# Intrinsic disentanglement: an invariance view for deep generative models

**Michel Besserve** [1 2] **Rémy Sun** [1 3] **Bernhard Schölkopf** [1]

## Abstract

Deep generative models such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) are important tools to capture and investigate the properties of complex empirical data. However, the complexity of their inner elements makes their functioning challenging to interpret and modify. In this respect, these architectures behave as black box models. In order to better understand the function of such network, we analyze the modularity of these system by quantifying the disentanglement of their intrinsic parameters. This concept relates to a notion of invariance to transformations of internal variables of the generative model, recently introduced in the field of causality. Our experiments on generation of human faces with VAEs supports that modularity between weights distributed over layers of generator architecture is achieved to some degree, and can be used to understand better the functioning of these architectures. Finally, we show that modularity can be enhanced during optimization.

## 1. Introduction

Deep generative models have proven powerful in learning to design realistic images in a variety of complex domains (handwritten digits, human faces, interior scenes). In particular, two approaches have recently emerged: Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), which train an image *generator* by having it fool a *discriminator* that should tell apart real from artificially generated images; and Variational Autoencoders (VAEs) (Kingma and Welling, 2013; Rezende et al., 2014) that learn both a mapping from latent variables to the data, the *decoder*, and the converse mapping from the data to the latent variables, the *encoder*, such that correspondences between latent variables and data features can be easily investigated. Although these architectures have been lately the subject of extensive investigations, understanding why and how they work, and how they can be improved, remains elusive. One major difficulty is the complexity of the function class entailed by their non-linearities and high dimensional parameter space. In order to improve this understanding, uncovering a modular structure in those architectures, such that each part of a network can be assigned a specific function, would be a major step forward.

In this paper, we propose that modularity can be quantified and exploited in a causal framework to infer whether modules within the architecture can be further disentangled. This hypothesis relies on recent work exploiting the postulate of Independence of Cause and Mechanism stating that Nature chooses independently the properties of a cause and those of the mechanism that generate effects from the cause (Janzing and Schölkopf, 2010; Lemeire and Janzing, 2012). It has been recently demonstrated that many approaches pertaining to this framework rely on a principle of invariance of the output of a mechanism with respect to transformations that are applied to its input (Besserve et al., 2018). We then show this invariance is a natural property to enforce in generative models, and that it extends the classical notion of disentangled representation investigated in the literature. Moreover, we propose that the Spectral Independence Criterion (SIC) (Shajarisales et al., 2015) can be used to quantify such invariance. We show empirically how VAEs trained on the CelebA face dataset express a form of invariance to perturbation of the intermediate activation maps. Finally we show how optimizing the SIC can improve their desirable invariance properties.

## 2. Modularity as intrinsic disentanglement

### 2.1. Forward-inverse optics in vision

We first introduce our framework in the context of a seminal example: work in neuroscience and computer vision has since long tried to address how a scene can be related to a high level internal representation. Such question can be framed using two objects: (1) the mapping of a 3D scene to its perceived (2D) image, called *forward optics*, (2) the converse mapping, called *inverse optics*, (see e.g. (Kawato et al., 1993)).

[1] MPI for Intelligent Systems, Tübingen, Germany [2] MPI for Biological Cybernetics, Tübingen, Germany [3] ENS Rennes, France. Correspondence to: Michel Besserve <michel.besserve@tuebingen.mpg.de>.

*Presented at the ICML 2018 workshop on Theoretical Foundations and Applications of Deep Generative Models*, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

A key difference between both maps is that forward optics can be concisely described with a restricted set of equations taking into account physical parameters of the scene, while inverse optics does not have an explicit form and relies heavily on prior assumptions to be solved numerically. This has led in particular to forward-inverse approaches to visual perception, relying on the assumption that forward optics can be implemented by feedback projections from higher to lower visual areas, while feedforward projections provide an initial rough estimate of the inverse optics (Kawato et al., 1993; Tajima and Watanabe, 2011). The forward optics is then further refined iteratively, following a predictive coding principle (Rao and Ballard, 1999).

Paralleling this framework, many computer vision algorithms have relied on *inverse graphics* approaches that model both forward and inverse optics simultaneously (Kulkarni et al., 2015). In recent years, emphasis has been put on producing compact latent description of the scene in terms of high level features, sometimes coined a graphics code, reflecting a *disentangled* latent representation. However, the resulting architectures do not reflect the fundamental asymmetry in complexity between the original forward and inverse optics maps. Taking as reference VAE and GAN architectures, their forward and inverse maps are indeed typically implemented as mirror convolutional networks, resulting in two densely connected multilayered architectures whose weights are mostly not interpretable. As simplicity of the forward optics can allow an agent to efficiently manipulate and update internal representations, e.g. in order to plan interactions with the outside world, we argue that modularity of the forward models implemented by generators should be enforced in order to be understood and manipulated easily. This is the objective of the framework presented in this paper.

