Self-Supervised Disentanglement by Leveraging Structure in Data Augmentations

Cian Eastwood[1,2,3] **Julius von Kügelgen**[1,4] **Linus Ericsson**[2]
Diane Bouchacourt[3] **Pascal Vincent**[3] **Mark Ibrahim**[3] **Bernhard Schölkopf**[1]

[1] Max Planck Institute for Intelligent Systems, Tübingen
[2] University of Edinburgh [3] Meta AI [4] University of Cambridge

Abstract

Self-supervised representation learning often uses data augmentations to induce some invariance to "style" attributes of the data. However, with downstream tasks generally unknown at training time, it is difficult to deduce *a priori* which attributes of the data are indeed "style" and can be safely discarded. To address this, we introduce a more principled approach that seeks to *disentangle* style features rather than discard them. The key idea is to add multiple style embedding spaces where: (i) each is invariant to *all-but-one* augmentation; and (ii) *joint* entropy is maximized. We formalize our structured data-augmentation procedure from a causal latent-variable-model perspective, and prove identifiability of both content *and* (multiple blocks of) style variables. We empirically demonstrate the benefits our approach on synthetic datasets and then present promising but limited results on ImageNet.

1 Introduction

Learning useful representations from unlabelled data is widely recognized as an important step towards more capable machine-learning systems (Bengio et al., 2013). In recent years, *self-supervised learning* (SSL) has made significant progress towards this goal, approaching the performance of supervised methods on many downstream tasks (Ericsson et al., 2021). The main idea is to leverage known data structures to construct proxy tasks or objectives that act as a form of (self-)supervision. This could involve predicting one part of an observation from another (Brown et al., 2020), or, as we focus on in this work, leveraging **data augmentations/transformations** to perturb different attributes of the data.

Most current approaches are based on the joint-embedding framework and use data augmentations as weak supervision to determine what information to retain (termed "content") and what information to discard (termed "style") (Bromley et al., 1994; Chen et al., 2020a; Zbontar et al., 2021; Bardes et al., 2022). In particular, they do so by optimizing for representation similarity or **invariance** across transformations of the same observation, subject to some form of **entropy** regularization, with this invariance-entropy trade-off tuned for some particular task (e.g., ImageNet object classification).

However, at pre-training time, it is unclear what information should be discarded as **one task's style may be another's content**. Ericsson et al. (2021) illustrated this point, finding ImageNet object-classification accuracy (the task optimized for in pre-training) to be poorly correlated with downstream object-detection and dense-prediction tasks, concluding that *"universal pre-training is still unsolved"*.

Example 1.1 (Color and Rotation)**.** Suppose we want to make use of color and rotation transformations. While some invariance to (or discarding of) an image's color and orientation features can be *beneficial* for ImageNet object classification (Chen et al., 2020a), it can also be *detrimental* for other tasks like segmentation or fine-grained species classification (Cole et al., 2022).

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

Figure 1: Framework overview. Given M atomic transformations like color distortion or rotation (here $M = 2$), we learn a "content" embedding space (Z_0) that is invariant to all transformations and M "style" embedding spaces (Z_1, Z_2) that are each invariant to all-but-the-m^{th} atomic transformation. To do so, we construct $M+1$ transformation pairs (t^m, t'^m) sharing different parameters and use these to create $M+1$ transformed image pairs $(\tilde{x}^m, \tilde{x}'^m)$ sharing different features. After routing each pair to a different space, we: (i) enforce invariance within each space; and (ii) maximize entropy across the joint spaces. The result is $M+1$ disentangled embedding spaces.

To address this and learn more universal representations, we introduce a new SSL framework which uses data augmentations to disentangle style attributes of the data rather than discard them. As illustrated in Fig. 1, we leverage M transformations to learn $M+1$ disentangled embedding spaces capturing both content and style information—with one style space per (group of) transformation(s).

2 Background: Using unstructured data augmentations to discard

Joint-embedding methods are often categorized as contrastive or non-contrastive; while both employ some invariance criterion L^{inv} to encourage the same embedding across different views of the same image (e.g., cosine similarity or mean squared error), they differ in how they regularize this invariance criterion to avoid collapsed or trivial solutions. In particular, contrastive methods (Chen et al., 2020a,b, 2021; He et al., 2020) do so by pushing apart the embeddings of different images, while non-contrastive methods do so by architectural design (Grill et al., 2020; Chen and He, 2021) or by regularizing the covariance of embeddings (Zbontar et al., 2021; Bardes et al., 2022; Ermolov et al., 2021). We focus on contrastive and covariance-based non-contrastive methods which can both be expressed as a combination of invariance L^{inv} and entropy L^{ent} terms (Garrido et al., 2023),

$$L^{SSL} = L^{inv} + L^{ent}. \tag{2.1}$$

Note these terms have also been called alignment and uniformity (Wang and Isola, 2020), respectively. For concreteness, Table 3 of App. B.2 specifies L^{inv} and L^{ent} for some common SSL methods.

In general, the joint-embedding framework involves an unlabelled dataset of observations or images x and M transformation distributions $T_1, \ldots, T_M$ from which to sample M atomic transformations $t_1, \ldots, t_M$, with $t_m \sim T_m$, composed together to form $t = t_1 \circ \cdots \circ t_M$. Critically, each atomic transformation t_m is designed to perturb a different "style" attribute of the data deemed nuisance for the task at hand. Returning to Example 1.1, this could mean sampling parameters for a color distortion $t_c \sim T_c$ and rotation $t_r \sim T_r$, and then composing them as $t = t_c \circ t_r$. For brevity, this sample-and-compose operation is often written as $t \sim T$.

For each image x, a pair of transformations $t, t' \overset{iid}{\sim} p_t$ is sampled and applied to form a pair of views $(\tilde{x}, \tilde{x}') = (t(x), t'(x))$. The views are then passed through a shared backbone network f to form a pair of representations (h, h'), with $h = f(\tilde{x})$, and then through a smaller projector g to form a pair of embeddings (z, z'), with $z = g(h) = g(f(\tilde{x})) \in Z$. Critically, the single embedding space Z seeks invariance to all transformations, thereby discarding each of the "style" attributes

3 Our Framework: Using structured data augmentations to disentangle

We now describe our framework for using data augmentations to disentangle style attributes of the data, rather than discard them—see Fig. 1 for an illustration. Given M transformations, we learn $M+1$ embedding spaces $\{Z_m\}_{m=0}^M$ capturing both content (Z_0) and style ($\{Z_m\}_{m=1}^M$) information—with one style space per (group of) atomic transformation(s).

Views. We start by constructing $M+1$ transformation pairs $\{(t^m, t'^m)\}_{m=0}^M$ which share different transformation parameters. For $m=0$, we independently sample two transformations $t'^0 \sim T$, which will generally not share any transformation parameters (i.e., $t^0_k \neq t'^0_k$ a.s. $\forall k$). For $1 \leq m \leq M$, we also independently sample two transformations $t^m, t'^m \sim T$, but then enforce that all the parameters of the m^{th} transformation are shared by setting $t'^m_m := t^m_m$. Finally, we apply each of these transformation pairs to a different image to form a pair of views $(\tilde{x}^m, \tilde{x}'^m) = (t^m(x^m), t'^m(x^m))$.

> Example 1.1 (continued). Suppose we can sample parameters for two transformations: color distortion $t_c \sim T_c$ and rotation $t_r \sim T_r$. As depicted in Fig. 1, we can then construct three transformation pairs sharing different parameters: $(t^0, t'^0) = ((t^0_c, t^0_r), (t'^0_c, t'^0_r))$ with no shared parameters; $(t^1, t'^1) = ((t^1_c, t^1_r), (t^1_c, t'^1_r))$ with shared color parameters; and $(t^2, t'^2) = ((t^2_c, t^2_r), (t'^2_c, t^2_r))$ with shared rotation parameters. Applying each transformation pair to a different image, we get three pairs of views $(\tilde{x}^0_{cr}, \tilde{x}'^0_{cr})$ for which only "content" information is shared as both color and rotation differ; $(\tilde{x}^1_{cr}, \tilde{x}'^1_{c})$ for which "content" and color information is shared, but rotation differs; and $(\tilde{x}^2_{cr}, \tilde{x}'^2_{r})$ for which "content" and rotation information is shared, but color differs.

