Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style

Anonymous Author(s)
Affiliation
Address
email

Abstract

Self-supervised representation learning has shown remarkable success in a number of domains. A common practice is to perform data augmentation via hand-crafted transformations intended to leave the semantics of the data invariant. We seek to understand the empirical success of this approach from a theoretical perspective. We formulate the augmentation process as a latent variable model by postulating a partition of the latent representation into a *content* component, which is assumed invariant to augmentation, and a *style* component, which is allowed to change. Unlike prior work on disentanglement and independent component analysis, we allow for both nontrivial statistical and causal dependencies in the latent space. We study the identifiability of the latent representation based on pairs of views of the observations and prove sufficient conditions that allow us to identify the invariant content partition up to an invertible mapping in both generative and discriminative settings. We find numerical simulations with dependent latent variables are consistent with our theory. Lastly, we introduce *Causal3DIdent*, a dataset of high-dimensional, visually complex images with rich causal dependencies, which we use to study the effect of data augmentations performed in practice.

1 Introduction

Learning good representations of high-dimensional observations from large amounts of unlabelled data is widely recognised as an important step for more capable and data-efficient learning systems [9, 67]. Over the last decade, *self-supervised learning* (SSL) has emerged as the dominant paradigm for such unsupervised representation learning [1, 18, 19, 32, 38, 44, 45, 84, 85, 109, 112, 115, 116]. The main idea behind SSL is to extract a supervisory signal from unlabelled observations by leveraging known structure of the data, which allows for the application of supervised learning techniques. A common approach is to directly predict some part of the observation from another part (e.g., future from past, or original from corruption), thus forcing the model to learn a meaningful representation in the process. While this technique has shown remarkable success in natural language processing [11, 21, 28, 75, 78, 80, 89, 93] and speech recognition [4, 5, 94, 98], where a finite dictionary allows one to output a distribution over the missing part, such *predictive* SSL methods are not easily applied to continuous or high-dimensional domains such as vision. Here, a common approach is to learn a *joint embedding* of similar observations or *views* such that their representation is close [6, 10, 20, 41]. Different views can come, for example, from different modalities (text & speech; video & audio) or time points. As still images lack such multi-modality or temporal structure, recent advances in representation learning have relied on generating similar views by means of *data augmentation*.

In order to be useful, data augmentation is thought to require the transformations applied to generate additional views to be generally chosen to *preserve the semantic characteristics* of an observation, while changing other "nuisance" aspects. While this intuitively makes sense and has shown remarkable empirical results, the success of data augmentation techniques in practice is still not very well understood from a theoretical perspective—despite some efforts [15, 17, 26]. In the present work, we seek to better understand the empirical success of SSL with data augmentation by formulating the generative process as a latent variable model (LVM) and studying *identifiability* of the representation, i.e., under which conditions the ground truth latent factors can provably be inferred from the data [71].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Related work and its relation to the current. Prior work on unsupervised representation learning from an LVM perspective often postulates *mutually independent latent factors*: this independence assumption is, for example, at the heart of independent component analysis (ICA) [22, 53] and disentanglement [9, 12, 16, 46, 61, 66]. Since it is impossible to identify the true latent factors without any supervisory signal in the general nonlinear case [54, 76], recent work has turned to weakly- or self-supervised approaches which leverage additional information in the form of multiple views [36, 77, 102, 119], auxiliary variables [55, 59], or temporal structure [42, 51, 52, 64]. To identify or disentangle the individual independent latent factors, it is typically assumed that there is a chance that *each factor changes* across views, environments, or time points.

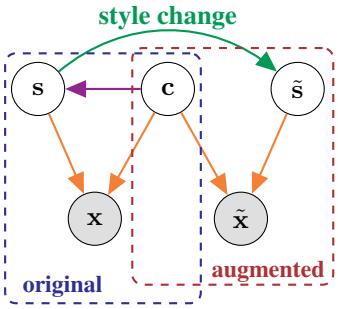


Figure 1: **Overview of our problem formulation.** We partition the latent variable $\mathbf{z}$ into content $\mathbf{c}$ and style $\mathbf{s}$, and allow for statistical and causal dependence of style on content. We assume that only style changes between the original view $\mathbf{x}$ and the augmented view $\tilde{\mathbf{x}}$, i.e., they are obtained by applying the same deterministic function $\mathbf{f}$ to $\mathbf{z} = (\mathbf{c}, \mathbf{s})$ and $\tilde{\mathbf{z}} = (\mathbf{c}, \tilde{\mathbf{s}})$, respectively.

Our work—being directly motivated by common practices in SSL with data augmentation—differs from these works in the following two key aspects (see Fig. 1 for an overview). First, we do not assume independence and instead *allow for both non-trivial statistical and causal relations between latent variables*. This is in line with a recently proposed [99] shift towards causal representation learning [70, 79, 81, 100, 101, 106, 113, 117], motivated by the fact that many underlying variables of interest may not be independent but causally related to each other.[1] Second, instead of a scenario wherein all latent factors may change as a result of augmentation, we assume a *partition of the latent space* into two blocks: a *content* partition which is shared or *invariant* across different augmented views, and a *style* partition that *may change*. This is aligned with the notion that augmentations leave certain semantic aspects (i.e., content) intact and only affect style, and is thus a more appropriate assumption for studying SSL.

Structure and contributions. Following a review of SSL with data augmentation and identifiability theory (§ 2), we formalise the process of data generation and augmentation as an LVM with content and style variables (§ 3). We then establish identifiability results of the invariant content partition (§ 4), validate our theoretical insights experimentally (§ 5), and discuss our findings in the broader context of SSL with data augmentation (§ 6). We highlight the following contributions:

- we prove that SSL with data augmentations identifies the invariant content partition of the representation in generative (Thm. 4.2) and discriminative learning with invertible (Thm. 4.3) and non-invertible encoders with entropy regularisation (Thm. 4.4); in particular, Thm. 4.4 provides a theoretical justification for the empirically observed effectiveness of contrastive SSL methods that use data augmentation and InfoNCE [85] as an objective, such as `SimCLR` [18];
- we show that our theory is consistent with results in simulating statistical dependencies within blocks of content and style variables, as well as with style causally dependent on content (§ 5.1);
- we introduce *Causal3DIdent*, a dataset of 3D objects which allows for the study of identifiability in a causal representation learning setting, and use it to perform a systematic study of data augmentations used in practice, yielding novel insights on what particular data augmentations are truly isolating as content and discarding as style when applied (§ 5.2).

2 Preliminaries and background

Self-supervised representation learning with data augmentation. Given an unlabelled dataset of observations (e.g., images) $\mathbf{x}$, data augmentation techniques proceed as follows. First, a set of observation-level transformations $\mathbf{t} \in \mathcal{T}$ are specified together with a distribution $p_{\mathbf{t}}$ over $\mathcal{T}$. Both $\mathcal{T}$ and $p_{\mathbf{t}}$ are typically designed using human intelligence and domain knowledge with the intention of *not changing the semantic characteristics* of the data (which arguably constitutes a form of weak supervision).[2] For images, for example, a common choice for $\mathcal{T}$ are combinations of random crops [107], horizontal or vertical flips, blurring, colour distortion [49, 107], or cutouts [29]; and $p_{\mathbf{t}}$ is a distribution over the parameterisation of these transformations, e.g., the centre and size of a crop [18, 29]. For each observation $\mathbf{x}$, a pair of transformations $\mathbf{t}, \mathbf{t}' \sim p_{\mathbf{t}}$ is sampled and applied separately to $\mathbf{x}$ to generate a pair of augmented views $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = (\mathbf{t}(\mathbf{x}), \mathbf{t}'(\mathbf{x}))$.