### 2.2. Causal generative models

In this paper, we will rely on the notion of causal generative models to represent any latent variable model used to fit observational data. Causality entails the idea that discovered relationships between variables have some degree of robustness to perturbations of the system under consideration. As a consequence a causal model allows predicting interventions and counterfactuals, and may thus generalize better. Causal models can be described based on Structural Equations (SEs) of the form

$$Y := f(X_1, X_2, \cdots, X_N, \epsilon)\,,$$

expressing the assignment of a value to variable $Y$ based on values of other variables $X_k$, with possibly additional exogenous effects accounted for through the random variable $\epsilon$. This expression thus stays valid if something selectively changes on the right hand side variables, and accounts for the robustness or invariance to interventions and counterfactuals expected from causal models as opposed to purely probabilistic ones (see for example (Peters et al., 2017; Pearl, 2000)). Such SEs can be combined to build a Structural Causal Model made of interdependent modules to represent a more complex system, for which dependencies between variables can be represented by a directed acyclic graph $\mathcal{G}$. Let us use such structural model to represent our generator:

**Definition 1** (Causal Generative Model (CGM))**.** *A causal generative model $\mathbb{G}(P_{\mathbf{Z}}, \boldsymbol{S}, G)$ consists in a distribution $P_{\mathbf{Z}}$ over latent variables $\mathbf{Z} = (Z_k)$, a collection $\boldsymbol{S}$ of structural equations assigning endogenous random variables $\mathbf{V} = (V_k)$ and output $I$ based on values of their endogenous or latent parents $\boldsymbol{Pa}_k$ in the directed acyclic graph $G$. We assume $I$ has no latent parent, such that it is assigned by two deterministic mappings using either latent or endogenous variables*

$$I = g(\mathbf{Z}) = \tilde{g}(\mathbf{V})\,.$$

The graphical representation of a CGM is exemplified on Fig. 1a.

### 2.3. Two forms of disentanglement

We introduce here a formal definition of the above concept of disentangled representation. In order to relate the definition to the concrete examples that will follow, we consider without loss of generality that the generated variable $I$ is meant to be an image.

**Definition 2** (Extrinsic disentanglement)**.** *A CGM $\mathbb{G}$ is extrinsically disentangled with respect to endomorphism $T$ and subset of latent variables $\mathcal{L}$, if there exists an endomorphism $T'$ of the latent variables such that for any image generated by a realization $\mathbf{z} = \mathbf{Z}(\omega)$ of the latent variables $(I = g(\mathbf{z}))$*

$$T(I) = g(T'(\mathbf{z}))\,, \tag{1}$$

*where $T'(\mathbf{z})$ only affects values of components of $\mathbf{z}$ in $\mathcal{L}$.*

The sparsity of the disentanglement is then reflected by the minimal size of the subset $\mathcal{L}$. Extrinsic disentanglement can be seen as a form of intervention on the CGM as illustrated in Fig. 1b. In this figure, we represent the effect of applying a transformation that affects only $Z_1$ (we thus abusively write $T'(Z_1)$), thus modifying descendant nodes, leading to a modified output $I' = T(I)$. We can easily see that this definition is compatible with the intuitive concept of disentangled representation as used for example in Kulkarni et al. (2015) in the context of inverse graphics, where $T$ would correspond to a change in e.g. illumination of the scene, while $T'$ would simply shift the values of the sparse set of latent variables controlling it.[1] In order for the definition

[1] In the context of inverse graphics, it could be argued that transformation $T$ should apply to the unobserved 3D scene (for example a 3D rotation of a face). This is compatible with our

to be useful in a quantitative setting, we would require an approximate version of equation (2), however we mostly use this definition for the purpose of contrasting it with the following.