Embedding spaces. As depicted in Fig. 1, the pairs of views $(\tilde{x}^m, \tilde{x}'^m)$ are passed through a shared backbone network f to form pairs of representations (h^m, h'^m) and subsequently through separate projectors g_l to form pairs of embeddings (z^m_l, z'^m_l), with

$$z^m_l = g_l(h^m) = g_l \circ f(\tilde{x}^m) = g_l \circ f \circ t^m(x^m) \in Z_l \tag{3.1}$$

the embedding of view $\tilde{x}^m$ in embedding space Z_l. We call Z_0 "content" space as it seeks invariance to all transformations, thereby discarding all style attributes and leaving only content. We call the other M spaces $\{Z_m\}_{m=1}^M$ "style" spaces as they seek invariance to all-but-one transformation t_m, thereby discarding all-but-one style attribute (that which is perturbed by t_m).

Loss. Given $M+1$ pairs of views $\{(\tilde{x}^m, \tilde{x}'^m)\}_{m=0}^M$ sharing different transformation parameters, we learn $M+1$ disentangled embedding spaces by minimizing the following objective:

$$L^{\text{ours}}(f, \{g_m\}_{m=0}^M; \{(\tilde{x}^m, \tilde{x}'^m)\}_{m=0}^M) = \underbrace{L^{\text{inv}}_{Z_0} + L^{\text{ent}}_{Z_0}}_{\text{standard loss (content } Z_0)} + \underbrace{\sum_{m=1}^M L^{\text{inv}}_{Z_m} + L^{\text{ent}}_{Z}}_{\text{additional terms (style } Z_m\text{'s)}}, \tag{3.2}$$

$$= \underbrace{\sum_{m=0}^M L^{\text{inv}}_{Z_m}}_{M+1 \text{ inv. terms}} + \underbrace{L^{\text{ent}}_{Z}}_{\text{joint entropy}} + \underbrace{L^{\text{ent}}_{Z_0}}_{\text{content entropy}}, \tag{3.3}$$

where the individual invariance ($L^{\text{inv}}_{Z_m}$) and (content / joint) entropy ($L^{\text{ent}}_{Z_0}$ / L^{ent}_{Z}) terms are given by

$$L^{\text{inv}}_{Z_m} = l_m \circ L^{\text{inv}}(z^m_m, z'^m_m), \quad L^{\text{ent}}_{Z_0} = L^{\text{ent}}(\{z^m_0, z'^m_0\}_{m=0}^M), \quad L^{\text{ent}}_{Z} = L^{\text{ent}}(\{z^m, z'^m\}_{m=0}^M),$$

with $z^m = [z^m_0, \ldots, z^m_M] \in Z_0 \times \cdots \times Z_M$ the concatenated embeddings of $\tilde{x}^m$ across all spaces. Eq. (3.2) highlights the additional terms we add to the standard contrastive loss. In particular, note that we require two different entropy terms to ensure disentangled embedding spaces. Since "content" is invariant to all transformations (by definition), we require L^{ent}_{Z} to prevent redundancy (of additional copies of content, one per style space) and $L^{\text{ent}}_{Z_0}$ to ensure content is indeed encoded in Z_0 (otherwise it could be spread across all $M+1$ spaces). As detailed in App. C.2, this is a key difference compared to Xiao et al. (2021), who learn multiple embedding spaces but do not achieve disentanglement.

4 Causal Representation Learning Perspective and Identifiability Analysis

In this section, we investigate what is actually learned by the structured use of data augmentations in § 3, through the lens of causal representation learning (Schölkopf et al., 2021). To this end, we first formalize the data generation and augmentation processes as a (causal) latent variable model, and then study the identifiability of different components of the latent representation. Our analysis strongly builds on and extends the work of von Kügelgen et al. (2021) by showing that the structure inherent to different augmentation transformations can be leveraged to identify not only the block of shared content variables, but also individual style components (subject to suitable assumptions).

Data-generation and augmentation processes. We assume that the observations $x \in \mathcal{X}$ result from underlying latent vectors $z \in \mathcal{Z}$ via an invertible nonlinear mixing function $f : \mathcal{Z} \to \mathcal{X}$,

$$z \sim p_z, \qquad x = f(z). \tag{4.1}$$

Here, $\mathcal{Z} = \mathbb{R}^d$ is a latent space capturing object properties such color or rotation, p_z is a distribution over latents; and $\mathcal{X}$ denotes the d-dimensional data manifold, which is typically embedded in a higher dimensional pixel space. In the same spirit, we model the way in which augmented views $(\tilde{x}, \tilde{x}')$ are generated from x through perturbations in the latent space:

$$\tilde{z}, \tilde{z}' \sim p_{\tilde{z}|z}, \qquad \tilde{x} = f(\tilde{z}), \qquad \tilde{x}' = f(\tilde{z}'). \tag{4.2}$$

The conditional $p_{\tilde{z}|z}$, from which the pair of augmented latents $(\tilde{z}, \tilde{z}')$ is drawn given the original latent z, constitutes the latent-space analogue of the image-level transformation (t, t') in § 3. More specifically, $\tilde{z} \sim p_{\tilde{z}|z}$ captures the behavior of $t \circ f^{-1}$ with $t \sim \mathcal{T}$ acting on $x = f(z)$.

Content-style partition. Typically, augmentations are designed to affect some semantic aspects of the data (e.g., color and rotation) and not others (e.g., object identity). We therefore partition the latents into *style* latents s, which are affected by the augmentations, and *shared content* latents c, which are not affected by the augmentations. Further, $p_{\tilde{z}|z}$ in (4.2) takes the form

$$p_{\tilde{z}|z}(\tilde{z} \mid z) = \delta(\tilde{c} - c)\, p_{\tilde{s}|s}(\tilde{s} \mid s) \tag{4.3}$$

for some style conditional $p_{\tilde{s}|s}$, such that z, $\tilde{z}$, and $\tilde{z}'$ in (4.2) are given by

$$z = (c, s), \qquad \tilde{z} = (c, \tilde{s}), \qquad \tilde{z}' = (c, \tilde{s}'). \tag{4.4}$$

For this setting, it has been shown that—under suitable additional assumptions—contrastive SSL recovers the shared content latents c up to an invertible function (von Kügelgen et al., 2021, Thm. 4.4).

Beyond content identifiability: separating and recovering individual style latents. Previous analyses of SSL with data augmentations considered style latents s as nuisance variables that should be discarded, thus seeking a pure content-based representation that is invariant to all augmentations (von Kügelgen et al., 2021; Daunhawer et al., 2023). The focus of our study, and its key difference to these previous analyses, is that we seek to also identify and disentangle different style variables s by leveraging available structure in data augmentations that has not been exploited thus far.

First, note that each class of atomic transformations $\mathcal{T}_m$ (e.g., color distortion or rotation) typically affects a different property, meaning that it should only affect a subset of style variables. Hence, we partition the style block into more fine-grained individual style components s_m,

$$s = (s_1, \ldots, s_M), \qquad \tilde{s} = (\tilde{s}_1, \ldots, \tilde{s}_M), \qquad \tilde{s}' = (\tilde{s}'_1, \ldots, \tilde{s}'_M), \tag{4.5}$$

and assume that the style conditional $p_{\tilde{s}|s}$ in (4.3) factorizes as follows:

$$p_{\tilde{s}|s}(\tilde{s} \mid s) = \prod_{m=1}^{M} p_{\tilde{s}_m|s_m}(\tilde{s}_m \mid s_m), \tag{4.6}$$

where each term $p_{\tilde{s}_m|s_m}$ on the RHS is the latent-space analogue of $\mathcal{T}_m$.