[1] E.g., [64], Fig. 11 where dependence between latents was demonstrated for multiple natural video data sets.

[2] Note that recent work has investigated automatically discovering good augmentations [24, 25].

The joint-embedding approach to SSL then uses a pair of encoder functions $(\mathbf{g}, \mathbf{g}')$, i.e. deep nets, to map the pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ to a typically lower-dimensional representation $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}') = (\mathbf{g}(\tilde{\mathbf{x}}), \mathbf{g}'(\tilde{\mathbf{x}}'))$. Often, the two encoders are either identical, $\mathbf{g} = \mathbf{g}'$, or directly related (e.g., via shared parameters or asynchronous updates). Then, the encoder(s) $(\mathbf{g}, \mathbf{g}')$ are trained such that the representations $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$ are "close", i.e., such that $\text{sim}(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$ is large for some similarity metric $\text{sim}(\cdot)$, e.g., the cosine similarity [18, 119], or negative L2 norm [119]. The advantage of directly optimising for similarity in representation space over generative alternatives is that reconstruction can be very challenging for high-dimensional data. The disadvantage is the problem of *collapsed representations*.[3] To avoid collapsed representations and force the encoder(s) to learn a meaningful representation, two main families of approaches have been used: (i) *contrastive learning* (CL) [18, 44, 45, 85, 109, 116]; and (ii) *regularisation-based* SSL [19, 38, 118].

The idea behind CL is to not only learn similar representations for augmented views $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}'_i)$ of the same $\mathbf{x}_i$, or *positive pairs*, but to also use other observations $\mathbf{x}_j$ $(j \neq i)$ to contrast with, i.e., to enforce a dissimilar representation across *negative pairs* $(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}'_j)$. In other words, CL pulls representations of positive pairs together, and pushes those of negative pairs apart. Since both aims cannot be achieved simultaneously with a constant representation, collapse is avoided. A popular CL objective function (used, e.g., in `SimCLR` [18][4]) is InfoNCE [85] (based on noise-contrastive estimation [39, 40]):

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}; \tau, K) = \mathbb{E}_{\{\mathbf{x}_i\}_{i=1}^K \sim p_{\mathbf{x}}} \Big[-\sum_{i=1}^K \log \frac{\exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_i)/\tau\}}{\sum_{j=1}^K \exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_j)/\tau\}} \Big] \tag{1}$$

where $\tilde{\mathbf{z}} = \mathbb{E}_{\mathbf{t} \sim p_{\mathbf{t}}}[\mathbf{g}(\mathbf{t}(\mathbf{x}))]$, τ is a temperature, and $K-1$ is the number of negative pairs. InfoNCE (1) has an interpretation as multi-class logistic regression, and lower bounds the mutual information across similar views $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$—a common representation learning objective [3, 8, 13, 47, 69, 73, 74, 91, 110]. Moreover, (1) can be interpreted as *alignment* (numerator) and *uniformity* (denominator) terms, the latter constituting a nonparametric entropy estimator of the representation as $K \to \infty$ [114]. CL with InfoNCE can thus be seen as alignment of positive pairs with (approximate) entropy regularisation.

Instead of using negative pairs, as in CL, a set of recent SSL methods only optimise for alignment and avoid collapsed representations through different forms of regularisation. For example, `BYOL` [38] and `SimSiam` [19] rely on "architectural regularisation" in the form of moving-average updates for a separate "target" net $\mathbf{g}'$ (`BYOL` only) or a stop-gradient operation (both). `BarlowTwins` [118], on the other hand, optimises the cross correlation between $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}')$ to be close to the identity matrix, thus enforcing redundancy reduction (zero off-diagonals) in addition to alignment (ones on the diagonal).

Identifiability of learned representations. In this work, we address the question of whether SSL with data augmentation can reveal or uncover properties of the underlying data generating process. Whether a representation learned from observations can be expected to match the true underlying latent factors—up to acceptable ambiguities and subject to suitable assumptions on the generative process and inference model—is captured by the notion of identifiability [71].

Within representation learning, identifiability has mainly been studied in the framework of (nonlinear) ICA which assumes a model of the form $\mathbf{x} = \mathbf{f}(\mathbf{z})$ and aims to recover the independent latents, or *sources*, $\mathbf{z}$, typically up to permutation or element-wise transformation. A crucial negative result states that, with i.i.d. data and without further assumptions, this is fundamentally impossible [54]. However, recent breakthroughs have shown that identifiability can be achieved if an auxiliary variable (e.g., a time stamp or environment index) renders the sources *conditionally* independent [42, 51, 52, 55]. These methods rely on constructing positive and negative pairs using the auxiliary variable and learning a representation with CL. This development has sparked a renewed interest in identifiability in the context of deep representation learning [59, 60, 64, 77, 96, 102, 103, 119].

Most closely related to SSL with data augmentation are works which study identifiability when given a second view $\tilde{\mathbf{x}}$ of an observation $\mathbf{x}$, resulting from a modified version $\tilde{\mathbf{z}}$ of the underlying latents or sources $\mathbf{z}$ [36, 64, 77, 95, 102, 119]. Here, $\tilde{\mathbf{z}}$ is either an element-wise corruption of $\mathbf{z}$ [36, 64, 95, 119] or may share a random subset of its components [77, 102]. Crucially, all previously mentioned works assume that *any* of the independent latents (are allowed to) change, and aim to identify the individual factors. However, in the context of SSL with data augmentation, where the semantic (content) part of the representation is intended to be shared between views, this assumption does not hold.

[3] If the only goal is to make representations of augmented views similar, a degenerate solution which simply maps any observation to the origin trivially achieves this goal.

[4] `SimCLR` uses a slight modification to InfoNCE in which the sum in the denominator is only over $j \neq i$.

3 Problem formulation

We specify our problem setting by formalising the processes of data generation and augmentation. We take a latent-variable model perspective and assume that observations $\mathbf{x}$ (e.g., images) are generated by a *mixing* function $\mathbf{f}$ which takes a latent code $\mathbf{z}$ as input. Importantly, we describe the augmentation process through changes in this latent space as captured by a conditional distribution $p_{\tilde{\mathbf{z}}|\mathbf{z}}$, as opposed to traditionally describing the transformations $\mathbf{t}$ as acting directly at the observation level.

Formally, let $\mathbf{z}$ be a continuous r.v. taking values in an open, simply-connected n-dim. *representation space* $\mathcal{Z} \subseteq \mathbb{R}^n$ with associated probability density $p_{\mathbf{z}}$. Moreover, let $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ be a *smooth and invertible* mapping to an *observation space* $\mathcal{X} \subseteq \mathbb{R}^d$ and let $\mathbf{x}$ be the continuous r.v. defined as $\mathbf{x} = \mathbf{f}(\mathbf{z})$.[5] The generative process for the dataset of original observations of $\mathbf{x}$ is thus given by:

$$\mathbf{z} \sim p_{\mathbf{z}}, \qquad \mathbf{x} = \mathbf{f}(\mathbf{z}). \tag{2}$$

Next, we formalise the data augmentation process. As stated above, we take a representation-centric view, i.e., we assume that an augmentation $\tilde{\mathbf{x}}$ of the original $\mathbf{x}$ is obtained by applying the same mixing or rendering function $\mathbf{f}$ to a modified representation $\tilde{\mathbf{z}}$ which is (stochastically) related to the original representation $\mathbf{z}$ of $\mathbf{x}$. Specifying the effect of data augmentation thus corresponds to specifying a conditional distribution $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ which captures the relation between $\mathbf{z}$ and $\tilde{\mathbf{z}}$.