**Definition 3** (Intrinsic disentanglement)**.** *A CGM $\mathbb{G}$ is intrinsically disentangled with respect to endomorphism $T$ and subset of endogenous variables $\mathcal{E}$ if there exists an endomorphism $T'$ such that for any image generated by a realization $\mathbf{z} = \mathbf{Z}(\omega)$ of the latent variables ($I = g(\mathbf{z}) = \tilde{g}(\mathbf{v})$),*

$$T(I) = \tilde{g}(T'(\mathbf{v})) \qquad (2)$$

*where $T'(\mathbf{v})$ only affects values of endogenous variables in $\mathcal{E}$.*

An illustration of this second notion of disentanglement is provided on Fig. 1c, where the split node indicates that the value of $V_3$ is computed as in the original CGM (Fig. 1a) before applying transformation $T'$ to the outcome. Intrinsic disentanglement directly relates to a causal interpretation of the generative model and its robustness to perturbation of its subsystems. To justify it, consider the case of Fig. 1d, where the GCM has an unaccounted latent variable $Z_3$. This may be due to the absence of significant variations of $Z_3$ in the training set, or simply bad performance of the estimation algorithm. If the remaining causal structure has been estimated in a satisfactory way, and the full structure is simple enough, a change in this missing variable can be ascribed to a change in only a small subset of the endogenous nodes. Then the transformation $T'$ from the definition can be seen as a proxy for the change in the structural equations induced by a change in $Z_3$. Broadly construed, appropriate transformations pairs $(T, T')$ emulate changes of unaccounted latent variables, allowing to check whether the fitted causal structure is likely to be robust to plausible changes in the dataset.

### 2.4. Related work

The issue of interpretability in convolutional neural networks has already been the topic of much research. Most of that research however has focused on discriminative neural networks, not generative ones. In the discriminative case, efforts have been made to find optimal activation patterns for filters ((Zeiler and Fergus, 2014),(Dosovitskiy and Brox, 2016)), to find correlation between intermediate feature space and data features ((Fong and Vedaldi, 2017),(Zhang et al., 2017b)) or to disentangle patterns detected by various filters to compute an explanatory graph (Zhang et al., 2017a). Furthermore, explicitly enforcing modularity in networks has been tried recently with Capsule networks architectures ((Sabour et al., 2017)), although Capsule network explicitly separate the architecture in different modules before training. A more detailed overview can found in review (Zhang and Zhu, 2018). It is important to emphasize discriminative and generative processes differ significantly, and working on generative processes allows to directly observe the effect of changes in intermediate representations on the generated picture rather than having to correlate it back input images.

The recent InfoGAN network ((Chen et al., 2016)) and other works ((Mathieu et al., 2016; Kulkarni et al., 2015; Higgins et al., 2017)) in disentanglement of latent variables in generative models can be seen as what we define as extrinsic disentanglement. As such, we believe our intrinsic disentanglement perspective should be complementary with such approaches and are not in direct competition.

Finally our approach relates to modularity and invariance principles formulated in the field of causality, in particular to (Besserve et al., 2018).

## 3. Quantifying modularity

Quantifying modularity of a given GCM presents several challenges. State of the art deep generative networks are made of densely connected layers, such that modules cannot be identified easily beyond the trivial distinction between successive layers. In addition, analysis of statistical dependencies between successive nodes in the graph is not likely to help, as the entailed relations are purely deterministic. In addition, the notion of intrinsic disentanglement is specific to transformations $T$ and $T'$, and the relationship between these functions may be very complex. However, the following simplified framework provides a practical solution.

### 3.1. Independence of cause and mechanism (ICM)

Elaborating on the missing variable example of Fig. 1d, a change in $Z_3$ induces $V_3$ to change to $T'(V3)$, which induces a change in $V1$. Given other variables are kept fixed, the structural equation

$$V_1 := m(V_3)\,,$$

is turned into a perturbed version

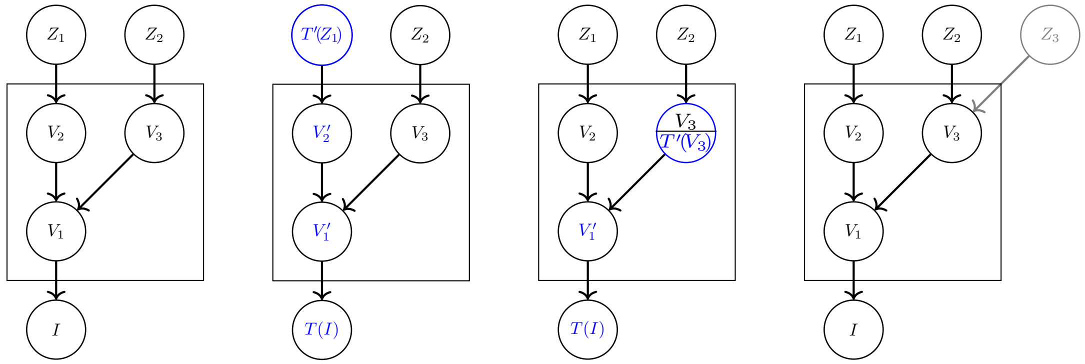
$$V_1^{pert} := m(T'(V_3))\,,$$

Assuming the GCM is correct, this perturbation should not affect the offspring vertices in a way compatible with generic properties of $I$. As a consequence, some property of $V_1$, that we call $C(V_1)$, specific to the modeled data will remain approximately invariant to any such transformation $T'$ such that

$$C(m(V_3)) \approx C(m(T'(V_3)))\,. \qquad (3)$$

This formula can be interpreted as the resulting output transformation $T$ not affecting the summary statistics computed

definition as long as it exists a (non-necessarily unique) solution of the inverse graphics that can then be composed with $T$.