Next, we wish for our latent variable model to capture the structured use of data augmentation through transformation pairs with shared parameters, as described in § 3. Specifically, note that—unlike most

prior approaches to SSL with data augmentations—we do not create a single dataset of ("positive") pairs $(\tilde{x}, \tilde{x}')$. Instead, we construct transformation pairs (t^m, t'^m) in $M+1$ different ways, giving rise to $M+1$ datasets of pairs $(\tilde{x}^m, \tilde{x}'^m)$, each differing in the shared (style) properties. In particular, the m^{th} atomic transformation is shared across (t^m, t'^m) by construction, such that $(\tilde{x}^m, \tilde{x}'^m)$ should share the same perturbed m^{th} style component $\tilde{s}_m = \tilde{s}'_m$—regardless of its original value s_m. To model this procedure, we define $M+1$ different ways of jointly perturbing the style variables as follows:

$$p^{(m)}(\tilde{s}, \tilde{s}' \mid s) = \prod_{l=1}^{M} p^{(m)}(\tilde{s}_l, \tilde{s}'_l \mid s_l) \quad \text{for} \quad m = 0, \ldots, M, \tag{4.7}$$

$$\text{where} \quad p^{(m)}(\tilde{s}_l, \tilde{s}'_l \mid s_l) = \begin{cases} p_{\tilde{s}_l \mid s_l}(\tilde{s}_l \mid s_l)\, \delta(\tilde{s}'_l - \tilde{s}_l) & \text{if } l = m \\ p_{\tilde{s}_l \mid s_l}(\tilde{s}_l \mid s_l)\, p_{\tilde{s}_l \mid s_l}(\tilde{s}'_l \mid s_l) & \text{otherwise} \end{cases}$$

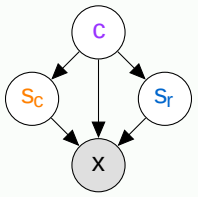
Together with $z = (c, s) \sim p_z$ as in (4.1), the conditionals in (4.7) induce $M+1$ different joint distributions $p^{(m)}_{\tilde{x},\tilde{x}'}$ over observation pairs $(\tilde{x}^m, \tilde{x}'^m)$: analogous to (4.2), we have for $m = 0, \ldots, M$,

$$\tilde{s}^m, \tilde{s}'^m \sim p^{(m)}_{\tilde{s},\tilde{s}'|s}, \qquad \tilde{x}^m = f([c, \tilde{s}^m]), \qquad \tilde{x}'^m = f\left[c, \tilde{s}'^m\right]. \tag{4.8}$$

Remark 4.1. In practice, we do not generate $M+1$ augmented pairs for each $x = f(z)$ as described above. Instead, each pair is constructed from a different observation $x^l = f(z^l)$ transformed according to $m := l \mod M+1$. In the limit of in nite data, these two options have the same effect.

Example 1.1 (continued). Denote the style component capturing color by s_c and that capturing rotation by s_r. For $m = 0, 1, 2$, let $z^m = (c^m, s^m_c, s^m_r)$ be the latents underlying separate images x^m. Then the augmentations shown in Fig. 1 (left) are captured by the following changes to the latents:

m	z^m	$\tilde{z}^m$	$\tilde{z}'^m$	Shared Latents
0	(c^0, s^0_c, s^0_r)	$(c^0, \tilde{s}^0_c, \tilde{s}^0_r)$	$(c^0, \tilde{s}'^0_c, \tilde{s}'^0_r)$	only content
1	(c^1, s^1_c, s^1_r)	$(c^1, \tilde{s}^1_c, \tilde{s}^1_r)$	$(c^1, \tilde{s}'^1_c, \tilde{s}'^1_r)$	content & color
2	(c^2, s^2_c, s^2_r)	$(c^2, \tilde{s}^2_c, \tilde{s}^2_r)$	$(c^2, \tilde{s}'^2_c, \tilde{s}'^2_r)$	content & rotation



Causal interpretation. The described augmentation procedure can also be interpreted in causal terms (Ilse et al., 2021; Mitrovic et al., 2021; von Kügelgen et al., 2021). Given a factual observation x, the augmented views $(\tilde{x}^m, \tilde{x}'^m)$ constitute pairs of counterfactuals under joint interventions on all style variables, provided that (i) c is a root note in the causal graph, to ensure content invariance in (4.3); and (ii) the style components s_m do not causally in uence each other, to justify the factorization in (4.6) and (4.7).[1] A causal graph compatible with these constraints is shown for Example 1.1 above on the right. As a structural causal model (SCM; Pearl, 2009), this can be written as

$$c := u_c, \qquad s_m := f_m(c, u_m), \qquad \text{for} \qquad m = 1, \ldots, M, \tag{4.9}$$

with jointly independent exogenous variables $\{u_m\}_{m=0}^{M}$. The style conditionals $p_{\tilde{s}_m|\tilde{s}_m}$ in (4.6) can then arise, e.g., from shift $do(s_m = f_m(c, u_m) + \tilde{u}_m)$ or perfect $do(s_m = \tilde{u}_m)$ interventions with independent augmentation noise $\tilde{u}_m$. Note that the latter renders $\tilde{s}_m$ independent of all other variables.

Style identi ability and disentanglement. By construction, $\{c, \tilde{s}_m\}$ is shared across $(\tilde{x}^m, \tilde{x}'^m)$ and can thus be identi ed up to nonlinear mixing by contrastive SSL on the m^{th} dataset (von Kügelgen et al., 2021, Thm. 4.4). However, it remains unclear how to disentangle the two and recover $\tilde{s}_m$ only i.e., how to "remove" c, which can separately be recovered as the only shared latent for $m = 0$. The following result, proven in App. A, shows that our approach from § 3 with $M+1$ alignment terms and joint entropy regularization indeed disentangles and recovers the individual style components.

[1] The allowed structure is similar to Suter et al. (2019, Fig. 1); Wang and Jordan (2021, Fig. 9). However, ours is more general as content does not only confound different s_m, but also directly in uences the observed x.

Theorem 4.2 (Identifiability). For the data generating process in (4.1), (4.7), (4.8), assume that

A_1. $\mathcal{Z}$ is open and simply connected, f is diffeomorphic onto its image, p_z is smooth and fully supported on $\mathcal{Z}$; each $p_{\tilde{s}_m|s_m}$ is smooth and supported on an open, non-empty set around any s_m;

A_2. p_z and $\{p_{\tilde{s}_m|s_m}\}_{m=1}^M$ are such that $\{c\} \cup \{\tilde{s}_m\}_{m=1}^M$ are jointly independent;

A_3. the latent dims $\{d_m\}_{m=0}^M$ are known and $\{f_m : \mathcal{X} \to (0,1)^{d_m}\}_{m=0}^M$ are smooth minimizers of

$$\sum_{m=0}^{M} \mathbb{E}_{p^{(m)}_{\tilde{x},\tilde{x}'}} \left\| f_m(\tilde{x}) - f_m(\tilde{x}') \right\|_2 - H_{p^{(0)}_{\tilde{x}}}([f_0(\tilde{x}), \ldots, f_M(\tilde{x})]) \,. \tag{4.10}$$

Then f_0 block-identifies (von Kügelgen et al., 2021, Defn. 4.1) the content c, and f_m block-identifies s_m in the sense that $\hat{s}_m = f_m(x) = g_m(s_m)$ for some invertible g_m for each $m = 1, \ldots, M$.