In terms of the transformation-centric view presented in § 2, we can view the modified representation $\tilde{\mathbf{z}} \in \mathcal{Z}$ as obtained by applying $\mathbf{f}^{-1}$ to a transformed observation $\tilde{\mathbf{x}} = \mathbf{t}(\mathbf{x}) \in \mathcal{X}$ where $\mathbf{t} \sim p_{\mathbf{t}}$, i.e., $\tilde{\mathbf{z}} = \mathbf{f}^{-1}(\tilde{\mathbf{x}})$. The conditional distribution $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ in the representation space can thus be viewed as being induced by the distribution $p_{\mathbf{t}}$ over transformations applied at the observation level.[6]

We now encode the notion that the set of transformations $\mathcal{T}$ used for augmentation is typically chosen such that any transformation $\mathbf{t} \in \mathcal{T}$ leaves certain aspects of the data invariant. To this end, we assume that *the representation* $\mathbf{z}$ *can be uniquely partitioned into two disjoint parts*:

(i) an *invariant* part $\mathbf{c}$ which will *always be shared* across $(\mathbf{z}, \tilde{\mathbf{z}})$, and which we refer to as *content*;
(ii) a *varying* part $\mathbf{s}$ which *may change* across $(\mathbf{z}, \tilde{\mathbf{z}})$, and which we refer to as *style*.

We assume that $\mathbf{c}$ and $\mathbf{s}$ take values in content and style subspaces $\mathcal{C} \subseteq \mathbb{R}^{n_c}$ and $\mathcal{S} \subseteq \mathbb{R}^{n_s}$, respectively, i.e., $n = n_c + n_s$ and $\mathcal{Z} = \mathcal{C} \times \mathcal{S}$. W.l.o.g., we let $\mathbf{c}$ corresponds to the first n_c dimensions of $\mathbf{z}$:

$$\mathbf{z} = (\mathbf{c}, \mathbf{s}), \qquad \mathbf{c} := \mathbf{z}_{1:n_c}, \qquad \mathbf{s} := \mathbf{z}_{(n_c+1):n},$$

We formalise the process of data augmentation with content-preserving transformations by defining the conditional $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ such that only a (random) subset of the style variables change at a time.

Assumption 3.1 (Content-invariance)**.** The conditional density $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ over $\mathcal{Z} \times \mathcal{Z}$ takes the form

$$p_{\tilde{\mathbf{z}}|\mathbf{z}}(\tilde{\mathbf{z}}|\mathbf{z}) = \delta(\tilde{\mathbf{c}} - \mathbf{c}) p_{\tilde{\mathbf{s}}|\mathbf{s}}(\tilde{\mathbf{s}}|\mathbf{s})$$

for some continuous density $p_{\tilde{\mathbf{s}}|\mathbf{s}}$ on $\mathcal{S} \times \mathcal{S}$, where $\delta(\cdot)$ is the Dirac delta function, i.e., $\tilde{\mathbf{c}} = \mathbf{c}$ a.e.

Assumption 3.2 (Style changes)**.** Let $\mathcal{A}$ be the set of subsets of style variables $A \subseteq \{1, ..., n_s\}$ and let p_A be a distribution on $\mathcal{A}$. Then, the style conditional $p_{\tilde{\mathbf{s}}|\mathbf{s}}$ is obtained via

$$A \sim p_A, \qquad p_{\tilde{\mathbf{s}}|\mathbf{s},A}(\tilde{\mathbf{s}}|\mathbf{s}, A) = \delta(\tilde{\mathbf{s}}_{A^c} - \mathbf{s}_{A^c}) p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\tilde{\mathbf{s}}_A|\mathbf{s}_A)\,,$$

where $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is a continuous density on $\mathcal{S}_A \times \mathcal{S}_A$, $\mathcal{S}_A \subseteq \mathcal{S}$ denotes the subspace of changing style variables specified by A, and $A^c = \{1, ..., n_s\} \setminus A$ denotes the complement of A.

Note that Assumption 3.2 is less restrictive than assuming that all style variables need to change, since it also allows for only a (possibly different) subset of style variables to change for any given observation. This is in line with the intuition that not all transformations affect all changeable (i.e., style) properties of the data: e.g., a colour distortion should not affect positional information, and, in the same vein, a (horizontal or vertical) flip should not affect the colour spectrum.

The generative process of an augmentation or transformed observation $\tilde{\mathbf{x}}$ is thus given by

$$A \sim p_A, \qquad \tilde{\mathbf{z}}|\mathbf{z}, A \sim p_{\tilde{\mathbf{z}}|\mathbf{z},A}, \qquad \tilde{\mathbf{x}} = \mathbf{f}(\tilde{\mathbf{z}}). \tag{3}$$

Our setting for modelling data augmentation differs from that commonly assumed in (multi-view) disentanglement and ICA in that *we do not assume that the latent factors* $\mathbf{z} = (\mathbf{c}, \mathbf{s})$ *are mutually (or conditionally) independent*, i.e., we allow for *arbitrary* (non-factorised) marginals $p_{\mathbf{z}}$ in (2).[7]

[5] While $\mathbf{x}$ may be high-dimensional $n \ll d$, invertibility of $\mathbf{f}$ implies that $\mathcal{X}$ is an n-dim. sub-manifold of $\mathbb{R}^d$.

[6] We investigate this correspondence between changes in observation and latent space empirically in § 5.

[7] The recently proposed Independently *Modulated* Component Analysis [60] extension of ICA is a notable exception, but only allows for trivial dependencies across $\mathbf{z}$ in the form of a shared base measure.

Causal interpretation: data augmentation as counterfactuals under soft style intervention. We now provide a causal account of the above data generating process by describing the (allowed) causal dependencies among latent variables using a structural causal model (SCM) [88]. As we will see, this leads to an interpretation of data augmentations as counterfactuals in the underlying latent SCM. The assumption that $\mathbf{c}$ stays invariant as $\mathbf{s}$ changes is consistent with the view that content may causally influence style, $\mathbf{c} \to \mathbf{s}$, but not vice versa, see Fig. 1. We therefore formalise their relation as:

$$\mathbf{c} := \mathbf{f}_{\mathbf{c}}(\mathbf{u}_{\mathbf{c}}), \qquad \mathbf{s} := \mathbf{f}_{\mathbf{s}}(\mathbf{c}, \mathbf{u}_{\mathbf{s}}), \qquad (\mathbf{u}_{\mathbf{c}}, \mathbf{u}_{\mathbf{s}}) \sim p_{\mathbf{u}_{\mathbf{c}}} \times p_{\mathbf{u}_{\mathbf{s}}}$$

where $\mathbf{u}_{\mathbf{c}}, \mathbf{u}_{\mathbf{s}}$ are independent exogenous variables, and $\mathbf{f}_{\mathbf{c}}, \mathbf{f}_{\mathbf{s}}$ are deterministic functions. The latent causal variables $(\mathbf{c}, \mathbf{s})$ are subsequently decoded into observations $\mathbf{x} = \mathbf{f}(\mathbf{c}, \mathbf{s})$. Given a factual observation $\mathbf{x}^{\mathtt{F}} = \mathbf{f}(\mathbf{c}^{\mathtt{F}}, \mathbf{s}^{\mathtt{F}})$ which resulted from $(\mathbf{u}_{\mathbf{c}}^{\mathtt{F}}, \mathbf{u}_{\mathbf{s}}^{\mathtt{F}})$, we may ask the counterfactual question: "*what would have happened if the style variables had been (randomly) perturbed, all else being equal?*". Consider, e.g., a *soft intervention* [33] on $\mathbf{s}$, i.e., an intervention that changes the mechanism $\mathbf{f}_{\mathbf{s}}$ to

