# Formalisms for Understanding Progress in Representation Learning

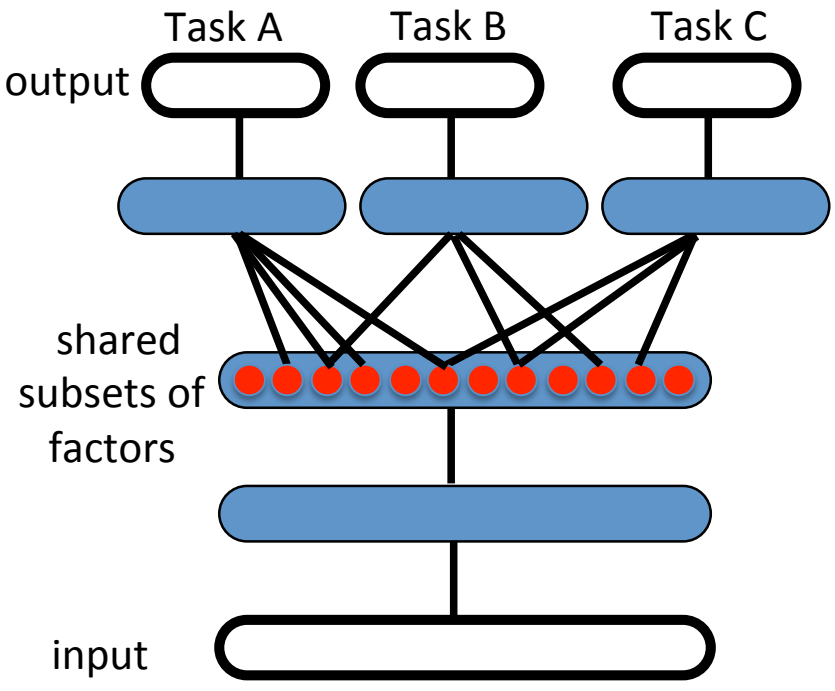
Yash Sharma[1,2,3]

[1] IMPRS for Intelligent Systems, [3] University of Tübingen, [3] MPI for Intelligent Systems Tübingen,

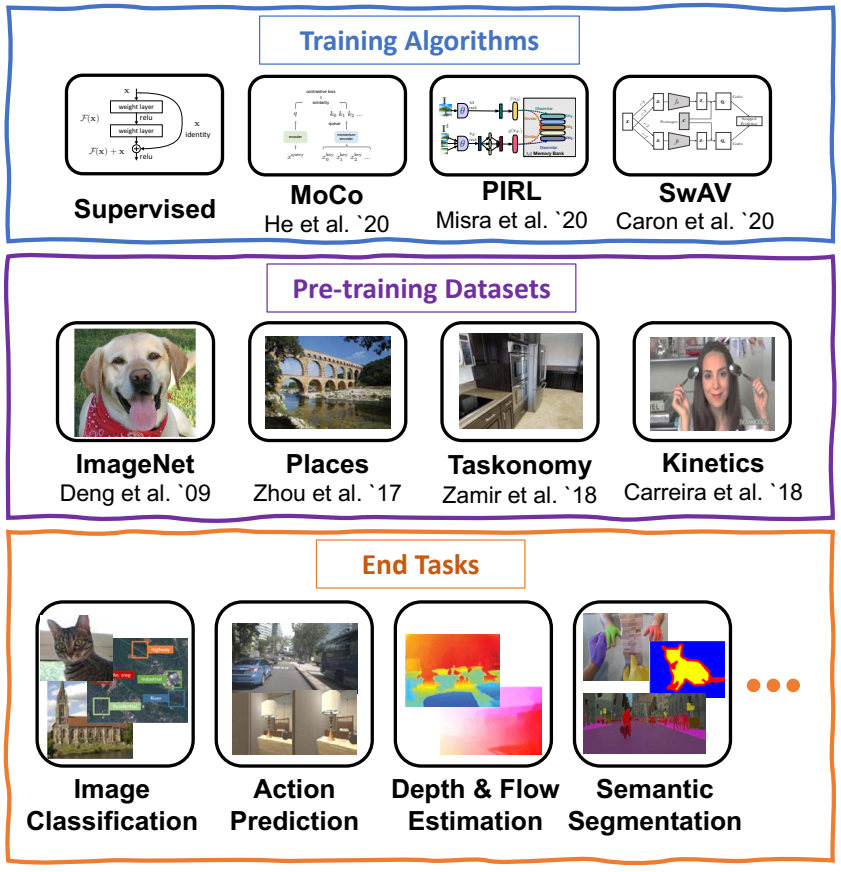


# Representation Learning

Learning representations of the data that facilitate the extraction of useful information for solving downstream tasks.




