossRef]
84. Nguyen, X.; Wainwright, M.; Jordan, M. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Trans. Inf. Theory* **2010**, *56*, 5847–5861. [CrossRef]
85. Sugiyama, M.; Suzuki, T.; Kanamori, T. Density-ratio matching under the Bregman divergence: A unified framework of density-ratio estimation. *Ann. Inst. Stat. Math.* **2011**, *64*, 1009–1044. [CrossRef]
86. Harrison, R. Introduction to monte carlo simulation. *AIP Conf. Proc.* **2010**, *1204*, 17–21. [PubMed]
87. Eastwood, C.; Williams, C. A framework for the quantitative evaluation of disentangled representations. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
88. Zaidi, J.; Boilard, J.; Gagnon, G.; Carbonneau, M. Measuring disentanglement: A review of metrics. *arXiv* **2020**, arXiv:2012.09276.
89. Sepliarskaia, A.; Kiseleva, J.; Rijke, M. Evaluating disentangled representations. *arXiv* **2019**, arXiv:1910.05587.
90. Ridgeway, K.; Mozer, M. Learning deep disentangled embeddings with the f-statistic loss. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 185–194.
91. Chen, X.; Kingma, D.; Salimans, T.; Duan, Y.; Dhariwal, P.; Schulman, J.; Sutskever, I.; Abbeel, P. Variational lossy autoencoder. *arXiv* **2016**, arXiv:1611.02731.

92. Zhao, S.; Song, J.; Ermon, S. Towards deeper understanding of variational autoencoding models. *arXiv* **2017**, arXiv:1702.08658.
93. Zhang, K. On mode collapse in generative adversarial networks. In Proceedings of the Artificial Neural Networks and Machine Learning—ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, 14–17 September 2021; pp. 563–574.
94. Alemi, A.; Poole, B.; Fischer, I.; Dillon, J.; Saurous, R.; Murphy, K. Fixing a broken ELBO. *Int. Conf. Mach. Learn.* **2018**, *80*, 159–168.
95. Liu, J.; Yuan, Z.; Pan, Z.; Fu, Y.; Liu, L.; Lu, B. Diffusion Model with Detail Complement for Super-Resolution of Remote Sensing. *Remote Sens.* **2022**, *14*, 4834. [CrossRef]
96. Benrhouma, O.; Alkhodre, A.; AlZahrani, A.; Namoun, A.; Bhat, W. Using Singular Value Decomposition and Chaotic Maps for Selective Encryption of Video Feeds in Smart Traffic Management. *Appl. Sci.* **2022**, *12*, 3917. [CrossRef]
97. Andriyanov, N. Methods for preventing visual attacks in convolutional neural networks based on data discard and dimensionality reduction. *Appl. Sci.* **2021**, *11*, 5235. [CrossRef]
98. Samuel, D.; Cuzzolin, F. Svd-gan for real-time unsupervised video anomaly detection. In Proceedings of the British Machine Vision Conference (BMVC), Virtual, 22–25 November 2021.