$$do(\mathbf{s} := \tilde{\mathbf{f}}_{\mathbf{s}}(\mathbf{c}, \mathbf{u}_{\mathbf{s}}, \mathbf{u}_A)),$$

where $\mathbf{u}_A$ is an additional source of stochasticity accounting for the randomness of the augmentation process ($p_A \times p_{\tilde{\mathbf{s}}|\mathbf{s},A}$). The resulting distribution over counterfactual observations $\mathbf{x}^{\mathtt{CF}} = \mathbf{f}(\mathbf{c}^{\mathtt{F}}, \mathbf{s}^{\mathtt{CF}})$ can be computed from the modified SCM by fixing the exogenous variables to their factual values and performing the soft intervention:

$$\mathbf{c}^{\mathtt{CF}} := \mathbf{c}^{\mathtt{F}}, \qquad \mathbf{s}^{\mathtt{CF}} := \tilde{\mathbf{f}}_{\mathbf{s}}(\mathbf{c}^{\mathtt{F}}, \mathbf{u}_{\mathbf{s}}^{\mathtt{F}}, \mathbf{u}_A), \qquad \mathbf{u}_A \sim p_{\mathbf{u}_A}.$$

This aligns with our intuition and assumed problem setting of data augmentations as style corruptions. We note that the notion of augmentation as (hard) style interventions is also at the heart of `ReLIC` [81], a recently proposed, causally-inspired SSL regularisation term for instance-discrimination [41, 116]. However, `ReLIC` assumes independence between content and style and does not address identifiability.

4 Theory: block-identifiability of the invariant content partition

Our goal is to prove that we can identify the invariant content partition $\mathbf{c}$ under a distinct, weaker set of assumptions, compared to existing results in disentanglement and nonlinear ICA [36, 64, 77, 102, 119]. We stress again that our primary interest is not to identify or disentangle individual (and independent) latent factors z_j, but instead to separate content from style, such that the content variables can be subsequently used for downstream tasks. We first define this distinct notion of *block-identifiability*.

Definition 4.1 (Block-identifiability)**.** We say that the true content partition $\mathbf{c} = \mathbf{f}^{-1}(\mathbf{x})_{1:n_c}$ is *block-identified* by a function $\mathbf{g} : \mathcal{X} \to \mathcal{Z}$ if the inferred content partition $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})_{1:n_c}$ contains *all* and *only* information about $\mathbf{c}$, i.e., if there exists an *invertible* function $\mathbf{h} : \mathbb{R}^{n_c} \to \mathbb{R}^{n_c}$ s.t. $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{c})$.

Defn. 4.1 is related to independent subspace analysis [14, 50, 68, 108], which also aims to identify blocks of random variables as opposed to individual factors, though under an *independence assumption across blocks*, and typically not within a multi-view setting as studied in the present work.

4.1 Generative self-supervised representation learning

First, we consider *generative* SSL, i.e., fitting a generative model to pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ of original and augmented views.[8] We show that under our specified data generation and augmentation process (§ 3), as well as suitable additional assumptions (stated and discussed in more detail below), it is possible to isolate (i.e., block-identify) the invariant content partition. Full proofs are included in Appendix A.

Theorem 4.2 (Identifying content with a generative model)**.** *Consider the data generating process described in § 3, i.e., the pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ of original and augmented views are generated according to (2) and (3) with $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ as defined in Assumptions 3.1 and 3.2. Assume further that*

(i) $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ is smooth and invertible with smooth inverse (i.e., a diffeomorphism);

(ii) $p_{\mathbf{z}}$ is a smooth, continuous density on $\mathcal{Z}$ with $p_{\mathbf{z}}(\mathbf{z}) > 0$ almost everywhere;

(iii) for any style coordinate $l \in \{1, ..., n_s\}$, $\exists A \subseteq \{1, ..., n_s\}$ s.t. $l \in A$ and $p_A(A) > 0$;

(iv) for any non-empty subset $A \subseteq \{1, ..., n_s\}$ s.t. $p_A(A) > 0$, the style conditional $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}$ is a smooth, continuous density on $\mathcal{S}_A \times \mathcal{S}_A$ with $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\tilde{\mathbf{s}}_A|\mathbf{s}_A) > 0$ almost everywhere.

If, for a given n_s ($1 \leq n_s < n$), a generative model $(\hat{p}_{\mathbf{z}}, \hat{p}_A, \hat{p}_{\tilde{\mathbf{s}}|\mathbf{s},A}, \hat{\mathbf{f}})$ assumes the same generative process (§ 3), satisfies the above assumptions (i)-(iv), and matches the data likelihood,

$$p_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) = \hat{p}_{\mathbf{x},\tilde{\mathbf{x}}}(\mathbf{x}, \tilde{\mathbf{x}}) \qquad \forall (\mathbf{x}, \tilde{\mathbf{x}}) \in \mathcal{X} \times \mathcal{X},$$

then it block-identifies the true content variables via $\mathbf{g} = \hat{\mathbf{f}}^{-1}$ in the sense of Defn. 4.1.

[8] For notational simplicity, we present our theory for pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ rather than for two augmented views $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$, as typically used in practice—see § 6 for further discussion.

Proof sketch. First, show (using *(i)* and the matching likelihoods) that the representation $\hat{\mathbf{z}} = \mathbf{g}(\mathbf{x})$ extracted by $\mathbf{g}$ is related to the true $\mathbf{z}$ by a smooth invertible mapping $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ such that $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{z})_{1:n_c}$ is invariant across $(\mathbf{z}, \tilde{\mathbf{z}})$ almost surely w.r.t. $p_{\mathbf{z},\tilde{\mathbf{z}}}$.[9] Second, show by contradiction (using *(ii)-(iv)*) that $\mathbf{h}(\cdot)_{1:n_c}$ can, in fact, only depend on the true content $\mathbf{c}$ and not on style $\mathbf{s}$, for otherwise the invariance from step 1 would be violated in a region of the style (sub)space of measure greater than zero.

Intuition. Thm. 4.2 assumes that the number of content (n_c) and style (n_s) variables is known, and that there is a positive probability that each *style* variable may change, though not necessarily on its own, according to *(iii)*. In this case, training a generative model of the form specified in § 3 (i.e., with an invariant content partition and subsets of changing style variables) by maximum likelihood on pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ will asymptotically (in the limit of infinite data) recover the true invariant content partition up to an invertible function, i.e., it isolates, or unmixes, content from style.

Discussion. The identifiability result of Thm. 4.2 for generative SSL is of potential relevance for existing variational autoencoder (VAE) [63] variants such as the `GroupVAE` [48],[10] or its adaptive version `AdaGVAE` [77]. Since, contrary to existing results, Thm. 4.2 does not assume independent latents, it may also provide a principled basis for generative causal representation learning algorithms [70, 101, 117]. However, an important limitation to its practical applicability is that generative modelling does not tend to scale very well to complex high-dimensional observations, such as images.

4.2 Discriminative self-supervised representation learning

We therefore next turn to a discriminative approach, i.e., directly learning an encoder function $\mathbf{g}$ which leads to a similar embedding across $(\mathbf{x}, \tilde{\mathbf{x}})$. As discussed in § 2, this is much more common for SSL with data augmentations. First, we show that if an invertible encoder $\mathbf{g}$ is used, then learning a representation which is aligned in the first n_c dimensions is sufficient to block-identify content.