*(Bengio et al., 2013)*




*(Kotar et al., 2021)*

# Recent Progress

| pre-train | COCO keypoint detection $AP^{kp}$ | $AP^{kp}_{50}$ | $AP^{kp}_{75}$ |
|---|---|---|---|
| random init. | 65.9 | 86.5 | 71.7 |
| super. IN-1M | 65.8 | 86.9 | 71.9 |
| **MoCo** IN-1M | 66.8 (+1.0) | 87.4 (+0.5) | 72.5 (+0.6) |
| **MoCo** IG-1B | 66.9 (+1.1) | 87.8 (+0.9) | 73.0 (+1.1) |

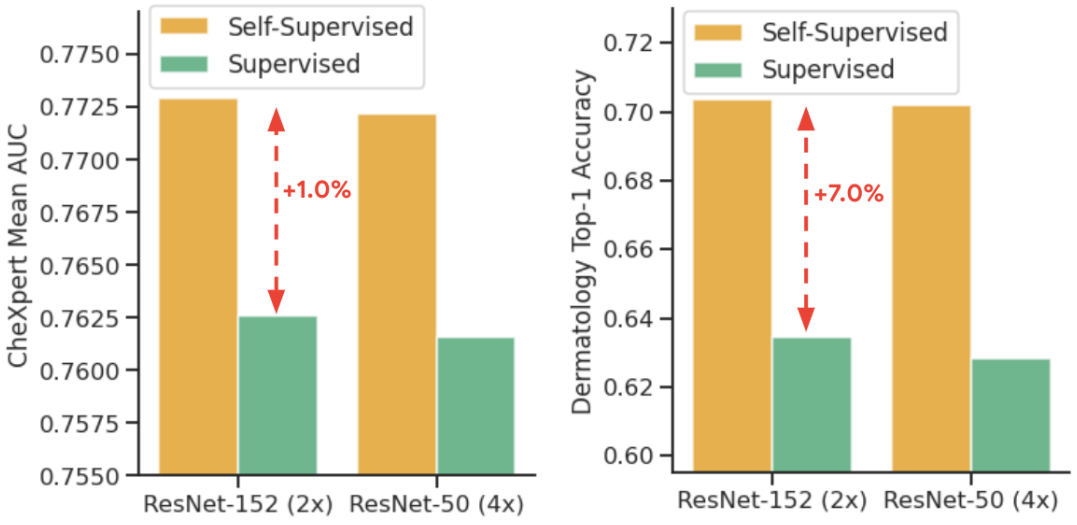
| pre-train | COCO dense pose estimation $AP^{dp}$ | $AP^{dp}_{50}$ | $AP^{dp}_{75}$ |
|---|---|---|---|
| random init. | 39.4 | 78.5 | 35.1 |
| super. IN-1M | 48.3 | 85.6 | 50.6 |
| **MoCo** IN-1M | 50.1 (+1.8) | 86.8 (+1.2) | 53.9 (+3.3) |
| **MoCo** IG-1B | 50.6 (+2.3) | 87.0 (+1.4) | 54.3 (+3.7) |

| pre-train | LVIS v0.5 instance segmentation $AP^{mk}$ | $AP^{mk}_{50}$ | $AP^{mk}_{75}$ |
|---|---|---|---|
| random init. | 22.5 | 34.8 | 23.8 |
| super. IN-1M† | 24.4 | 37.8 | 25.8 |
| **MoCo** IN-1M | 24.1 (−0.3) | 37.4 (−0.4) | 25.5 (−0.3) |
| **MoCo** IG-1B | 24.9 (+0.5) | 38.2 (+0.4) | 26.4 (+0.6) |

| pre-train | Cityscapes instance seg. $AP^{mk}$ | $AP^{mk}_{50}$ | Semantic seg. (mIoU) Cityscapes | VOC |
|---|---|---|---|---|
| random init. | 25.4 | 51.1 | 65.3 | 39.5 |
| super. IN-1M | 32.9 | 59.6 | 74.6 | 74.4 |
| **MoCo** IN-1M | 32.3 (−0.6) | 59.3 (−0.3) | 75.3 (+0.7) | 72.5 (−1.9) |
| **MoCo** IG-1B | 32.9 ( 0.0) | 60.3 (+0.7) | 75.5 (+0.9) | 73.6 (−0.8) |

Table 6. **MoCo *vs.* ImageNet supervised pre-training, fine-tuned on various tasks**. For each task, the same architecture and