Discussion of Thm. 4.2. The technical assumption A_1 is also needed to prove content identifiability (von Kügelgen et al., 2021). Assumption A_2, which requires that the augmentation process renders c and $\{\tilde{s}_m\}_{m=1}^M$ independent, is specific to our extended analysis. It holds, e.g., if $p_z(z)$ is such that c and $\{s_m\}_{m=1}^M$ are independent to begin with; or if (b) $p_{\tilde{s}|s} = p_{\tilde{s}}$ does not depend on s, as would be the case for perfect interventions. As discussed in more detail in App. C, (a) relates to work on multi-view latent correlation maximization (Lyu et al., 2021), nonlinear ICA (Gresele et al., 2019), and disentanglement (Locatello et al., 2020; Ahuja et al., 2022), whereas (b) relates to work in weakly supervised causal representation learning (Brehmer et al., 2022). In case (b), we could actually also allow for causal relations among individual style components s_m and $s_{m'}$, as such links are broken by perfect interventions. When A_2 does not hold (e.g., for content-dependent style and imperfect interventions—arguably the most realistic setting), block-identifiability of the style components seems infeasible, consistent with existing negative results (Brehmer et al., 2022; Squires et al., 2023). However, in this case we posit that the exogenous style variables in (4.9), which capture any style information not due to c and are jointly independent by assumption, are recovered in place of s_m.

5 Experiments

We now present our experimental results which: (i) use a numerical dataset and a synthetic-image dataset to illustrate how adapting our ℓ hyperparameter helps to fully disentangle content (see App. B.3); and (ii) use ImageNet to illustrate the downstream performance benefits of retaining more style information. App. B gives full implementation details for all experiments.

Numerical dataset: Recovering only content. Following von Kügelgen et al. (2021, Sec. 5.1), we generate synthetic data pairs $(x, \tilde{x}) = (f(c, s), f(c, \tilde{s}))$ with shared content c and perturbed style $s, \tilde{s}$ (see App. B.4 for details.). We then train a simple encoder f (3-layer MLP) with SimCLR using (i) fixed ℓ (ii) our adaptive ℓ (see App. B.3) to get learned embeddings $z = f(x)$. We then report the r^2 coefficient of determination in predicting the ground-truth c and s from z. Fig. 2 shows how varying the dimensionality of z affects the recovery of content c and style s, focusing on the scenario where we have sufficient capacity to encode (all of) content c (i.e., $\dim(z) \geq \dim(c)$). Similar to von Kügelgen et al. (2021, Fig. 10), we find that with standard SimCLR (ℓ fixed), excess capacity is used to encode some style information (since that increases entropy). However, by adapting ℓ using the procedure of App. B.3, we prevent style "leaking in", allowing us to recover only content in z.

ColorDSprites: Sensitivity to augmentation strengths. We now make use of a colored version of the DSprites dataset (Locatello et al., 2019) which contains images of 2D shapes generated from 6 independent ground-truth factors (# values): color (10), shape (3), scale (6), orientation (40), x-position (32) and y-position (32). We first train SimCLR and VICReg models on the (unlabelled) dataset using different augmentation strengths (see Fig. 3 of App. B.1). We then train linear classifiers on top of frozen embeddings z to predict the ground-truth factor values. Table 1 shows that for (i) fixed ℓ, the augmentation strengths severely affect SimCLR's invariance-entropy trade-off and, as a result, the amount (and type) of style information captured in z, and (ii) adapting ℓ (see App. B.3) makes SimCLR's invariance-entropy trade-off much more robust to the augmentation strengths, ensuring that z captures all of content ($C = 1$) and almost no style ($\bar{S} = 0$)—regardless of the augmentation strengths. Table 5 of App. D.1 gives the corresponding results for VICReg.

Table 1: SimCLR's sensitivity to augmentation strengths with fixed and adaptive l on ColorDSprites. r^2 in predicting the ground-truth factor values from the post-projector embedding z with a linear classifier. Adapting l ensures that z captures all of content ($\overline{C} = 1$) and almost no style ($\overline{S} = 0$), regardless of the augmentation strengths.

l	Augm. Strength	Content (C) Shape	Style (S) Color	Scale	Orient.	PosX	PosY	$\overline{S}$ (#)
fixed	weak	1.0	0.93	0.89	0.30	0.82	0.83	0.75
	medium	1.0	0.73	1.00	0.19	0.89	0.89	0.74
	strong	1.0	0.31	1.00	0.05	0.23	0.30	0.38
adaptive	weak	1.0	0.21	0.17	0.00	0.01	0.01	0.08
	medium	1.0	0.10	0.16	0.00	0.00	0.00	0.05
	strong	1.0	0.10	0.11	0.00	0.00	0.00	0.04

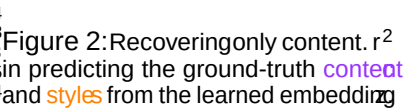


Figure 2: Recovering only content. r^2 in predicting the ground-truth content and styles from the learned embedding z

Table 2: Linear evaluation on ImageNet and a broad range of downstream tasks. We show top-1 accuracies (%) for all but CUB_{bbox} (r^2), CUB_{kpt} (r^2), and VOC (mAP_{50}). We use frozen representations h and embeddings z (post-projector). FT: our framework fine-tuned as a base SimCLR model. Ct101: Caltech101. Cf10: CIFAR10.

Alg.	Feat.	ImNt	Acft	Ct101	Cars	Cf10	Cf100	CUB_{bbox}	CUB_{cls}	CUB_{kpt}	DTD	Flwrs	Pets	SUN	VOC	Avg.
SimCLR	z	56.5	14.6	70.9	13.0	76.7	50.5	35.6	22.5	12.0	66.4	66.8	70.3	48.9	74.6	47.9
SimCLR-Ours	z	49.0	25.9	77.6	14.4	81.8	56.2	60.5	15.1	17.6	64.4	63.4	60.7	45.9	73.6	50.6
SimCLR-Ours-FT	z	57.8	15.9	72.4	14.6	77.8	53.6	36.2	22.5	12.6	67.0	67.3	70.6	49.5	74.9	48.8
SimCLR	h	68.1	50.9	88.2	50.7	89.3	73.0	71.3	48.6	32.5	75.2	93.5	82.4	60.3	79.7	68.9
SimCLR-Ours	h	61.7	46.2	86.5	37.5	87.2	67.4	70.6	29.6	23.6	72.6	86.8	73.3	55.1	77.2	62.6
SimCLR-Ours-FT	h	67.9	51.0	88.1	51.0	89.4	72.9	71.4	48.5	32.9	75.9	93.5	82.6	60.2	79.7	69.0

ImageNet: Downstream performance. We train all models for 100 epochs on a blurred-face ImageNet1k (Russakovsky et al., 2014) dataset using the standard transformations (random crop, horizontal flip, color jitter, grayscale and blur). For our method, we group these into spatial (crop, flip) and appearance (color jitter, greyscale, blur) transformations and thus learn $n = 2$ style spaces. We then follow the setup of Ericsson et al. (2021) to evaluate models on a broad range of downstream tasks, covering object/texture/scene classification, localization, and keypoint estimation. With the post-projector z, Table 2 shows that: (i) using our framework from scratch improves downstream performance at the cost of ImageNet performance; and (ii) using our framework to fine-tune a SimCLR (Chen et al., 2020a) model (i.e., add in style spaces) leads to improved performance both downstream and on ImageNet. Unfortunately, this improved performance with z did not translate into improved performance with the pre-projector h. This highlights the importance of the projector, but also our poor understanding of its role and impact on the retention of style information. Table 6 of App. D.2 gives the corresponding results for VICReg (Bardes et al., 2022).

6 Discussion

Related work. Xiao et al. (2021) also learn multiple embedding spaces in order to capture style information. In our work, we further develop these ideas towards a fully disentangled embedding space through a different use of augmentations and embedding spaces, as well as a different objective function—see App. C.2 for a detailed comparison with Xiao et al. (2021). Other prior work sought to retain some style information by predicting the augmentation parameters (Lee et al., 2020, 2021), seeking transformation equivariance (Dangovski et al., 2022), or employing techniques that improve performance when using a linear projector (Jing et al., 2022). Importantly, we seek to both retain and separate/disentangle style information using a theoretically-grounded framework. Further related work on disentanglement in generative models and identifiable CRL is discussed in App. C.1.

Outlook. Our framework leverages structured data augmentations to identify not only invariant content latents, but also varying style latents from multiple transformed views. Future work may investigate new types of data augmentations and their use for self-supervised causal representation learning.

