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

Julius von Kügelgen*[1,2], Yash Sharma*[3,4], Luigi Gresele*[1], Wieland Brendel[3], Bernhard Schölkopf†[1], Michel Besserve†[1], Francesco Locatello†[5]
*equal contribution.
†equal supervision.

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany; [2]University of Cambridge; [3]University of Tübingen; [4]IMPRS for Intelligent Systems; [5]Amazon.

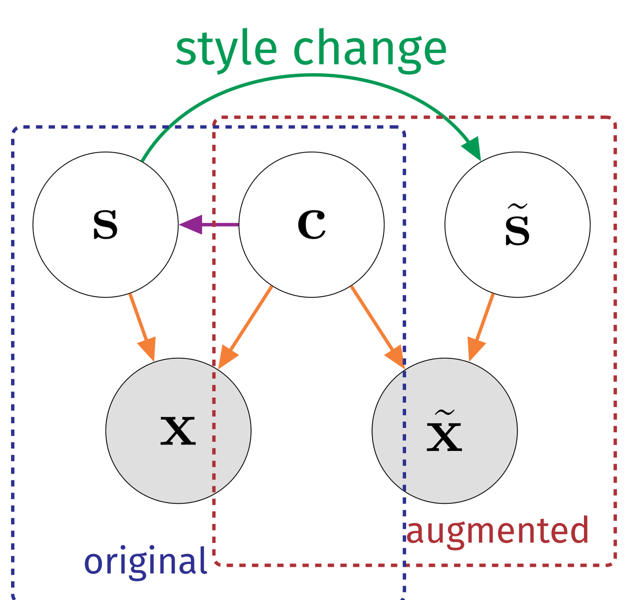
## Background

- **Self-supervised representation learning** has shown remarkable success in a numberof 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**.

## Data augmentation

For each observation $\mathbf{x}$, a pair of observation level transformations $\mathbf{t}, \mathbf{t}' \in \mathcal{T}$, $\mathbf{t}, \mathbf{t}' \sim p_{\mathbf{t}}$. is sampled and applied separately to $\mathbf{x}$ to generate a pair of augmented views $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = (\mathbf{t}(\mathbf{x}), \mathbf{t}'(\mathbf{x}))$.

Both $\mathcal{T}$ and $p_{\mathbf{t}}$ are designed using domain knowledge with the intention of **not changing the semantic characteristics** of the data (weak supervision).




We describe this data generating process through a **latent variable model**, and partition the latent variable $\mathbf{z}$ into content $\mathbf{c}$ and style $\mathbf{s}$, allowing 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.

*The choice of augmentations implicitly determines the partition!*

## Self-supervised representation learning (SSL)

A popular SSL objective function (used e.g., in `SimCLR` [1]) is InfoNCE [2] :

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

where $\mathbf{g}$ is an encoder, $\tilde{\mathbf{z}} = \mathbb{E}_{\mathbf{t} \sim p_{\mathbf{t}}}[\mathbf{g}(\mathbf{t}(\mathbf{x}))]$, $\tau$ is a temperature, and $K-1$ is the number of negative pairs, and $\text{sim}(\cdot, \cdot)$ is a similarity metric.

Contrastive SSL with negative samples using objective 1 can asymptotically be understood as **alignment with entropy regularisation** [3].

## Block-identifiability

Typical results in nonlinear ICA discuss **identifiability at the level of individual latent variables** [4]. We consider **block-identifiability** of the content partition.

**Definition: 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. $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{c})$.

## Our contributions

- We formulate the **augmentation process as a latent variable model**;
- **Content** is **invariant** to augmentation; **style** is **allowed to change**;
- We allow for both **nontrivial statistical and causal dependencies** in the latent space.
- We prove that generative and discriminative self-supervised learning with data augmentations isolates what is invariant across views, in the presence of nontrivial statistical and causal dependencies.
- We introduce `Causal3DIdent`, a dataset of high-dimensional, visually complex images with rich causal dependencies.

## Assumptions

- **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 $\mathcal{S} \times \mathcal{S}$.
- **Style changes.** Let $A$ be the set of subsets of style variables $A \subseteq \{1, ..., n_s\}$ and let $p_A$ be a distribution on $A$. 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 $A$.
- **Additional technical assumptions.**
  (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.

## Main theorem

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

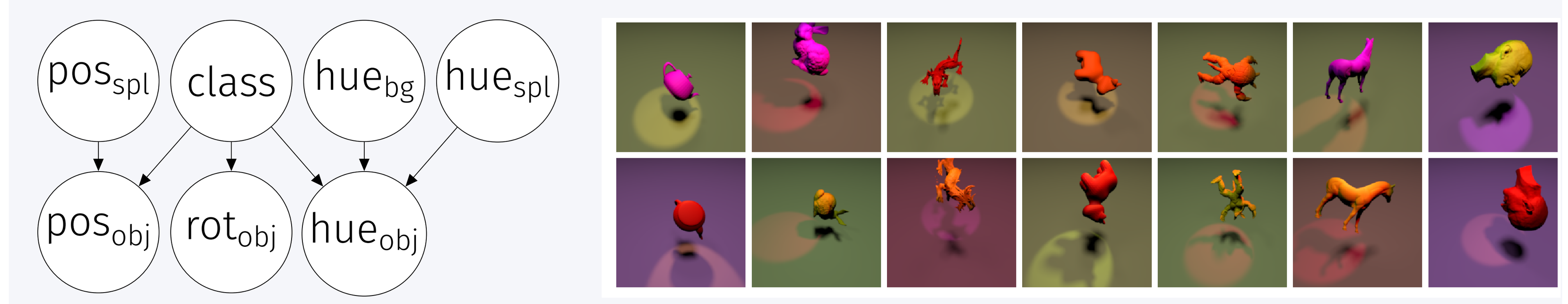
$$\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x}, \tilde{\mathbf{x}}) \sim p_{\mathbf{x}, \tilde{\mathbf{x}}}} \left[ \left( \mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}}) \right)^2 \right] - H(\mathbf{g}(\mathbf{x})) \tag{2}$$

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.

Significance:

- **Our theorem provides a theoretical justification for the empirically observed effectiveness of SSL with InfoNCE.**
- Also interesting connection with `BarlowTwins` [5], which only uses positive pairs and combines alignment with redundancy reduction.
- See paper for additional details, and a related theorem for the generative (as opposed to discriminative) setting.

## `Causal3DIdent` dataset




To empirically study identifiability in a causal representation learning context, we also introduce the `Causal3DIdent` dataset, containing rendered $224 \times 224$ images of 7 different 3D objects with 10 additional ground truth latent factors and causal dependence structure.

## Augmentations vs. latent space manipulations

*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.

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

**Summary of experimental findings:**

(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.

**Paper:** `https://arxiv.org/abs/2106.04619`
**Dataset:** `https://zenodo.org/record/4784282`

## References

[1] 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.

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

[3] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[4] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[5] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.