Theorem 4.3 (Identifying content with an invertible encoder)**.** *Assume the same data generating process (§ 3) and conditions (i)-(iv) as in Thm. 4.2. Let* $\mathbf{g} : \mathcal{X} \to \mathcal{Z}$ *be any smooth and* invertible *function which minimises the following functional:*

$$\mathcal{L}_{\text{Align}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left(\mathbf{g}(\mathbf{x})_{1:n_c} - \mathbf{g}(\tilde{\mathbf{x}})_{1:n_c} \right)^2 \right] \tag{4}$$

Then $\mathbf{g}$ *block-identifies the true content variables in the sense of Definition 4.1.*

Proof sketch. First, we show that the global minimum of (4) is reached by the smooth invertible function $\mathbf{f}^{-1}$. Thus, any other minimiser $\mathbf{g}$ must satisfy the same invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ used in step 1 of the proof of Thm. 4.2. The second step uses the same argument by contradiction as in Thm. 4.2.

Intuition. Thm. 4.3 states that if—under the same assumptions on the generative process as in Thm. 4.2—we directly learn a representation with an *invertible* encoder, then enforcing alignment between the first n_c latents is sufficient to isolate the invariant content partition. Intuitively, invertibility guarantees that all information is preserved, thus avoiding a collapsed representation.

Discussion. According to Thm. 4.3, content can be isolated if, e.g., a flow-based architecture [30, 31, 62, 86, 87] is used, or invertibility is enforced otherwise during training [7, 56]. However, the applicability of this approach is limited as it *places strong constraints on the encoder architecture which makes it hard to scale these methods up to high-dimensional settings*. As discussed in § 2, state-of-the-art SSL methods such as `SimCLR` [18], `BYOL` [38], `SimSiam` [19], or `BarlowTwins` [118] do not use invertible encoders, but instead avoid collapsed representations—which would result from naively optimising (4) for arbitrary, non-invertible $\mathbf{g}$—using different forms of regularisation.

To close this gap between theory and practice, finally, we investigate how to block-identify content without assuming an invertible encoder. We show that, if we add a regularisation term to (4) that encourages maximum entropy of the learnt representation, the invertibility assumption can be dropped.

Theorem 4.4 (Identifying content with discriminative learning and a non-invertible encoder)**.** *Assume the same data generating process (§ 3) and conditions (i)-(iv) as in Thm. 4.2. Let* $\mathbf{g} : \mathcal{X} \to (0,1)^{n_c}$ *be any smooth function which minimises the following functional:*

$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\left(\mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}}) \right)^2 \right] - H\left(\mathbf{g}(\mathbf{x})\right) \tag{5}$$

where $H(\cdot)$ *denotes the differential entropy of the random variable* $\mathbf{g}(\mathbf{x})$ *taking values in* $(0,1)^{n_c}$. *Then* $\mathbf{g}$ *block-identifies the true content variables in the sense of Defn. 4.1.*

[9] This step is partially inspired by [77]; the technique used to prove the second *main* step is entirely novel.

[10] which also uses a fixed content-style partition for multi-view data, but assumes that all latent factors are mutually independent, and that all style variables change between views, independent of the original style;

Proof sketch. First, use the Darmois construction [27, 54] to build a function $\mathbf{d} : \mathcal{C} \to (0,1)^{n_c}$ mapping $\mathbf{c} = \mathbf{f}^{-1}(\mathbf{x})_{1:n_c}$ to a uniform random variable. Then $\mathbf{g}^\star = \mathbf{d} \circ \mathbf{f}^{-1}_{1:n_c}$ attains the global minimum of (5) because $\mathbf{c}$ is invariant across $(\mathbf{x}, \tilde{\mathbf{x}})$ and the uniform distribution is the maximum entropy distribution on $(0,1)^{n_c}$. Thus, any other minimiser $\mathbf{g}$ of (5) must satisfy invariance across $(\mathbf{x}, \tilde{\mathbf{x}})$ and map to a uniform r.v. Then, use the same step 2 as in Thms. 4.2 and 4.3 to show that $\mathbf{h} = \mathbf{g} \circ \mathbf{f}^{-1} : \mathcal{Z} \to (0,1)^{n_c}$ cannot depend on style, i.e., it is a function from $\mathcal{C}$ to $(0,1)^{n_c}$. Finally, we show that $\mathbf{h}$ must be invertible since it maps $p_\mathbf{c}$ to a uniform distribution, using a result from [119].

Intuition. Thm. 4.4 states that if we do not explicitly enforce invertibility of $\mathbf{g}$ as in Thm. 4.3, additionally maximising the entropy of the learnt representation (i.e., optimising alignment *and* uniformity [114]) avoids a collapsed representation and recovers the invariant content block. Intuitively, this is because any function that only depends on $\mathbf{c}$ will be invariant across $(\mathbf{x}, \tilde{\mathbf{x}})$, so it is beneficial to preserve all content information to maximise entropy.

Discussion. Of our theoretical results, Thm. 4.4 requires the weakest set of assumptions, and is most closely aligned with common SSL practice. As discussed in § 2, contrastive SSL with negative samples using InfoNCE (1) as an objective can asymptotically be understood as alignment with entropy regularisation [114], i.e., objective (5). *Thm. 4.4 thus provides a theoretical justification for the empirically observed effectiveness of CL with InfoNCE*: subject to our assumptions, CL with InfoNCE asymptotically isolates content, i.e., the part of the representation that is always left invariant by augmentation. For example, the strong image classification performance based on representations learned by `SimCLR` [18], which uses color distortion and random crops as augmentations, can be explained in that object class is a content variable in this case. We extensively evaluate the effect of various augmentation techniques on different ground-truth latent factors in our experiments in § 5. There is also an interesting connection between Thm. 4.4 and `BarlowTwins` [118], which only uses positive pairs and combines alignment with a redundancy reduction regulariser that enforces decorrelation between the inferred latents. Intuitively, redundancy reduction is related to increased entropy: $\mathbf{g}^\star$ constructed in the proof of Thm. 4.4—and thus also any other minimiser of (5)—attains the global optimum of the `BarlowTwins` objective, though the reverse implication may not hold.

5 Experiments

We perform two sets of experiments. First, we numerically test our main result, Thm. 4.4, in a fully-controlled, finite sample setting, using CL to estimate the entropy term in (5). Second, we seek to better understand the effect of data augmentations used in practice. To this end, we introduce a new dataset of 3D objects with rich dependencies between a number of known ground-truth factors, and use it to evaluate the effect of different data augmentation techniques on what is identified as content.

5.1 Numerical data

Experimental setup. We generate synthetic data as described in § 3. We consider $n_c = n_s = 5$, with content and style latents distributed as $\mathbf{c} \sim \mathcal{N}(0, \Sigma_c)$ and $\mathbf{s}|\mathbf{c} \sim \mathcal{N}(\mathbf{a} + B\mathbf{c}, \Sigma_s)$, thus allowing for *statistical dependence* within the two blocks (via Σ_c and Σ_s) and *causal dependence* between content and style (via B). For $\mathbf{f}$, we use a 3-layer MLP with LeakyReLU activation functions.[11] The distribution p_A over subsets of changing style variables is obtained by independently flipping the same biased coin for each s_i. The conditional style distribution is taken as $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A} = \mathcal{N}(\mathbf{s}_A, \Sigma_A)$. We train an encoder $\mathbf{g}$ on pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ with InfoNCE using the negative L2 loss as the similarity measure, i.e., we approximate (5) using empirical averages and negative samples. For evaluation, we use kernel ridge regression [82] to predict the ground truth $\mathbf{c}$ and $\mathbf{s}$ from the learnt representation $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})$ and report the R^2 coefficient of determination. For a more detailed account, we refer to Appendix D.