(a) (b) (c) (d)

Figure 1: Graphical representation of CGMs. (a) Example CGM with 2 latent variables. (b) Illustration of extrinsic disentanglement with $\mathcal{L} = \{1\}$. (c) Illustration of intrinsic disentanglement with $\mathcal{E} = \{3\}$. (d) Illustration of unaccounted latent variable emulated by (c). Nodes modified by intervention on the graph are indicated in blue.

by $C(V_1)$. If in addition, mimicking the effect of intrinsic disentanglement on output $I$, we assume a transformation $T''$ such that $T'' = m \circ T'$, then the above formula relates directly to invariance of the function $C$ itself as equation (3) implies

$$C(m(V_3)) \approx C(T''(m(V_3)))\,.$$

This approach relates to the concept of ICM, in the sense that the family of transformations $T'$ decouples the cause $V_3$ from the mechanism $m$ (Besserve et al., 2018). Several causal inference methods rely on the postulate that properties of cause and mechanism are “independent” in a sense that can be formalized in various ways (Peters et al., 2017). Broadly construed, it reflects the idea that the mechanism does not adapt its properties to the specific input it receives, such that the global properties of the effect are generic (i.e. they are similar to what would have been produced with another cause of the same kind). In the present context it reflects the fact that $V_3$ is modulated by a latent factor that have nothing to do with the sub mechanism that computes $V_1$ from $V_3$.

Now that we have shown ICM is a natural criterion to evaluate intrinsic disentanglement, a quantifiable measure of how well that criterion is respected has to be chosen.

### 3.2. Spectral Independence

Shajarisales et al. (2015) introduce a specific formalization of ICM in the context of time series that will be suited to the study of convolutional layers in deep neural network. This relies on analyzing signals or images in the Fourier domain (see supplemental information for a background on Fourier analysis).

Assume now that our cause-effect pair $(X,\,Y)$ is a weakly stationary time series. This implies that the power of these signals can be decomposed in the frequency domain using their Power Spectral Densities (PSD) $S_x(\nu)$ and $S_y(\nu)$. If $Y$ results from the filtering of $X$ with convolution kernel $h$

$$Y := \left\{\textstyle\sum_{\tau\in\mathbb{Z}} h_\tau X_{t-\tau}\right\} = h * X\,. \tag{4}$$

then PSDs are related by the formula $S_y(\nu) = |\widehat{\mathrm{h}}(\nu)|^2 S_x(\nu)$ for all frequencies $\nu$. The Spectral Independence Postulate consists then in assuming that the power amplification of the filter at each frequency $|\widehat{\mathrm{h}}(\nu)|^2$ does not adapt to the input power spectrum $S_x(\nu)$, i.e. the filter will not tend to selectively amplify or attenuate the frequencies with particularly large or low power. This can be formalized by stating that the total output power (integral of the PSD) factorized into the product of input power and the energy of the filter, leading to the criterion (Shajarisales et al., 2015):

**Postulate 1** (Spectral Independence Criterion (SIC))**.** *Let $S_x$ be the Power Spectral Density (PSD) of a cause $X$ and $h$ the impulse response of the causal system of (4), then*

$$\int_{-1/2}^{1/2} S_x(\nu)|\widehat{\mathrm{h}}(\nu)|^2 d\nu = \int_{-1/2}^{1/2} S_x(\nu)d\nu \cdot \int_{-1/2}^{1/2} |\widehat{\mathrm{h}}(\nu)|^2 d\nu\,, \tag{5}$$

*holds approximately.*

This equation relates to an invariance to particular transformation as defined above. Indeed, if we consider the family of translations $(\tau_\nu)_{\nu\in[0,]}$ modulo 1 of the frequency axis, then equation (5) amounts to invariance of the output signal power $\mathcal{P}$ in the sense that

$$\mathcal{P}(h * X) = \mathbb{E}_\nu \mathcal{P}(h * \tau_\nu X)$$

where the frequency translation parameter $\nu$ is drawn from a uniform distribution of the unit interval. We can define a scale invariant quantity $\rho_{X\to Y}$ measuring the departure from this SIC assumption, i.e. the dependence between input power spectrum and frequency response of the filter: the

Spectral Dependency Ratio (SDR) from $X$ to $Y$ is defined as

$$\rho_{X\to Y} := \frac{\langle S_x \cdot |\widehat{\mathrm{h}}|^2\rangle}{\langle S_x\rangle\langle|\widehat{\mathrm{h}}|^2\rangle}\,, \tag{6}$$

where $\langle . \rangle$ denotes the integral (and also the average) over the unit frequency interval. As a consequence, $\rho_{X\to Y}\approx 1$ reflects spectral independence.