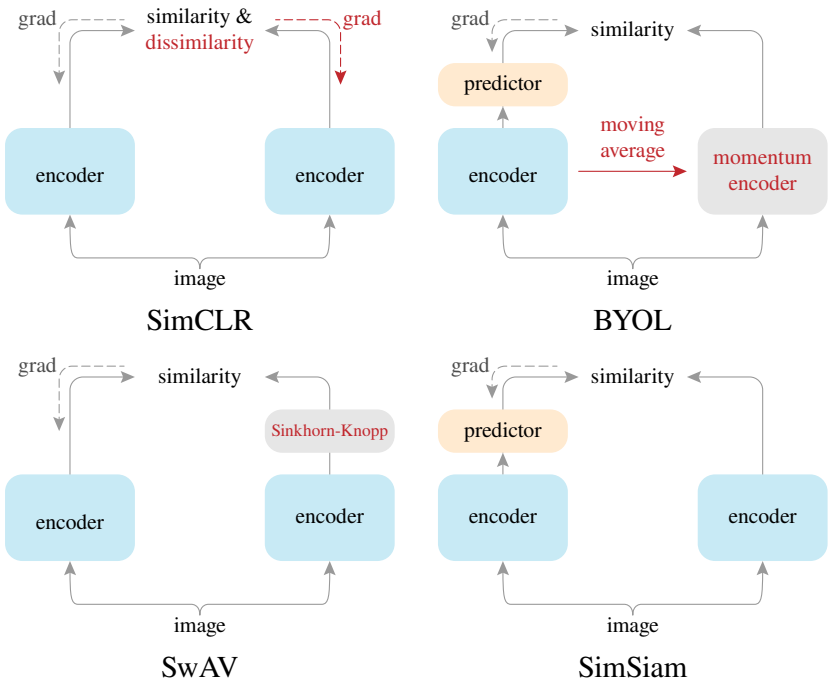
*(He et al., 2020)*




*(Azizi et al., 2021)*

# Architecture Comparison

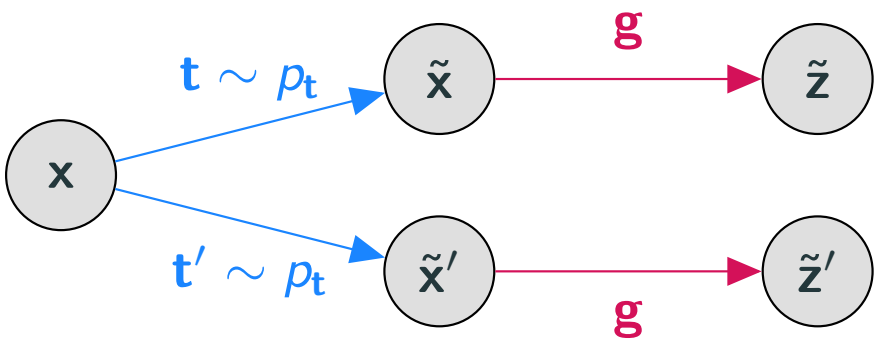
All methods maximize similarity between augmentations, subject to different conditions for avoiding collapse.




*(Chen & He, 2020)*

*(Chen et al., 2020)*

# Preliminaries: SSL with Data Augmentation




1. Specify a set $\mathcal{T}$ of transformations with distribution $p_{\mathbf{t}}$.
2. For each observation $\mathbf{x}$, sample pair of transformations $\mathbf{t}, \mathbf{t}' \sim p_{\mathbf{t}}$ and apply to $\mathbf{x}$ to form positive pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = (\mathbf{t}(\mathbf{x}), \mathbf{t}'(\mathbf{x}))$.
3. Learn encoder $\mathbf{g}$ s.t. representations $(\tilde{\mathbf{z}}, \tilde{\mathbf{z}}') = (\mathbf{g}(\tilde{\mathbf{x}}), \mathbf{g}'(\tilde{\mathbf{x}}'))$ are similar while *avoiding a collapsed (trivial) representation*.
   - Contrastive learning (CL): e.g., SimCLR *(Chen et al., 2020)*