Generative process			R^2 (nonlinear)	
p(chg.)	Stat.	Cau.	Content $\mathbf{c}$	Style $\mathbf{s}$
1.0	✗	✗	**1.00** $\pm$ 0.00	0.07 $\pm$ 0.00
0.75	✗	✗	**1.00** $\pm$ 0.00	0.06 $\pm$ 0.05
0.75	✓	✗	**0.99** $\pm$ 0.00	0.02 $\pm$ 0.04
0.75	✓	✓	**0.96** $\pm$ 0.00	0.00 $\pm$ 0.00

Results. In the inset table, we report mean $\pm$ std. dev. over 3 random seeds across four generative processes of increasing complexity covered by Thm. 4.4: "p(chg.)", "Stat.", and "Cau." denote respectively the change probability for each s_i, statistical dependence within blocks ($\Sigma_c \neq I \neq \Sigma_s$), and causal dependence of style on content ($B \neq 0$). An R^2 close to one indicates that almost all variation is explained by $\hat{\mathbf{c}}$, i.e., that there is a 1-1 mapping, as required by Defn. 4.1. As can be seen, *across all settings, content is block-identified and style is not.* Moreover, we only observe a slight drop-off in performance with increasing complexity of the generative process, i.e., richer dependencies across latents. Finally, we show in Appendix C that a high R^2 score can be obtained even if we use linear regression to predict $\mathbf{c}$ from $\hat{\mathbf{c}}$ ($R^2 = 0.93 \pm 0.04$, for the last row).

[11] chosen to lead to invertibility almost surely by following the settings used by previous work [51, 52]

Table 1: *Causal3DIdent* results: R^2 mean $\pm$ std. dev. over 3 random seeds. DA: data augmentation, LT: latent transformation, bold: $R^2 \geq 0.5$, red: $R^2 < 0.25$. Results for individual axes of object position & rotation are aggregated, see Appendix C for the full table.

Views generated by	**Class**	**Positions**		**Hues**			**Rotations**
		object	spotlight	object	spotlight	background	
DA: colour distortion	0.42 ± 0.01	$\mathbf{0.61} \pm 0.10$	0.17 ± 0.00	0.10 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	0.33 ± 0.02
LT: change hues	$\mathbf{1.00} \pm 0.00$	$\mathbf{0.59} \pm 0.33$	$\mathbf{0.91} \pm 0.00$	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.30 ± 0.01
DA: crop (large)	0.28 ± 0.04	0.09 ± 0.08	0.21 ± 0.13	$\mathbf{0.87} \pm 0.00$	0.09 ± 0.02	$\mathbf{1.00} \pm 0.00$	0.02 ± 0.02
DA: crop (small)	0.14 ± 0.00	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{1.00} \pm 0.00$	0.00 ± 0.00
LT: change positions	$\mathbf{1.00} \pm 0.00$	0.16 ± 0.23	0.00 ± 0.01	0.46 ± 0.02	0.00 ± 0.00	$\mathbf{0.97} \pm 0.00$	0.29 ± 0.01
DA: crop (large) + colour distortion	$\mathbf{0.97} \pm 0.00$	$\mathbf{0.59} \pm 0.07$	$\mathbf{0.59} \pm 0.05$	0.28 ± 0.00	0.01 ± 0.01	0.01 ± 0.00	$\mathbf{0.74} \pm 0.03$
DA: crop (small) + colour distortion	$\mathbf{1.00} \pm 0.00$	$\mathbf{0.69} \pm 0.04$	$\mathbf{0.93} \pm 0.00$	0.30 ± 0.01	0.00 ± 0.00	0.02 ± 0.03	$\mathbf{0.56} \pm 0.03$
LT: change positions + hues	$\mathbf{1.00} \pm 0.00$	0.22 ± 0.22	0.07 ± 0.08	0.32 ± 0.02	0.00 ± 0.01	0.02 ± 0.03	0.34 ± 0.06
DA: rotation	0.33 ± 0.06	0.17 ± 0.09	0.23 ± 0.12	$\mathbf{0.83} \pm 0.01$	0.30 ± 0.12	$\mathbf{0.99} \pm 0.00$	0.05 ± 0.03
LT: change rotations	$\mathbf{1.00} \pm 0.00$	$\mathbf{0.53} \pm 0.33$	$\mathbf{0.90} \pm 0.00$	0.41 ± 0.00	0.00 ± 0.00	$\mathbf{0.97} \pm 0.00$	0.28 ± 0.00
DA: rotation + colour distortion	$\mathbf{0.59} \pm 0.01$	$\mathbf{0.58} \pm 0.06$	0.21 ± 0.01	0.12 ± 0.02	0.01 ± 0.00	0.01 ± 0.00	0.33 ± 0.04
LT: change rotations + hues	$\mathbf{1.00} \pm 0.00$	$\mathbf{0.57} \pm 0.34$	$\mathbf{0.91} \pm 0.00$	0.30 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.28 ± 0.00

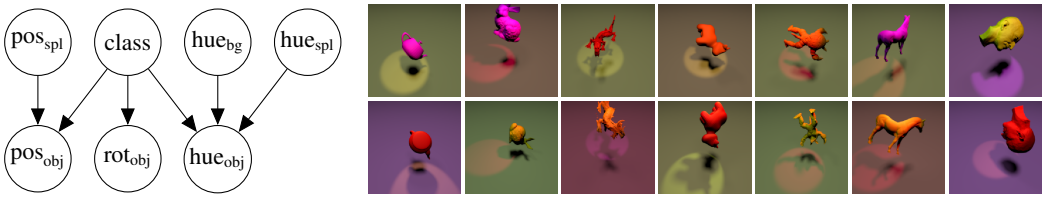
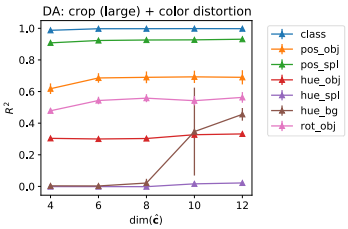


Figure 2: *(Left)* Causal graph for the *Causal3DIdent* dataset. *(Right)* Two samples from each object class.

5.2 High-dimensional images: *Causal3DIdent*

***Causal3DIdent* dataset.** *3DIdent* [119] is a benchmark for evaluating identifiability with rendered 224×224 images which contains hallmarks of natural environments (e.g. shadows, different lighting conditions, a 3D object). For influence of the latent factors on the renderings, see Fig. 2 of [119]. In *3DIdent*, there is a single object class (Teapot [83]), and all 10 latents are sampled independently. For *Causal3DIdent*, we introduce **six** additional classes: Hare [111], Dragon [104], Cow [58], Armadillo [65], Horse [92], and Head [105]; and impose a causal graph over the latent variables, see Fig. 2. While object class and all environment variables (spotlight position & hue, background hue) are sampled independently, all object latents are dependent,[12] see Appendix B for details.[13]

Experimental setup. For $\mathbf{g}$, we train a convolutional encoder composed of a ResNet18 [43] and an additional fully-connected layer, with LeakyReLU activation. As in `SimCLR` [18], we use InfoNCE with cosine similarity, and train on pairs of augmented examples $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$. As n_c is unknown and variable depending on the augmentation, we fix $\dim(\hat{\mathbf{c}}) = 8$ throughout. Note that we find the results to be, for the most part, robust to the choice of $\dim(\hat{\mathbf{c}})$, see inset figure. We consider the following data augmentations (DA): crop, resize & flip; colour distortion (jitter & drop); and rotation $\in \{90°, 180°, 270°\}$. For comparison, we also consider directly imposing a content-style partition by performing a latent transformation (LT) to generate views. For evaluation, we use linear logistic regression to predict object class, and kernel ridge to predict the other latents from $\hat{\mathbf{c}}$.[14]



Results. The results are presented in Tab. 1. Overall, our main findings can be summarised as:

(i) it can be difficult to design image-level augmentations that leave *specific* latent factors invariant;
(ii) augmentations & latent transformations generally have a similar effect on groups of latents;
(iii) augmentations that yield good classification performance induce variation in all other latents.