Acknowledgments and Disclosure of Funding

This work was supported by the Tübingen AI Center (FKZ: 01IS18039B) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. The authors declare no competing interests.

References

Ahuja, K., Hartford, J. S., and Bengio, Y. (2022). Weakly supervised representation learning with sparse perturbations. In *Advances in Neural Information Processing Systems*, volume 35, pages 15516–15528. [Cited on p. 6 and 16.]

Bardes, A., Ponce, J., and LeCun, Y. (2022). VICReg: variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*. [Cited on p. 1, 2, 7, 15, and 16.]

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828. [Cited on p. 1.]

Bouchacourt, D., Tomioka, R., and Nowozin, S. (2018). Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Arti cial Intelligence*, volume 32. [Cited on p. 16.]

Brehmer, J., De Haan, P., Lippe, P., and Cohen, T. (2022). Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 38319–38331. [Cited on p. 6 and 16.]

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1994). Signature veri cation using a "siamese" time delay neural network. *Advances in Neural Information Processing Systems*, 6. [Cited on p. 1.]

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. [Cited on p. 1.]

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. [Cited on p. 1, 2, 7, and 15.]

Chen, T., Luo, C., and Li, L. (2021). Intriguing properties of contrastive losses. *Advances in Neural Information Processing Systems*, 34:11834–11845. [Cited on p. 2.]

Chen, X., Fan, H., Girshick, R., and He, K. (2020b). Improved baselines with momentum contrastive learning. *arXiv:2003.04297*. [Cited on p. 2.]

Chen, X. and He, K. (2021). Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758. [Cited on p. 2.]

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , and Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. [Cited on p. 14.]

Cole, E., Yang, X., Wilber, K., Mac Aodha, O., and Belongie, S. (2022). When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14755–14764. [Cited on p. 1.]

Dangovski, R., Jing, L., Loh, C., Han, S., Srivastava, A., Cheung, B., Agrawal, P., and Soljacic, M. (2022). Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*. [Cited on p. 7.]

Darmois, G. (1951). Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231. [Cited on p. 13.]

Daunhawer, I., Bizeul, A., Palumbo, E., Marx, A., and Vogt, J. E. (2023). Identi ability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*. [Cited on p. 4 and 16.]

Desjardins, G., Courville, A., and Bengio, Y. (2012). Disentangling factors of variation via generative entangling. *arXiv preprint arXiv:1210.5474*. [Cited on p. 16.]

Eastwood, C., Nicolicioiu, A. L., Von Kügelgen, J., Kekić, A., Träuble, F., Dittadi, A., and Schölkopf, B. (2023). Dci-es: An extended disentanglement framework with connections to identi ability. In The Eleventh International Conference on Learning Representations. [Cited on p. 16.]

Eastwood, C. and Williams, C. K. I. (2018). A framework for the quantitative evaluation of disentangled representations. In International Conference on Learning Representations. [Cited on p. 16.]

Ericsson, L., Gouk, H., and Hospedales, T. M. (2021). How well do self-supervised models transfer? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5414–5423. [Cited on p. 1, 7, and 16.]

Ermolov, A., Siarohin, A., Sangineto, E., and Sebe, N. (2021). Whitening for self-supervised representation learning. In International Conference on Machine Learning, pages 3015–3024. [Cited on p. 2.]

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2007). The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. [Cited on p. 14.]

Garrido, Q., Chen, Y., Bardes, A., Najman, L., and LeCun, Y. (2023). On the duality between contrastive and non-contrastive self-supervised learning. In International Conference on Learning Representations. [Cited on p. 2.]

Goyal, P., Dollár, P., Girshick, R. B., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. (2017). Accurate, large minibatch sgd: Training imagenet in 1 hour. ArXiv. [Cited on p. 16.]

Gresele, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Schölkopf, B. (2019). The incomplete rosetta stone problem: Identi ability results for multi-view nonlinear ICA. In Proceedings of the 35th Conference on Uncertainty in Arti cial Intelligence, volume 115, pages 217–227. PMLR. [Cited on p. 6 and 16.]

Gresele, L., Von Kügelgen, J., Stimper, V., Schölkopf, B., and Besserve, M. (2021). Independent mechanism analysis, a new concept? Advances in Neural Information Processing Systems, volume 34, pages 28233–28248. [Cited on p. 13.]

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Pires, B. A., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. (2020). Bootstrap Your Own Latent: A new approach to self-supervised learning. Advances in Neural Information Processing Systems. [Cited on p. 2.]

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9729–9738. [Cited on p. 2.]

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In CVPR. [Cited on p. 16.]

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2017). b-VAE: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations. [Cited on p. 14 and 16.]

Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. Neural networks, 12(3):429–439. [Cited on p. 13 and 16.]

Ilse, M., Tomczak, J. M., and Forré, P. (2021). Selecting data augmentation for simulating interventions. In International Conference on Machine Learning, pages 4555–4562. PMLR. [Cited on p. 5.]

Jing, L., Vincent, P., LeCun, Y., and Tian, Y. (2022). Understanding dimensional collapse in contrastive self-supervised learning. In International Conference on Learning Representations. [Cited on p. 7.]

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. (2013). 3d object representations for ne-grained categorization. In 2013 IEEE International Conference on Computer Vision Workshops. [Cited on p. 14.]

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. [Cited on p. 14.]

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. (2015). Deep convolutional inverse graphics network. Advances in neural information processing systems, 28. [Cited on p. 16.]

Lee, H., Hwang, S. J., and Shin, J. (2020). Self-supervised label augmentation via input transformations. In International Conference on Machine Learning, pages 5714–5724. [Cited on p. 7.]

Lee, H., Lee, K., Lee, K., Lee, H., and Shin, J. (2021). Improving transferability of representations via augmentation-aware self-supervision. Advances in Neural Information Processing Systems. [Cited on p. 7 and 16.]

Li, F.-F., Andreeto, M., Ranzato, M., and Perona, P. (2022). Caltech 101. [Cited on p. 14.]

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In 36th International Conference on Machine Learning, pages 7247–7283. Curran Associates, Inc. [Cited on p. 6, 14, and 16.]

Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. International Conference on Machine Learning pages 6348–6359. PMLR. [Cited on p. 6 and 16.]

Loshchilov, I. and Hutter, F. (2017). Sgdr: Stochastic gradient descent with restarts. ICLR. [Cited on p. 16.]

Lyu, Q., Fu, X., Wang, W., and Lu, S. (2021). Understanding latent correlation-based multiview learning and self-supervision: An identi ability perspective. In International Conference on Learning Representations. [Cited on p. 6 and 16.]

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. (2013). Fine-grained visual classi cation of aircraft. arXiv preprint arXiv:1306.5151. [Cited on p. 14.]

Mitrovic, J., McWilliams, B., Walker, J. C., Buesing, L. H., and Blundell, C. (2021). Representation learning via invariant causal mechanisms. In International Conference on Learning Representations. [Cited on p. 5.]

Nilsback, M.-E. and Zisserman, A. (2008). Automated ower classi cation over a large number of classes. In Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing. [Cited on p. 14.]

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing ows for probabilistic modeling and inference. The Journal of Machine Learning Research, 22(1):2617–2680. [Cited on p. 13.]

Parkhi, O., Vedaldi, A. ., Zisserman, A., and Jawahar, C. V. (2012). Cats and dogs. IEEE Conference on Computer Vision and Pattern Recognition. [Cited on p. 14.]

Pearl, J. (2009). Causality: Models, Reasoning, and Inference. Cambridge University Press. [Cited on p. 5.]

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. (2014). Imagenet large scale visual recognition challenge. CoRR. [Cited on p. 7 and 14.]

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5):612–634. [Cited on p. 4.]

Squires, C., Seigal, A., Bhate, S., and Uhler, C. (2023). Linear causal disentanglement via interventions. In 40th International Conference on Machine Learning. [Cited on p. 6.]