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

*(Gutmann & Hyvärinen, 2010; van den Oord et al., 2018)*

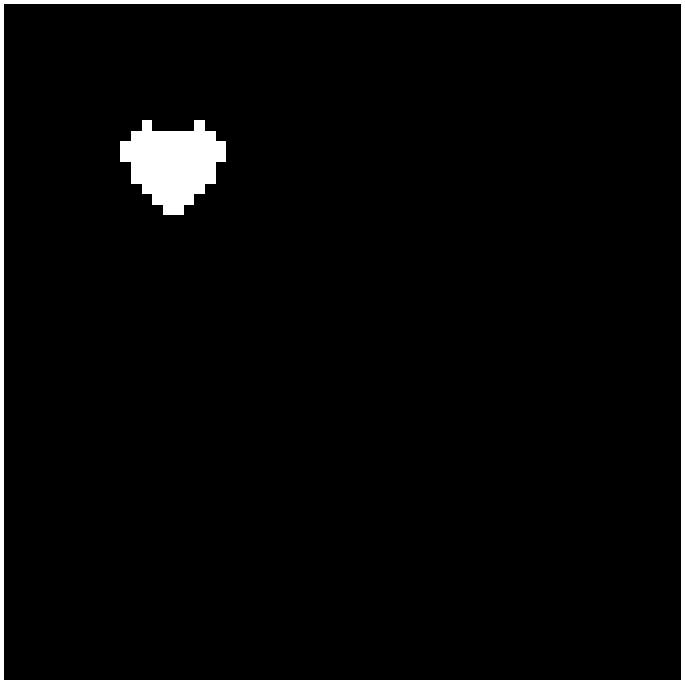
# Problem Setting I: Latent Variable Model

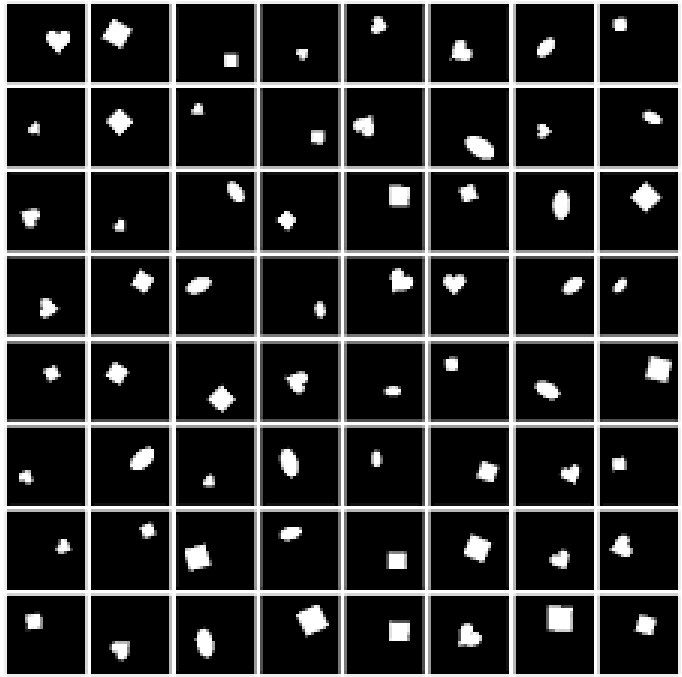
Formalise problem setting from latent variable model perspective:

**Data generation:** $\mathbf{z} \sim p_{\mathbf{z}}, \qquad \mathbf{x} = \mathbf{f}(\mathbf{z}),$

**Augmentation:** $\tilde{\mathbf{z}} \sim p_{\tilde{\mathbf{z}}|\mathbf{z}}, \qquad \tilde{\mathbf{x}} = \mathbf{f}(\tilde{\mathbf{z}}),$

with invertible decoder or mixing function $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$.