## 4. Independence of mechanisms in deep networks

We now introduce our causal reasoning in the context of deep convolutional networks, where the output of successive layers are often interpreted as different levels of representation of an image, from detailed low level features to abstract concepts. We thus investigate whether a form of modularity between successive layers can be identified using the above framework.

### 4.1. SIC in convolutional layers

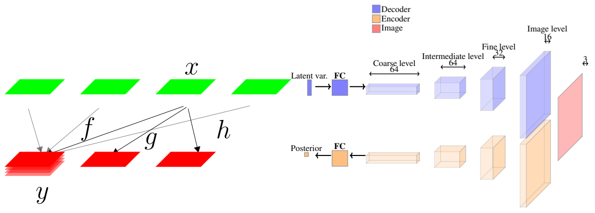
The SIC framework is well suited to the analysis of convolutional layers, since it assumes deterministic convolution mechanisms. Indeed, as can be seen looking at the leftmost activation map of the lower layer in Fig.2a, an activation map $y$ (corresponding to one channel in the considered layer) is generated from the $n$ channels' activation maps $x = (x^1, \dots, x^n)$ in the previous layer through the filter kernel $f = (f_1, \dots, f_n)$ according to

$$y = \sum_{i=1}^{n} f_i * x^i + b\,. \tag{7}$$

By looking only at 'partial' activation maps $y_k^i = f_i * x^i$, it is possible to view the relationship between a given (partial) activation map and the downstream convolutional layer as an actual convolution (with some additive constant bias). Therefore, unless specified otherwise, the term filters will refer specifically to partial filters $(f_i)$ in the rest of this paper.

One difference appears with respect to the original SIC assumptions: the striding possibly adds spacing between input pixels in order to progressively increase the dimension and resolution of the image from one layer to the next. Striding can be easily modeled, as it amounts to upsampling the input image before convolution. We denote $.^{\uparrow s}$ the upsampling operation with integer factor[2] $s$ that turns the 2D activation map $x$ into

$$x^{\uparrow s}[k,l] = \begin{cases} x[k/s, l/s], & k \text{ and } l \text{ multiple of s,}\\ 0, & \text{otherwise.}\end{cases}$$



(a) (b)

Figure 2: (a) Convolution pathways between layers. (b) Architecture of the pretrained VAE generator used in our experiments.

leading to a compression of the normalized frequency axis in the Fourier domain such that $\widehat{\mathrm{x}^{\uparrow \mathrm{s}}}(u,v) = \widehat{\mathrm{x}}(su, sv)$. The convolution relation in Fourier domain thus translates to $\widehat{\mathrm{y}}(u,v) = \widehat{\mathrm{h}}(u,v)\widehat{\mathrm{x}}(su,sv)$. As a consequence, the SDR measure needs to be adapted to upsampling by contracting rescaling the frequency axis of the activation map with respect to the one of the filter. Using power spectral density estimates based on Bartlett's method, we use a batch of input images of size $B$ leading to $B$ values of activation map $x$, $x_0, \dots, x_{B-1}$, to obtain the following SDR estimate:

$$\rho_{\{x_i\}\to f} = \frac{\left\langle \frac{1}{B}\sum_{i=0}^{(B-1)} |\widehat{\mathrm{f}}(u,v)\widehat{\mathrm{x}_\mathrm{i}}(su,sv)|^2 \right\rangle}{\left\langle |\widehat{\mathrm{f}}(u,v)|^2 \right\rangle \left\langle \frac{1}{B}\sum_{i=0}^{(B-1)} |\widehat{\mathrm{x}}(u,v)|^2 \right\rangle}\,. \tag{8}$$

As per our intrinsic disentanglement approach, modularity with respect to activation maps is the specific point we would like to attach most importance to. However, as can be seen on Fig.2a, a single activation map will be transformed in to multiple activation maps through convolution with different filters $f$,$g$ and $h$ leading to multiple SDR statistic. Therefore, we consider the average of all those SDR statistics when we refer to an activation map's SDR statistic. Now that quantitative evaluation of SIC can be computed for activation maps, we can try to further enforce spectral independence.