Suter, R., Miladinovic, D., Schölkopf, B., and Bauer, S. (2019). Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. International Conference on Machine Learning, pages 6056–6065. [Cited on p. 5.]

Tenenbaum, J. and Freeman, W. (1996). Separating style and content. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press. [Cited on p. 16.]

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. (2021). Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, pages 16451–16467. [Cited on p. 4, 5, 6, 13, 14, 15, and 16.]

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology. [Cited on p. 14.]

Wang, T. and Isola, P. (2020). Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR. [Cited on p. 2 and 15.]

Wang, Y. and Jordan, M. I. (2021). Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*. [Cited on p. 5.]

Xiao, J., Hays, J., Ehinger, K., Oliva, A., and Torralba, A. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. [Cited on p. 14.]

Xiao, T., Wang, X., Efros, A. A., and Darrell, T. (2021). What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*. [Cited on p. 3, 7, 12, 16, 17, and 18.]

You, Y., Gitman, I., and Ginsburg, B. (2017). Large batch training of convolutional networks. *arXiv: Computer Vision and Pattern Recognition*. [Cited on p. 16.]

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. [Cited on p. 1, 2, and 15.]

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. (2021). Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*. [Cited on p. 13.]

Appendices

Table of Contents

A Proof of Thm. 4.2

Theorem 4.2 (Identifiability). *For the data generating process (4.1), (4.7), (4.8), assume that*

- A_1. *$\mathcal{Z}$ is open and simply connected, g is diffeomorphic onto its image, p_z is smooth and fully supported on $\mathcal{Z}$; each $p_{\tilde{s}_m|s_m}$ is smooth and supported on an open, non-empty set around any s_m,*
- A_2. *p_z and $\{p_{\tilde{s}_m|s_m}\}_{m=1}^M$ are such that $\{c\} \cup \{\tilde{s}_m\}_{m=1}^M$ are jointly independent;*
- A_3. *the latent dims $\{d_m\}_{m=0}^M$ are known and $\{f_m : \mathcal{X} \to (0,1)^{d_m}\}_{m=0}^M$ are smooth minimizers of*

$$\sum_{m=0}^{M} \mathbb{E}_{p^{(m)}_{\tilde{x},\tilde{x}^0}} \left\| f_m(\tilde{x}) - f_m(\tilde{x}^0) \right\|_2 - H_{p^{(0)}_{\tilde{x}}}([f_0(\tilde{x}), \ldots, f_M(\tilde{x})]) \,. \tag{4.10}$$

Then f_0 block-identifies (von Kügelgen et al., 2021, Defn. 4.1) the content c, and f_m block-identifies s_m in the sense that $\hat{s}_m = f_m(x) = \gamma_m(s_m)$ for some invertible γ_m for each $m = 1, \ldots, M$.

Proof. The proof follows a similar argument as that of von Kügelgen et al. (2021, Thm. 4.4), extended to our setting with $M+1$ alignment terms instead of a single one, and without entropy regularization.

Step 1. First, we show the existence of a solution $\{f_m\}_{m=0}^M$ attaining the global minimum of zero of the objective in (4.10). To this end, we construct each f_m by composing the inverse of the true mixing function with the cumulative distribution function (CDF) transform[2] to map each latent block to a uniform version of itself. Specifically, let $f_0 := F_c \circ f^{-1}_{1:d_0}$, and for $m = 1, \ldots, M$, let $f_m := F_{s_m} \circ f^{-1}_{a_m:b_m}$ with $a_m = 1 + \sum_{l=0}^{m-1} d_l$ and $b_m = \sum_{l=0}^{m} d_l$, where F_v denotes the CDF of v. By construction, $f_0(\tilde{x})$ is a function of c only, and uniformly distributed on $(0,1)^{d_0}$; similarly, $f_m(\tilde{x})$ is a function of $\tilde{s}_m$ only and uniform on $(0,1)^{d_m}$ for $m = 1, \ldots, M$. Recall that, with probability one, c is shared across $(\tilde{x}, \tilde{x}^0) \sim p^{(0)}_{\tilde{x},\tilde{x}^0}$ and $\tilde{s}_m$ is shared across $(\tilde{x}, \tilde{x}^0) \sim p^{(m)}_{\tilde{x},\tilde{x}^0}$. Hence, all the alignment (expectation) terms in (4.10) are zero. Finally, since $\{c\} \cup \{\tilde{s}_m\}_{m=1}^M$ are mutually independent by assumption A_2, and since each f_m for $m = 0, \ldots, M$ is uniform on $(0,1)^{d_m}$, it follows that $[f_0(\tilde{x}), \ldots, f_M(\tilde{x})]$ is jointly uniform on $(0,1)^d$. Hence, the entropy term in (4.10) is also zero.

Step 2. Next, let $\{f_m\}_{m=0}^M$ be any other solution attaining the global minimum of (4.10). By the above existence argument, this implies that (i) $f_m(\tilde{x}) = f_m(\tilde{x}^0)$ almost surely w.r.t. $p^{(m)}_{\tilde{x},\tilde{x}^0}$ for $m = 0, \ldots, M$; and (ii) $[f_0(\tilde{x}), \ldots, f_M(\tilde{x})]$ is jointly uniform on $(0,1)^d$. As shown by von Kügelgen et al. (2021), the invariance constraint (i) together with the postulated data generating process and assumption A_1 implies that each $f_m \circ f$ can only be a function of the latents that are shared almost surely across $(\tilde{x}, \tilde{x}^0) \sim p^{(m)}_{\tilde{x},\tilde{x}^0}$. That is, $f_0 \circ f(x) = \gamma_0(c)$ and $f_m \circ f(x) = \gamma_m(c, s_m)$ for $m = 1, \ldots, M$. By A_1 and constraint (ii), γ_0 maps a regular density to another regular density and thus must be invertible (Zimmermann et al., 2021, Prop. 5).

Step 3. It remains to show that γ_m is invertible and actually cannot depend on c for $m = 1, \ldots, M$, for this would otherwise violate the maximum entropy (uniformity) constraint (ii). Suppose for a contradiction that γ_k depends on c for some $k \in \{1, \ldots, M\}$. By constraint (ii), $[f_0(\tilde{x}), f_k(\tilde{x})] = [\gamma_0(c), \gamma_k(c, \tilde{s}_k)]$ is jointly uniform on $(0,1)^{d_0+d_k}$. Hence, $\gamma_0(c)$ and $\gamma_k(c, \tilde{s}_k)$ are independent. Since γ_0 is invertible, this implies that c and $\gamma_k(c, \tilde{s}_k)$ are independent, which (by smoothness of $\gamma_k = f_k \circ f$ and independence of c and $\tilde{s}_k$) contradicts the assumption that γ_k depends on c.

Thus, by contradiction, we have that $f_m(\tilde{x}) = \gamma_m(c, \tilde{s}_m) = \gamma_m(\tilde{s}_m)$ for $m = 1, \ldots, M$. Finally, invertibility of $\gamma_m(\tilde{s}_m)$ for $m = 1, \ldots, M$ follows from A_1 and Prop. 5 of Zimmermann et al. (2021). Together with $f_0(\tilde{x}) = \gamma_0(c)$ (established above) concludes the proof of block-identifiability. □

[2] Sometimes also referred to as "Darmois construction" (Darmois, 1951; Hyvärinen and Pajunen, 1999; Gresele et al., 2021; Papamakarios et al., 2021).

(a) Weak 1 (b) Weak 2

(c) Strong 1 (d) Strong 2

Figure 3: Augmentation strengths on ColorDSprites. Columns show augmentation pairs of the same strength. Note that images are more similar across (a) & (b) than across (c) & (d), in terms of color, orientation, scale, translation and X-Y position.