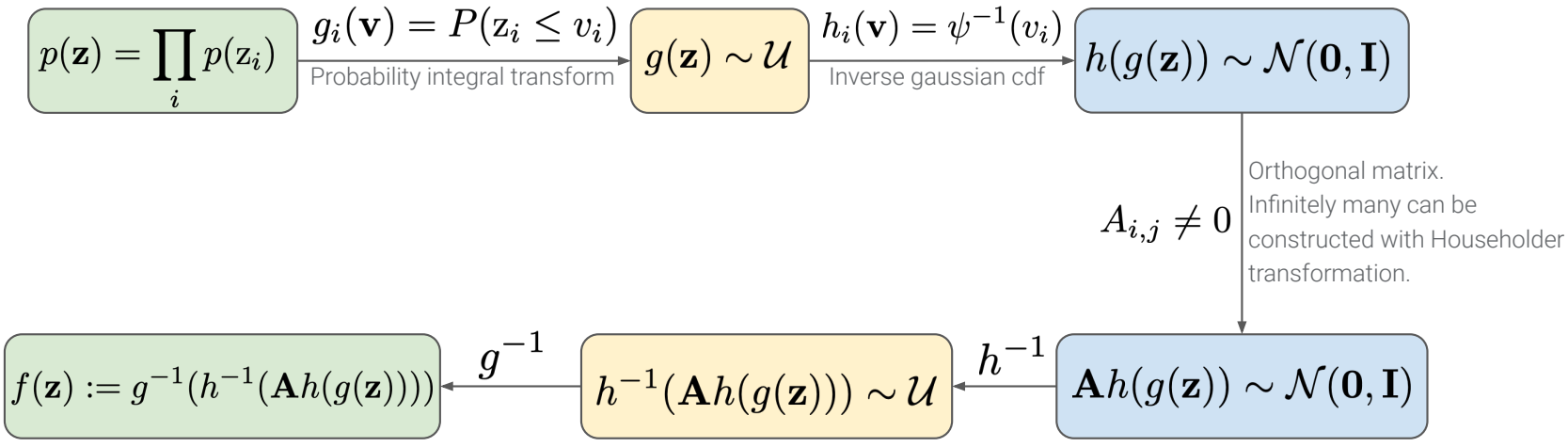




*(Matthey et al., 2017)*

# Identifiability: Impossible?

**Identifiability:** recover properties of the true generative process.



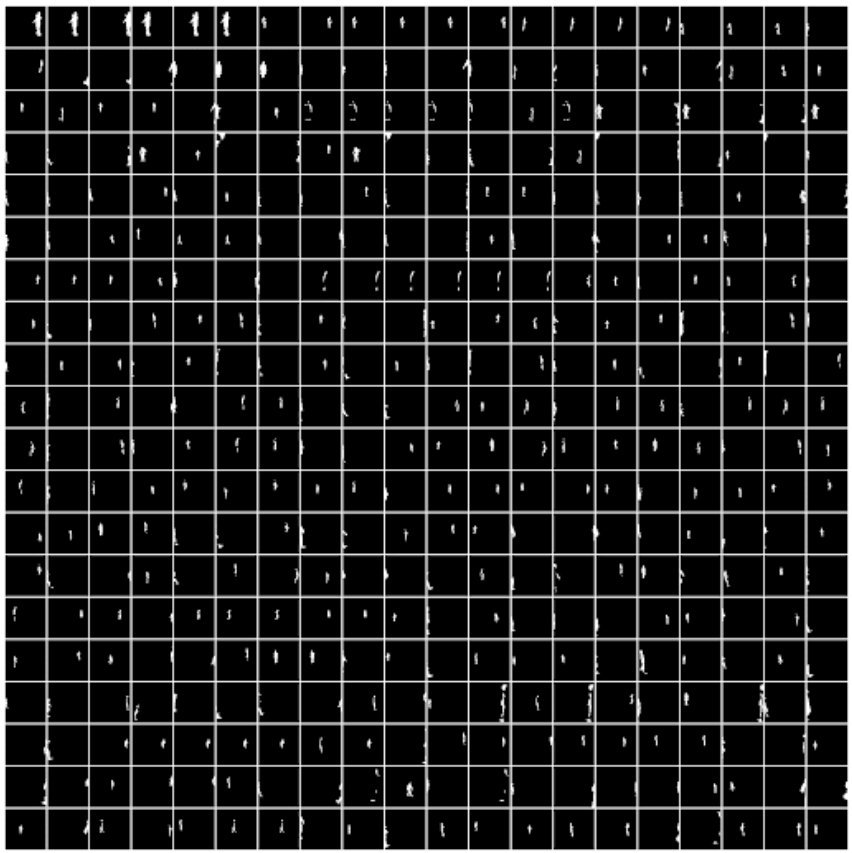

**Issue:** $P(\mathbf{z} \leq \boldsymbol{u}) = P(f(\mathbf{z}) \leq \boldsymbol{u}), \qquad \frac{\partial f_i(\mathbf{z})}{\partial \mathbf{z}_j} \neq 0$