We observe that, similar to directly varying the hue latents, colour distortion leads to a discarding of hue information as style, and a preservation of (object) position as content. Crops, similar to varying the position latents, lead to a discarding of position as style, and a preservation of background and object hue as content, the latter assuming crops are sufficiently large. In contrast, image-level rotation affects both the object rotation and position, and thus deviates from only varying the rotation latents.

[12] e.g., our causal graph entails hares blend into the environment (object hue centered about background & spotlight hue), a form of active camouflage observed in Alaskan [72], Arctic [2], & Snowshoe hares.

[13] We made the Causal3DIdent dataset publicly available at this URL.

[14] See Appendix C for results with linear regression, as well as evaluation using a higher-dimensional intermediate layer by considering a projection head [18].

Whereas class is always preserved as content when generating views with latent transformations, when using data augmentations, we can only reliably decode class when crops & colour distortion are used in conjunction—a result which mirrors evaluation on ImageNet [18]. As can be seen by our evaluation of crops & colour distortion in isolation, while colour distortion leads to a discarding of hues as style, crops lead to a discarding of position & rotation as style. Thus, when used in conjunction, class is isolated as the sole content variable. See Appendix C for additional analysis.

6 Discussion

Theory vs practice. We have made an effort to tailor our problem formulation (§ 3) to the setting of data augmentation with content-preserving transformations. However, some of our more technical assumptions, which are necessary to prove block-identifiability of the invariant content partition, may not hold exactly in practice. This is apparent from our second experiment (§ 5.2), where we observe that—while class should, in principle, always be invariant across views (i.e., content)—when using *only* crops, colour distortion, or rotation, $\mathbf{g}$ appears to encode *shortcuts* [35, 90].[15] Data augmentation, unlike latent transformations, generates views $\tilde{\mathbf{x}}$ which are not restricted to the 11-dim. image manifold $\mathcal{X}$ corresponding to the generative process of *Causal3DIdent*, but may introduce additional variation: e.g., colour distortion leads to a rich combination of colours, typically a 3-dim. feature, whereas *Causal3DIdent* only contains one degree of freedom (hue). With additional factors, any introduced invariances may be encoded as content in place of class. Image-level augmentations also tend to change multiple latent factors in a correlated way, which may violate assumption *(iv)* of our theorems, i.e., that $p_{\tilde{\mathbf{s}}_A|\mathbf{s}_A}(\tilde{\mathbf{s}}_A|\mathbf{s}_A) > 0$ a.e. Moreover, our theory holds asymptotically and at the global optimum, whereas in practice we solve a non-convex optimisation problem with a finite sample and need to approximate the entropy term in (5), e.g., using a finite number of negative pairs. The resulting challenges for optimisation may be further accentuated by the higher dimensionality of $\mathcal{X}$ induced by image-level augmentations. Finally, we remark that while, for simplicity, we have presented our theory for pairs $(\mathbf{x}, \tilde{\mathbf{x}})$ of original and augmented examples, in practice, using pairs $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ of two augmented views typically yields better performance. All of our assumptions (content invariance, changing style, etc) and theoretical results still apply to the latter case. Using two augmented views helps because it leads to *increased variability* across the pair: for if $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}'$ differ from $\mathbf{x}$ in style subsets A and A', respectively, then $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ differ from each other (a.s.) in $(A \cup A')$.

Beyond entropy regularisation. We have shown a clear link between an identifiable maximum entropy approach to SSL (Thm. 4.4) and `SimCLR` [18] based on the analysis of [114], and have discussed an intuitive connection to the notion of redundancy reduction used in `BarlowTwins` [118]. Whether other types of regularisation such as the architectural approach pursued in `BYOL` [38] and `SimSiam` [19] can also be linked to entropy maximisation, remains an open question. Deriving similar results to Thm. 4.4 with other regularisers is a promising direction for future research.

The choice of augmentation technique implicitly defines content and style. As we have defined content as the part of the representation which is always left invariant across views, the choice of augmentation implicitly determines the content-style partition. This is particularly important to keep in mind when applying SSL with data augmentation to safety-critical domains, such as medical imaging. We also advise caution when using data augmentation to identify specific latent properties, since, as observed in § 5.2, image-level transformations may affect the underlying ground-truth factors in unanticipated ways. Also note that, *for a given downstream task*, we may not want to discard all style information since style variables may still be correlated with the task of interest and may thus help improve predictive performance. *For arbitrary downstream tasks*, however, where style may change in an adversarial way, it can be shown that only using content is optimal [97].

Conclusion. Existing representation learning approaches typically assume mutually independent latents, though dependencies clearly exist in nature [100]. We demonstrate that in a non-*i.i.d.* scenario, i.e., constructing multiple views of the same example with data augmentation, we can learn useful representations in the presence of this neglected phenomenon. More specifically, the present work contributes, to the best of our knowledge, the first: (i) identifiability result under arbitrary dependence between latents; and (ii) empirical study that evaluates the effect of data augmentations not only on classification, but also on different *continuous* ground-truth latents. Unlike existing identifiability results which rely on *change as a learning signal*, our approach aims to identify what is always shared across views, i.e., using *invariance as a learning signal*. We hope that this change in perspective will provide a fruitful basis for future applications, such as optimal style transfer or disentangling shape from pose in vision, as well as allowing further theoretical insights.

[15] class is distinguished by shape, a feature commonly unused in downstream tasks on natural images [34]

References

[1] Pulkit Agrawal, Joao Carreira, and Jitendra Malik. Learning to see by moving. In *Proceedings of the IEEE international conference on computer vision*, pages 37–45, 2015.

[2] Arctic Wildlife. Churchill Polar Bears . https://churchillpolarbears.org/churchill/, 2021.

[3] Philip Bachman, R. Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems 32*, pages 15509–15519, 2019.

[4] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33*, 2020.

[6] Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

[7] Jens Behrmann, Will Grathwohl, Ricky TQ Chen, David Duvenaud, and Jörn-Henrik Jacobsen. Invertible residual networks. In *International Conference on Machine Learning*, pages 573–582. PMLR, 2019.

[8] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

[9] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[10] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. *Advances in neural information processing systems*, 6: 737–744, 1993.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

[12] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

[13] J-F Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal processing letters*, 4(4):112–114, 1997.

[14] Michael A Casey and Alex Westner. Separation of mixed audio sources by independent subspace analysis. In *ICMC*, pages 154–161, 2000.

[15] Olivier Chapelle and Bernhard Schölkopf. Incorporating invariances in nonlinear SVMs. In *Advances in Neural Information Processing Systems 14*, pages 609–616, Cambridge, MA, USA, 2002. MIT Press.

[16] Ricky TQ Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in vaes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2615–2625, 2018.

[17] Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

[18] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[19] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. *arXiv preprint arXiv:2011.10566*, 2020.

[20] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546, 2005.

[21] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2461–2505, 2011.

[22] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[23] Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[24] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[25] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020.

[26] Tri Dao, Albert Gu, Alexander Ratner, Virginia Smith, Chris De Sa, and Christopher Ré. A kernel theory of modern data augmentation. In *International Conference on Machine Learning*, pages 1528–1537. PMLR, 2019.

[27] G Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231, 1951.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[29] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[30] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[31] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

[32] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

[33] Frederick Eberhardt and Richard Scheines. Interventions and causal inference. *Philosophy of science*, 74 (5):981–995, 2007.