B Implementation Details

B.1 Datasets

The numerical data is based on the experiments of von Kügelgen et al. (2021), and the data samples are generated programmatically. ColorDSprites is a synthetic image dataset based on DSprites (Higgins et al., 2017), and extended by Locatello et al. (2019). The rest of the experiments are based on models pretrained on ImageNet1k (Russakovsky et al., 2014), which are then evaluated on the downstream datasets FGVC Aircraft (Maji et al., 2013), Caltech-101 (Li et al., 2022), Stanford Cars (Krause et al., 2013), CIFAR10 (Krizhevsky, 2009), CIFAR100 (Krizhevsky, 2009), CUB (Wah et al., 2011), DTD (Cimpoi et al., 2014), Oxford Flowers (Nilsback and Zisserman, 2008), Oxford-IIIT Pets (Parkhi et al., 2012), SUN397 (Xiao et al., 2010) and VOC2007 (Everingham et al., 2007).

ColorDSprites samples. Fig. 3 depicts samples from the ColorDprites dataset when transformed with transformations/augmentations of different strengths.

B.2 Invariance and entropy terms

To remain general and apply to any contrastive (e.g., SimCLR, Chen et al. 2020a) or non-contrastive covariance-based SSL method (e.g., VICReg, Bardes et al. 2022), both Eq. (2.1) and Eq. (3.2) are expressed as a combination of invariance L^{inv} and entropy L^{ent} terms. For concreteness, Table 3 specifies these terms for some common SSL methods, namely SimCLR (Chen et al., 2020a), BarlowTwins (BTs, Zbontar et al. 2021), and VICReg (Bardes et al., 2022). Note a slight misalignment between $L^{ent}(Z, Z^0)$ in Table 3 and our usage of it in Eq. (3.2). In particular, we write $L^{ent}\{z^m, z'^m\}_{m=0}^M$ for brevity, but should write $L^{ent}(Z, Z^0)$ with $Z = [z^0, z^1, \ldots, z^M]$ to align with Table 3.

Table 3: Unified Perspective on SSL Objectives Through Invariance and Entropy. Many SSL methods can be expressed as a (weighted) combination of invariance L^{inv} and entropy L^{ent} terms. Here, $Z = [z^1, z^2, \ldots, z^n]$ and $Z^0 = [z'^1, z'^2, \ldots, z'^n]$ are two batches of n vectors of d-dimensional representations with $Z, Z^0 \in \mathbb{R}^{n \times d}$; $Z_j \in \mathbb{R}^n$ is a vector composed of the values at dimension j for all n vectors in Z; $C(Z) = 1/(n-1) \sum_i (z^i - \bar{z})(z^i - \bar{z})^T$ is the (sample) covariance matrix of Z with $\bar{z} = 1/n \sum_{i=1}^n z^i$; and λ_v, λ_c are hyperparameters for weighting the variance and covariance terms, respectively.

Algorithm	$L^{inv}(Z, Z^0)$	$L^{ent}(Z, Z^0)$
SimCLR	$\frac{1}{n}\sum_{i=1}^n \frac{z_i^T z_i'}{\|z_i\|\|z_i'\|}$	$\frac{1}{n}\sum_{i=1}^n \log \sum_{j=1, j\neq i}^n \exp \frac{z_i^T z_j'}{\|z_i\|\|z_j'\|}$
BTs	$\sum_{j=1}^d \left(1 - \frac{(Z_j)^T Z_j'}{\|Z_j\|\|Z_j'\|}\right)^2$	$\sum_{j=1}^d \sum_{k=1, k\neq j}^d \left(\frac{(Z_j)^T Z_k'}{\|Z_j\|\|Z_k'\|}\right)^2$
VICReg	$\frac{1}{n}\sum_{i=1}^n \|z_i - z_i'\|_2^2$	$\frac{\lambda_v}{d}\sum_{j=1}^d \max\left(0, 1 - \sqrt{\mathrm{Var}(Z_j) + \epsilon}\right) + \max\left(0, 1 - \sqrt{\mathrm{Var}(Z_j') + \epsilon}\right) + \frac{\lambda_c}{d}\sum_{i=1}^n \sum_{j=1, j\neq i}^n [C(Z)]_{i,j}^2 + [C(Z')]_{i,j}^2$

B.3 Adaptive λ_m

We now describe our procedure for adaptively updating our hyperparameters in Eq. (3.2) using a dual-ascent approach. To motivate this approach, first note that our placement of the invariance terms L^{inv} differs from the standard approach of placing it on the entropy term (Wang and Isola, 2020; Zbontar et al., 2021). Doing so allows us to:

- View L^{inv} as a constraint that should be satisfied. We view the goal of Eq. (2.1) as the soft/unconstrained version of the following constrained problem: maximize L^{ent} subject to $L^{inv} = 0$. As a result, we then view λ_m as a Lagrange multiplier which should be set such that the invariance constraint is satisfied to within some acceptable tolerance, i.e., $L^{inv} < \epsilon$. This way of choosing λ diverges from the standard approach to choosing the invariance-entropy trade-off in SSL (implicitly or explicitly), where it is chosen to maximize performance on some downstream task (e.g., ImageNet object classification accuracy).
- Iteratively updating λ_m during training using a dual-ascent approach. While we could take a standard grid-search approach to choosing λ_m such that this invariance constraint is satisfied at the end of training, we instead iteratively adapt λ_m during training using a dual-ascent approach. In particular, given a step size or learning rate η and tolerance level ϵ, we perform iterative gradient-based updates of both the model parameters (inner loop) and λ_m (outer loop) with $\lambda_m^t \leftarrow \lambda_m^{t-1} + \eta \cdot \mathrm{relu}(L^{inv}(q^t) - \epsilon)$.

B.4 Numerical dataset

Following von Kügelgen et al. (2021, Sec. 5.1), we generate synthetic data $(x, \tilde{x}) = (f(c, s), f(c, \tilde{s}))$ with content $c \sim N(0, S_c)$, style $s|c \sim N(a + Bc, S_s)$, and perturbed style $\tilde{s} \sim N(s, S_{\tilde{s}})$. We choose the simplest setup with a, B, S_c, S_s and $S_{\tilde{s}}$ set to the identity. See von Kügelgen et al. (2021, App. D) for further details on the data-generation process.

B.5 ImageNet

Pretraining. Our ImageNet1k pretraining setup is based on the settings in (Bardes et al., 2022), which can be consulted for full details. We train ResNet50 (He et al., 2016) models for only 100 epochs, with 3-layer projectors of dimension 8196. The optimizer is LARS (You et al., 2017; Goyal et al., 2017), the batch size is 2048 and the learning rate follows a cosine decay schedule (Loshchilov and Hutter, 2017).

The data augmentation also follows Bardes et al. (2022) and is applied asymetrically to the two views. It includes crops, ips, color jitter, grayscale, solarize and blur. These atomic augmentations are split into two groups: spatial (crops and ips) and appearance (color jitter, grayscale, solarize and blur). Thus, the number of "style" attributes in this setting are $m = 2$.

While we aim for fair experiments that use default hyperparameters, projectors, and augmentation settings, we note that these are optimized for existing SSL methods that prioritize information removal. Perhaps other settings, such as different augmentations explored in Xiao et al. (2021) and Lee et al. (2021), can be bene cial in our framework which instead aims to retain and disentangle information.

Downstream evaluation. Our downstream evaluation follows that of Ericsson et al. (2021). We train linear models (logistic or ridge regression) on frozen pre-projector representations h and post-projector embeddings z. Images are cropped to 224×224, with L2 regularization searched using 5-fold cross-validation over 45 logarithmically spaced values in the range 10^{-6} to 10^{5}.

C Further Related Work

C.1 Disentangled and Identi able Representation Learning

Disentanglement in generative models. In a generative setting, disentangled representations are commonly sought (Desjardins et al., 2012; Higgins et al., 2017; Eastwood and Williams, 2018; Eastwood et al., 2023). In the vision-as-inverse-graphics paradigm, separating or disentangling "content" and "style" has a long history (Tenenbaum and Freeman, 1996; Kulkarni et al., 2015). Purely based on i.i.d. data and without assumptions on the model-class, disentangled representation learning is generally impossible (Hyvärinen and Pajunen, 1999; Locatello et al., 2019). More recently, generative disentanglement has thus been pursued with additional weak supervision in the form of paired data (Bouchacourt et al., 2018; Locatello et al., 2020).