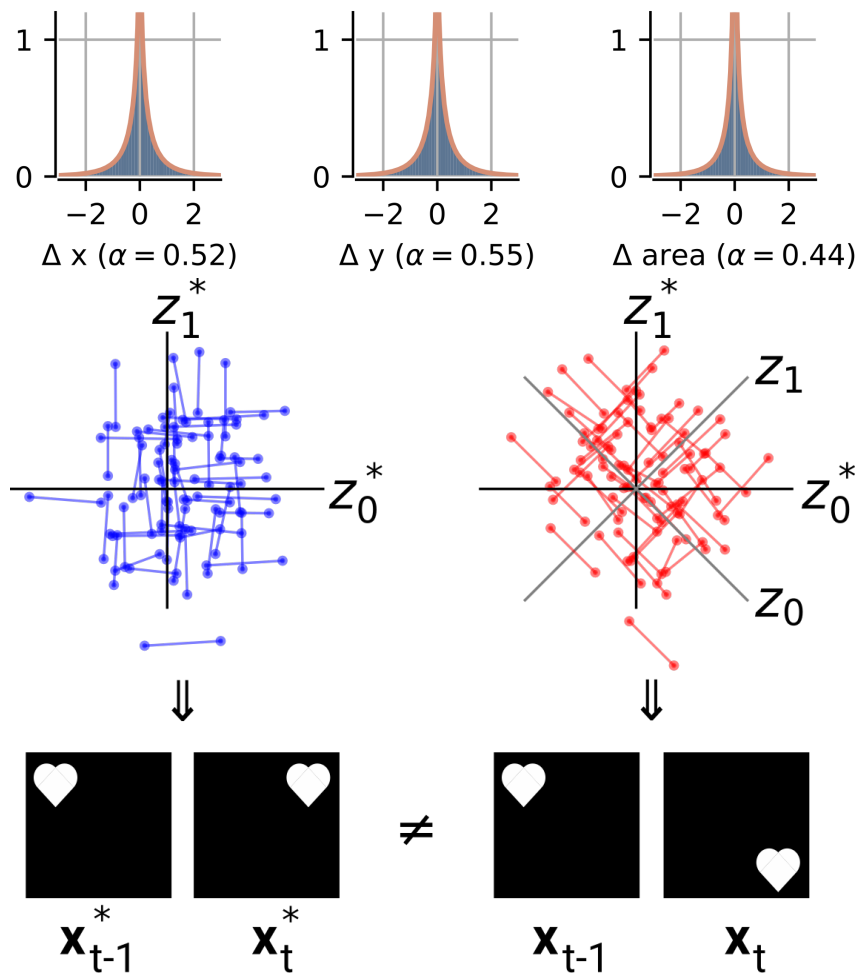
*(Locatello, 2020)*

Earlier work proved similar results *(Hyvarinen & Pajunen, 1999)*.

# Identifiability: Additional Assumptions

Learn a generative model for which both the marginal $p_{\mathbf{z}}$ **and** the conditional $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ distributions match ground-truth.