### 4.2. Optimizing filters to be more independent from activation maps

Direct optimization euclidian distance to 1 of the SDR statistic is challenging due to the normalization term in equation (6). To avoid this, for a fixed activation map, we therefore simply minimize the square difference between the SDR and its ideal value of 1, but multiplied by the normalization term $\left\langle |\widehat{f}|^2 \right\rangle$. For a single (filter,map) pair, this leads to the minimization of the objective

$$\left\langle |\widehat{\mathrm{f}}(u,v)|^2 \cdot \left( \frac{|\frac{1}{B}\sum_{i=0}^{(B-1)} \widehat{\mathrm{x}_\mathrm{i}}(su,sv)|^2}{\langle |\frac{1}{B}\sum_{i=0}^{(B-1)} \widehat{\mathrm{x}_\mathrm{i}}(su,sv)|^2 \rangle} - 1 \right) \right\rangle^2 . \tag{9}$$

[2] $s$ is the inverse of the stride parameter; the latter is fractional in that case

When multiple pairs are considered we simply minimize the sum of all corresponding objectives.

## 5. Experiments

We investigate in this section the above notions on real data in the form of the CelebFaces Attributes Dataset (CelebA)[3]. We used a plain VAE with least square reconstruction loss .[4] The general structure of the VAE is summarized in Fig. 2b. We distinguish the different layers with 4 levels indicated in Fig.2b: coarse (closest to latent variables), intermediate, fine and image level (closest to the image). Complete architecture details are provided in the supplemental material.

### 5.1. Evaluating invariance with a simple transformation

We apply a 1.5 fold horizontal stretching transformation to all maps of a single convolutional layer (coarse, intermediate, fine or image level) and compare the resulting distorted image to the result of applying such a stretch to the normally generated output image. The edges of the stretched maps and images are cropped symmetrically to keep to the right dimensions.

We suggest applying such transformation to intermediate layers and observing how it affects the output as a way to get some insights in the internal organization of such networks. This approach is in addition justified on a theoretical ground in section 2.

#### 5.1.1. Scale of convolutional layers

The images obtained by distorting various convolutional layers' activation maps are presented in Fig. 3a for the VAE trained with 10000 iterations. We can observe how the distortion affects differentially successive scales of the picture: A concrete example can be observed directly by considering the eyes generated by stretching the intermediate level activation maps (thrid row of Fig. 3a). Although the location of each eye changes according to the stretching, the eyes themselves seem to mostly keep their original dimensions. This supports the assumption that successive convolutional layers build upon each other to encode different scales of features. Interestingly, Fig. 3b replicating the result but after 40000 additional training iterations, show perturbed image of poorer quality, suggesting a loss in modularity when the number of iterations grows. In particular, grid like periodic interference patterns appearing at the fine and image levels are stronger, and may correspond to imperfect alignment of the ouput of convolution kernels encoding neighboring region of the image. Such artifact could possibly be tempered by a different implementation of upsampling (e.g. using nearest neighbor interpolation instead of fractional striding).

We next sought to quantify these initial observations using a discrete Haar wavelet transform of the images. For a given image, by zeroing wavelet coefficients at all but one scale and applying the inverse wavelet transform, we generated five component images containing features at successive spatial scales, from coarse to fine. For each generated original image and each scale, we then compute the difference between the component of the stretched original image and the component image obtained through distortion of a layer. Resulting examples are plotted on Fig. 3c for 10000 training iterations. The patterns of deformations for each level of distorted activation map is in accordance with the previous qualitative observations. In particular, we notice the perturbation localized at the level of eyes, mouth and nose for the intermediate level (Fig. 3c, second row), reflecting that the dimensions of these patterns are not fully rescaled, although their position is. We then computed the mean squared error resulting from the above differences over all pixels of 64 images of a batch. The resulting histograms for each perturbed layer on Fig. 3d shows that the mismatch is more concentrated on the finer scales, corresponding to scales encoded below the distorted layer, as the kernels encoding this finer scale are not affected by the stretch. For comparison, we show the same analysis for the case of 50000 training iteration of the VAE on Fig. 3d. It confirms the above qualitative observations: modularity is less satisfied, especially for perturbations at the fine and image levels (last two rows).