[34] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations (ICLR)*, 2019.

[35] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, Nov 2020.

[36] Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The Incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI*, 2019.

[37] Luigi Gresele, Giancarlo Fissore, Adrián Javaloy, Bernhard Schölkopf, and Aapo Hyvarinen. Relative gradient optimization of the Jacobian term in unsupervised deep learning. *Advances in Neural Information Processing Systems*, 33, 2020.

[38] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33*, 2020.

[39] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304, 2010.

[40] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 13:307–361, 2012.

[41] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006.

[42] Hermanni Hälvä and Aapo Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition, CVPR*, 2016.

[44] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 9726–9735. IEEE, 2020.

[45] Olivier J. Hénaff. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*,

volume 119 of *Proceedings of Machine Learning Research*, pages 4182–4192. PMLR, 2020.

[46] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations*, 2017.

[47] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations*, 2019.

[48] Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *IJCAI*, pages 2506–2513, 2019.

[49] Andrew G. Howard. Some improvements on deep convolutional neural network based image classification, 2013.

[50] Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720, 2000.

[51] Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, pages 3765–3773, 2016.

[52] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

[53] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[54] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

[55] Aapo Hyvarinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[56] Jörn-Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. i-RevNet: Deep invertible networks. In *International Conference on Learning Representations*, 2018.

[57] Edwin T Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9): 939–952, 1982.

[58] Keenan's 3D Model Repository. Keenan's 3D Model Repository . https://www.cs.cmu.edu/ kmcrane/Projects/ModelRepository/, 2021.

[59] Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 108, pages 2207–2217, 2020.

[60] Ilyes Khemakhem, Ricardo Pio Monti, Diederik P. Kingma, and Aapo Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. In *Advances in Neural Information Processing Systems 33*, 2020.

[61] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[62] Diederik P Kingma and Prafulla Dhariwal. Glow: generative flow with invertible $1\times$ 1 convolutions. In *Advances in Neural Information Processing Systems*, pages 10236–10245, 2018.

[63] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014.

[64] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *International Conference on Learning Representations (ICLR)*, 2021.

[65] Venkat Krishnamurthy and Marc Levoy. Fitting smooth surfaces to dense polygon meshes. In John Fujii, editor, *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1996, New Orleans, LA, USA, August 4-9, 1996*, pages 313–324. ACM, 1996.

[66] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

[67] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[68] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR 2011*, pages 3361–3368. IEEE, 2011.

[69] Te-Won Lee, Mark Girolami, and Terrence J Sejnowski. Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation*, 11 (2):417–441, 1999.

[70] Felix Leeb, Yashas Annadani, Stefan Bauer, and Bernhard Schölkopf. Structural autoencoders improve representations for generation and transfer. *arXiv preprint arXiv:2006.07796*, 2020.

[71] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[72] Lepus Othus. Animal Diversity Web . https://animaldiversity.org/accounts/Lepus_othus/, 2021.

[73] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

[74] Ralph Linsker. An application of the principle of maximum information preservation to linear systems. In *Advances in neural information processing systems*, pages 186–194, 1989.

[75] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[76] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124, 2019.

[77] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6348–6359. PMLR, 2020.

[78] Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations*, 2018.

[79] Chaochao Lu, Yuhuai Wu, Jośe Miguel Hernández-Lobato, and Bernhard Schölkopf. Nonlinear invariant risk minimization: A causal approach. *arXiv preprint arXiv:2102.12353*, 2021.

[80] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pages 3111–3119, 2013.

[81] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020.

[82] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT Press, Cambridge, MA, 2012.

[83] Martin Edward Newell. *The Utilization of Procedure Models in Digital Image Synthesis.* PhD thesis, The University of Utah, 1975. AAI7529894.

[84] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016.

[85] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[86] George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, pages 2338–2347, 2017.

[87] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[88] Judea Pearl. *Causality*. Cambridge university press, 2009.

[89] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[90] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks, 2020.

[91] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

[92] Emil Praun, Adam Finkelstein, and Hugues Hoppe. Lapped textures. In Judith R. Brown and Kurt Akeley, editors, *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000*, pages 465–470. ACM, 2000.

[93] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *Technical report, OpenAI*, 2018.

[94] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio. Multi-task self-supervised learning for robust speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993, 2020.

[95] H. Richard, L. Gresele, A. Hyvarinen, B. Thirion, A. Gramfort, and P. Ablin. Modeling shared responses in neuroimaging studies through multiview ica. In *Advances in Neural Information Processing Systems 33*, pages 19149–19162. Curran Associates, Inc., December 2020.

[96] Geoffrey Roeder, Luke Metz, and Diederik P Kingma. On linear identifiability of learned representations. *arXiv preprint arXiv:2007.00810*, 2020.

[97] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research*, 19(1):1309–1342, 2018.

[98] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *CoRR*, abs/1904.05862, 2019.

[99] Bernhard Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

[100] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021.

[101] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *arXiv preprint arXiv:2010.02637*, 2020.

[102] Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

[103] Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). In *International Conference on Learning Representations*, 2020.

[104] Stanford Scanning Repository. The Stanford 3D Scanning Repository. http://graphics.stanford.edu/data/3Dscanrep/, 2021.

[105] Suggestive Contour Gallery. Suggestive Contour Gallery . https://gfx.cs.princeton.edu/proj/sugcon/models/, 2021.

[106] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pages 6056–6065. PMLR, 2019.

[107] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[108] Fabian Theis. Towards a general independent subspace analysis. *Advances in Neural Information Processing Systems*, 19:1361–1368, 2006.

[109] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[110] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[111] Greg Turk and Marc Levoy. Zippered polygon meshes from range images. In Dino Schweitzer, Andrew S. Glassner, and Mike Keeler, editors, *Proceedings of the 21th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1994, Orlando, FL, USA, July 24-29, 1994*, pages 311–318. ACM, 1994.

[112] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pages 1096–1103, 2008.

[113] Julius von Kügelgen, Ivan Ustyuzhaninov, Peter Gehler, Matthias Bethge, and Bernhard Schölkopf. Towards causal generative scene models via competition of experts. *arXiv preprint arXiv:2004.12906*, 2020.

[114] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 2020.

[115] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

[116] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3733–3742. IEEE Computer Society, 2018.

[117] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*, 2020.

[118] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.

[119] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. *arXiv preprint arXiv:2102.08850*, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the "Theory vs practice" paragraph, as well as lines 409–411 in § 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] While our work is theoretical in nature, we discuss some potential negative impacts in lines 405–409 where we caution against a naive use of data-augmentation in safety-critical domains without extensive prior study.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Our problem formulation is formally described in detail in § 3. All our theoretical results are accompanied by a clear and complete account of all necessary assumptions.
   (b) Did you include complete proofs of all theoretical results? [Yes] Due to space constraints, and since our proofs introduce novel techniques and are nontrivial, we only sketch the high-level ideas behind our proofs in the main paper. The full proofs with detailed explanations are included in Appendix A.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The *Causal3DIdent* dataset is included as a public URL, see the hyperlink in footnote 13. We include all code to reproduce our experiments in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Provided in Appendix D
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All results presented in § 5 report the mean and standard deviation over 3 random seeds as also stated in the main paper.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Provided in Appendix D
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The 3DIdent dataset, as well as the 3D models used to extend it are cited in § 5.2.
   (b) Did you mention the license of the assets? [Yes] Provided in Appendix B
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] See question 3 (a): the newly-introduced *Causal3DIdent* dataset has been publicly released.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We did not use/curate any data from other people, see Appendix B for discussion of this point.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] The data does not contain any personally identifiable information or offensive content, see Appendix B for discussion of this point.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]