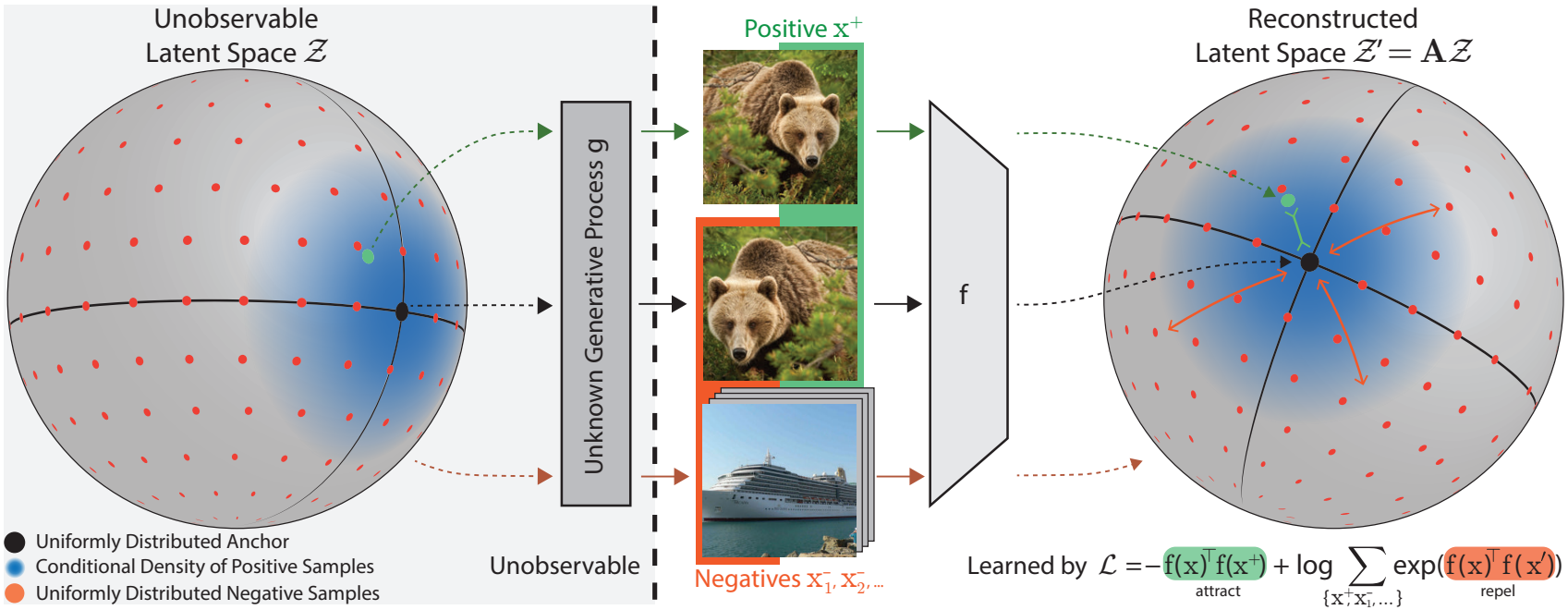





*(Klindt*, Schott*, Sharma* et al., 2021)*

# Identifiability: Additional Assumptions

Can we obtain a similar result if we solely learn an encoder via contrastive learning?




*(Zimmermann*, Sharma*, Schneider* et al., 2021)*

# Identifiability: Additional Assumptions

Can we obtain a similar result if we solely learn an encoder via contrastive learning?

$\Longrightarrow$ **Yes, if** the distribution encoded in $\mathcal{L}_{\text{InfoNCE}}(\mathbf{g})$ matches that of ground-truth

**Theorem 1** ($\mathcal{L}_{\text{contr}}$ converges to the cross-entropy between latent distributions). *If the ground-truth marginal distribution $p$ is uniform, then for fixed $\tau > 0$, as the number of negative samples $M \to \infty$, the (normalized) contrastive loss converges to*

$$\lim_{M\to\infty} \mathcal{L}_{\text{contr}}(f;\tau,M) - \log M + \log|\mathcal{Z}| = \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))\right] \tag{14}$$

*where $H$ is the cross-entropy between the ground-truth conditional distribution $p$ over positive pairs and a conditional distribution $q_{\text{h}}$ parameterized by the model $f$, and $C_h(\mathbf{z}) \in \mathbb{R}^+$ is the partition function of $q_{\text{h}}$ (see Appendix A.1.1):*

$$q_{\text{h}}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\mathbf{z})^{-1} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \quad \text{with} \quad C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}\, \mathrm{d}\tilde{\mathbf{z}}. \tag{15}$$

**Theorem 2.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, the ground-truth marginal be uniform, and the conditional a vMF distribution (cf. Eq. 2). Let the restriction of the mixing function $g$ to $\mathcal{Z}$ be injective and $h$ be differentiable in a vicinity of $\mathcal{Z}$. If the assumed form of $q_{\text{h}}$, as defined above, matches that of $p$, and if $f$ is differentiable and minimizes the CL loss as defined in Eq. (1), then for fixed $\tau > 0$ and $M \to \infty$, $h = f \circ g$ is linear, i.e., $f$ recovers the latent sources up to an orthogonal linear transformation and a constant scaling factor.*

**Theorem 6.** *Let $\mathcal{Z}$ be a convex body in $\mathbb{R}^N$, $h : \mathcal{Z} \to \mathcal{Z}$, and $\delta$ be an $L^\alpha$ metric or semi-metric (cf. Lemma 1 in Appx. A.2.4) for $\alpha \geq 1, \alpha \neq 2$. Further, let the ground-truth marginal distribution be uniform and the conditional distribution be as Eq. (5), and let the mixing function $g$ be differentiable and invertible. If the assumed form of $q_{\text{h}}(\cdot|\mathbf{z})$ matches that of $p(\cdot|\mathbf{z})$, i.e., both use the same metric $\delta$ up to a constant scaling factor, and if $f$ is differentiable and minimizes the $\mathcal{L}_{\delta\text{-contr}}$ objective in Eq. (6) for $M \to \infty$, we find that $h = f \circ g$ is a composition of input independent permutations, sign flips and rescaling.*

*(Zimmermann*, Sharma*, Schneider* et al., 2021)*

# How well do our assumptions match practice? (e.g., SimCLR)

Prior work on identifiable representation learning and multi-view nonlinear ICA *(e.g., Hyvarinen & Morioka, 2016, 2017; Gresele et al., 2019; Locatello et al., 2020; Klindt et al., 2020; Zimmermann et al., 2021)*

1. **Assume factorized $p_{\mathbf{z}}$, i.e. independent latents**
   (unknown statistical/causal dependence between latents of interest)
2. **Assume that all latents $z_i$ change across views** $(\mathbf{x}, \tilde{\mathbf{x}})$
   (augmentations chosen to leave certain aspects invariant)

Can we perform identifiability analysis without these assumptions?