#### 5.1.2. Evolution of distortions with training

As the objective function of a VAE's generator is not directly linked to the multiresolution structure observed previously, the VAE may not naturally enforce modularity between its layers during training. This is suggested by the deterioration of modularity observed after 50000 iterations. To quantify this effect, we tracked the evolution (as the number of iterations grows) of the mean square errors at different wavelet scales (Fig. 4a), as well as for the complete picture (Fig. 4b), resulting from the stretch of the fine level convolutional layer. Interestingly, this difference clearly grows as the training progresses. Looking at the time evolution of the errors at multiple scales, it seems to mostly be due to a progressive rise of the distance in coarser scales. The same overall increasing trend can be observed for the mean squared error of the complete picture. Overall, this suggest that the optimization algorithm does not enforce modularity of the network, and can possibly be improved to take this aspect into account. We thus investigated whether we can encourage more independence from layer to layer during optimization.

[3] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[4] https://github.com/yzwxx/vae-celebA

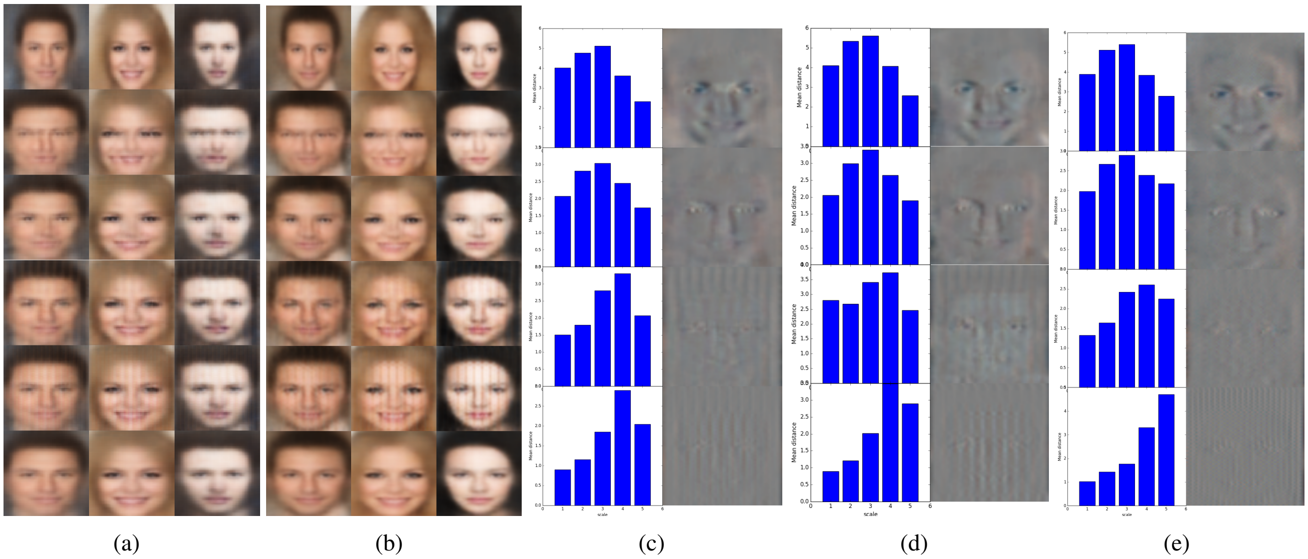


Figure 3: VAE distortion experiment. (a) From top to bottom: normal image generated by the VAE after 10000 training iterations, image resulting from distorting coarse/intermediate/fine/image level layer, stretched orginal. (b) Same as a, for 50000 training iterations. (c) Quantitative analysis for 10000 iterations. Left: Distortion levels at different wavelet scales (1:coarsest, 5:finest). Right: Difference with original stretched. From top to bottom: perturbation on coarse, intermediate, fine and image level, respectively. (d) Same as c for 50000 iterations. (e) Same as d but **with SDR optimization**.

### 5.2. Optimizing modularity

To enforce a better modularity of the network while still optimizing the VAE objective, we trained a VAE for which we alternatively optimized filters of image, fine and intermediate level (leaving one normal VAE iteration between each) by minimizing the sum of the squared scalar product presented in subsection 4.2 in addition to the normal weight update resulting from the VAE objective.

To observe the effect of the new training process over modularity, we repeated the previous experiment and track the mean squared error resulting from the difference between distorted generated pictures and stretched normal images when modifying the fine level convolutional layer when comparing complete pictures (Fig.4c). As can be observed, there is now a significant delay in the time at which the the difference starts rising again. Moreover, the increase, when it starts, seems slower. This is confirmed by the analysis of example images at multiple scales for 50000 iterations, as seen in Fig.3e, deformations at the intermediate, fine and image level exhibit a better modularity, compared to the what was obtain at the same number of iterations (Fig. 3d) with classical VAE training. This supports a link between intrinsic disentanglement and spectral independence between filters and activation maps.