Identi ability in disentangled and causal representation learning. Our Thm. 4.2 can be viewed as an extension of the content block-identi ability result of von Kügelgen et al. (2021, Thm. 4.4), which was generalized to a multi-modal setting with distinct mixing functions $f_1 \neq f_2$ and additional modality-speci c latents by Daunhawer et al. (2023, Thm. 1). The two options discussed at the end of § 4 for satisfying assumption A_2—(a) independent style variables, and (b) perfect interventions—can be used to draw additional links to existing identi ability results. Option (a) relates to a result of Lyu et al. (2021, Thm. 2) showing that $\tilde{z}$ and $\tilde{s}^0$ can be block-identi ed through latent correlation maximization with invertible encoders, provided that c, s, and $\tilde{s}^0$ are mutually independent. Thm. 4.2 establishes a more ne-grained disentanglement into individual style components. On the other hand, Gresele et al. (2019) and Locatello et al. (2020) prove identi ability of individual latents for the setting in which all latents are mutually independent and subject to change (with probability > 0), i.e., without an invariant block of content latents. Option (b) relates to a result of Brehmer et al. (2022, Thm. 1) showing that all variables (and the graph) in a causal representation learning setup can be identi ed through weak supervision in the form of pairs $(x, \tilde{x})$ arising from single-node perfect interventions by tting a generative model via maximum likelihood. Perhaps most closely related is the work of Ahuja et al. (2022) who do not assume independence of latents, and also consider learning from M views arising from sparse perturbations, but require perturbations on all latent blocks for full identi ability.

C.2 Detailed comparison with Xiao et al. (2021)

Table 4 presents the key differences between the framework of Xiao et al. (2021) and ours. Both rely on structured augmentations, by which views are constructed using augmentations that either share or do not share the same parameters. Both frameworks also learn multiple embedding spaces to capture style attributes of the data. However, our goal is not only to learn multiple embedding spaces but to

Table 4: High-level comparison with Xiao et al. (2021). While both use structured augmentations and multiple embedding spaces to capture style attributes of the data, only ours seeks disentangled embedding spaces and provides theoretical grounding/analyses.

Method	Structured augmentations	Multiple embeddings	Disentangled embeddings	Theoretical underpinning
Xiao et al. (2021)	3	3	7	7
Ours	3	3	3	3

fully disentangle them. This is achieved in our framework by the careful combination of invariance and entropy terms in Eq. (3.2), including the removal of redundant information with an entropy term across the joint embedding space. Furthermore, in § 4 and App. A, we provide theoretical analysis with the conditions under which our framework identifies the underlying style attributes or features.

In addition to these key, high-level differences, there are several smaller differences at the implementation level. In particular, we use an optimization procedure that adaptively sets the hyperparameters to guarantee the disentanglement of content and style (see App. B.3). We also adopt a more general construction of our framework, instantiating it with multiple different SSL methods, SimCLR, BTs and VICReg. Finally, our construction of image views allows more negative samples and in a given batch, compared to the query-key construction of Xiao et al. (2021)—see Fig. 4.

D Further Results

We now present additional results.

D.1 ColorDSprites

Table 5 shows VICReg's sensitivity to augmentation strengths with fixed and adaptive λ on ColorDSprites, complementing the results for SimCLR in Table 1.

Table 5: VICReg's sensitivity to augmentation strengths with fixed and adaptive λ on ColorDSprites. r^2 in predicting the ground-truth factor values from the post-projector embedding z with a linear classifier. Adapting λ ensures that z captures all of content ($\overline{C} = 1$) and almost no style ($\overline{S} = 0$), regardless of the augmentation strengths. SimCLR results in Table 1.

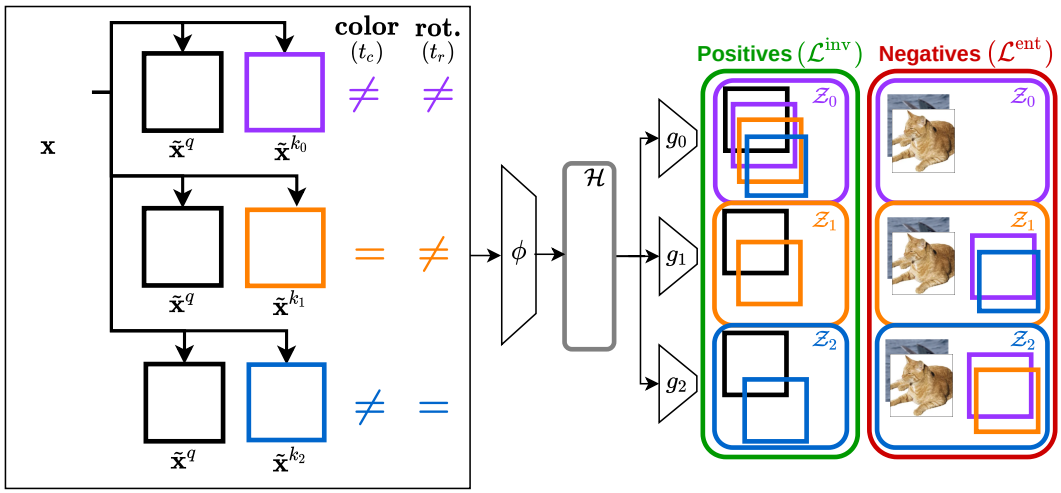
Algorithm	λ	Augm. Str.	Content (C)	Style (S)					$\overline{S}$ (#)
			Shape	Color	Scale	Orient.	PosX	PosY	
VICReg	fixed	Weak	1.0	0.87	0.71	0.29	0.45	0.45	0.55
	fixed	Medium	1.0	0.40	1.00	0.05	0.56	0.56	0.51
	fixed	Strong	1.0	0.12	0.99	0.08	0.62	0.62	0.49
	adaptive	Weak	1.0	0.20	0.17	0.00	0.00	0.00	0.07
	adaptive	Medium	1.0	0.10	0.52	0.00	0.00	0.01	0.13
	adaptive	Strong	1.0	0.10	0.53	0.00	0.00	0.00	0.13

D.2 ImageNet

Table 6 gives the downstream results for VICReg, analogous to Table 2 which gives the results for SimCLR.

Table 6: Linear evaluation on ImageNet and a broad range of downstream tasks. We show top-1 accuracies (%) for all but CUB_{bbox} (r^2), CUB_{kpt} (r^2), and VOC (AP_{50}). We use frozen representations h and embeddings z (post-projector). FT: our framework fine-tuned as a base VICReg model. Ct101: CalTech101. Cf10: CIFAR10.

Alg.	Feat.	ImNt	Acft	Ct101	Cars	Cf10	Cf100	CUB_{bbox}	CUB_{cls}	CUB_{kpt}	DTD	Flwrs	Pets	SUN	VOC	Avg.
VICReg	z	55.7	10.6	69.5	9.5	75.0	48.3	27.6	17.1	10.8	64.6	61.3	68.4	46.7	75.6	45.0
VICReg-Ours-FT	z	55.3	11.7	72.6	11.1	78.5	54.4	32.5	17.8	11.4	66.5	66.8	68.3	48.7	75.5	47.3
VICReg	h	67.2	51.1	87.6	52.6	88.3	70.1	69.1	47.4	31.9	75.3	93.4	83.1	59.7	79.6	68.4
VICReg-Ours-FT	h	66.9	50.1	87.5	52.4	88.4	70.3	69.8	47.2	32.8	75.7	93.6	82.7	59.8	79.6	68.5



(a) Xiao et al. (2021)

(b) Ours

Figure 4: **Comparison with Xiao et al. (2021).** Note the differences in data augmentation modules, as well as the embedding spaces in which positives and negatives are compared. See Xiao et al. (2021, Sec. 3) for details on their query-key notation. See App. C.2 for further details on this comparison.