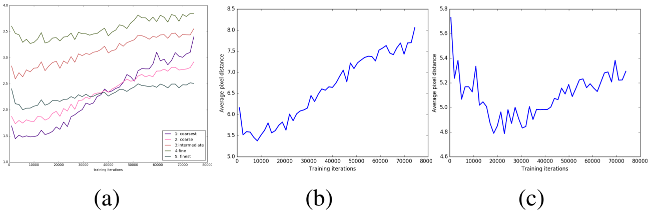


Figure 4: (a) Evolution of the residual error between distorted outputs at the fine level and stretched original image at different wavelet scales during VAE training. (b) Same as a for the complete picture (all scales). (c) Same as b but **with SDR optimization**

## 6. Conclusion

We propose approaching modularity in generative networks using a principle of invariance with respect to transformation of their internal variables. To assess this notion of *intrinsic disentanglement*, we analyzed generative networks as Causal Generative Models and adapt a metric tracking independence of cause and mechanism to evaluate the disentanglement of successive layers of the network. We found evidence of modularity in VAEs trained to generated images of human faces. Moreover, imposing a additional step in the optimization procedure to favor independence between activation maps and filters helped bolster existing modular behavior.

## References

Michel Besserve, Naji Shajarisales, Bernhard Schölkopf, and Dominik Janzing. Group invariance principles for causal generative models. In *AISTATS*, 2018.

X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and

P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

A. Dosovitskiy and T. Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4829–4837, 2016.

R. C. Fong and A. Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296*, 2017.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR 2017*, 2017.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *Information Theory, IEEE Transactions on*, 56(10):5168–5194, 2010.

Mitsuo Kawato, Hideki Hayakawa, and Toshio Inui. A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network: Computation in Neural Systems*, 4(4): 415–422, 1993.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pages 2539–2547, 2015.

J. Lemeire and D. Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, pages 1–23, 7 2012. doi: 10.1007/s11023-012-9283-1.

M. F. Mathieu, J. J. Zhao, A. Ramesh, P. Sprechmann, and Y. LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pages 5041–5049, 2016.

J. Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, 2017.

Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79, 1999.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

S. Sabour, N. Frosst, and G. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.

N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve. Telling cause from effect in deterministic linear dynamical systems. In *ICML 2015*, 2015.

Satohiro Tajima and Masataka Watanabe. Acquisition of nonlinear forward optics in generative models: two-stage ”downside-up” learning for occluded vision. *Neural Netw*, 24(2):148–158, 2011.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Q. Zhang and S. Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.

Q. Zhang, R. Cao, Y. Wu, and S. Zhu. Growing interpretable part graphs on convnets via multi-shot learning. In *AAAI*, pages 2898–2906, 2017a.

Q. Zhang, W. Wang, and S. Zhu. Examining cnn representations with respect to dataset bias. *arXiv preprint arXiv:1710.10577*, 2017b.

# Supplemental information

## Background

### Fourier analysis of discrete signals and images

The Discrete-time Fourier Transform (DTFT) of a sequence $a = \{a[k], k \in \mathbb{Z}\}$ is defined as

$$\widehat{\mathrm{a}}(\nu) = \sum_{k\in\mathbb{Z}} a[k] e^{-\mathbf{i}2\pi\nu k},\ \nu \in \mathbb{R}\,.$$

Note that the DTFT of such sequence is a continuous 1-periodic function of the normalized frequency $\nu$. By Parseval's theorem, the energy (sum of squared coefficients) of the sequence can be expressed in the Fourier domain by $\|a\|_2^2 = \int_{-1/2}^{1/2} |\widehat{\mathrm{a}}(\nu)|^2 d\nu$. The Fourier transform can be easily generalized to 2D signals of the form $\{b[k,l], (k,l) \in \mathbb{Z}^2\}$, leading to a 2D function, 1-periodic with respect to both arguments

$$\widehat{\mathrm{b}}(u,v) = \sum_{k\in\mathbb{Z},l\in\mathbb{Z}} b[k,l] e^{-\mathbf{i}2\pi(uk+vl)},\ (u,v) \in \mathbb{R}^2\,.$$

## Network hyperparameters

Default network hyperparameters are summarized in Table 1 (they apply unless otherwise stated in main text).

| Architecture | VAE |
|---|---|
| Nb. of deconv. layers/channels of generator | 4/(64,64,32,16,3) |
| Size of activation maps of generator | (8,16,32,64) |
| Optimization algorithm | Adam ($\beta = 0.5$) |
| Minimized objective | VAE loss (Gaussian posteriors) |
| batch size | 64 |
| Beta parameter | 0.0005 |

Table 1