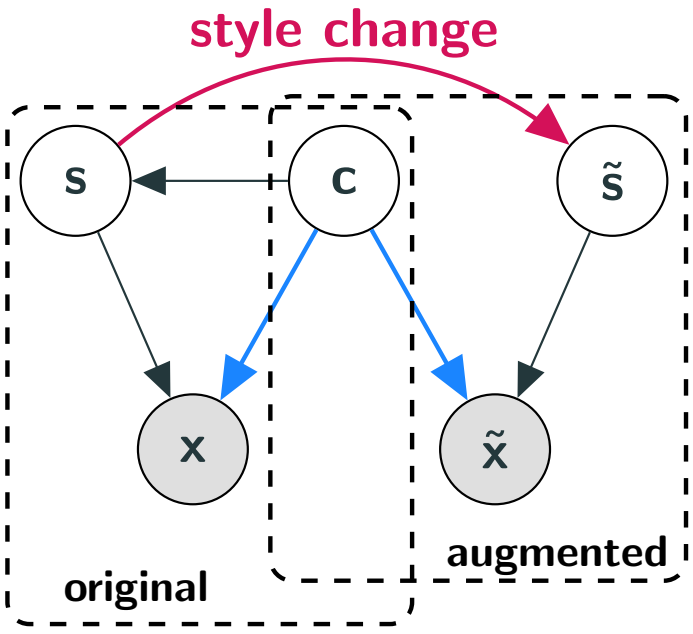
$\Longrightarrow$ **Yes!** *(von Kügelgen$^*$, Sharma$^*$, Gresele$^*$ et al., 2021)*

# Problem Setting II: Content-Style Partition

Partition the latent representation $\mathbf{z}$ into two disjoint blocks:

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

- **invariant content block c is always shared across** $(\mathbf{x}, \tilde{\mathbf{x}})$
- **varying style block s may change across** $(\mathbf{x}, \tilde{\mathbf{x}})$




**Note:**

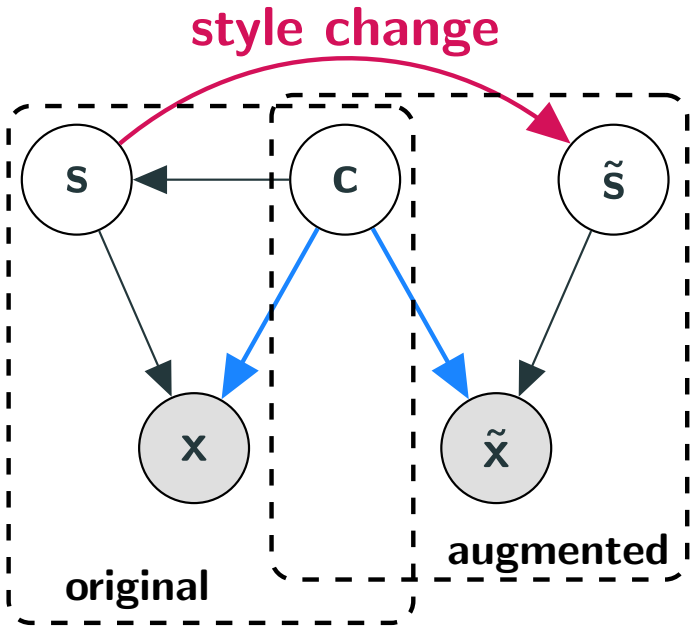
*In practice*, the content-style partition is **implicitly defined** by the set $\mathcal{T}$ of transformations $\mathbf{t} \sim p_{\mathbf{t}}$ used for data augmentation.

# Problem Setting III: Main Assumptions

## Assumption 1: Content Invariance

The conditional $p_{\tilde{\mathbf{z}}|\mathbf{z}}$ takes the form

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




## Assumption 2: Random Style Subsets Change

Let $p_A$ be a distribution over style subsets $A \subseteq \{1, ..., n_s\}$.
Then, the style conditional $p_{\tilde{\mathbf{s}}|\mathbf{s}}$ is given by

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

# Theorem

## Theorem: Identifying Content with a Non-Invertible Encoder

Assume the same data generating process and conditions.
Let $\mathbf{g} : \mathcal{X} \to (0,1)^{n_c}$ be any smooth function that minimises

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

where $H(\cdot)$ denotes differential entropy.
Then $\mathbf{g}$ block-identifies the true content variables.

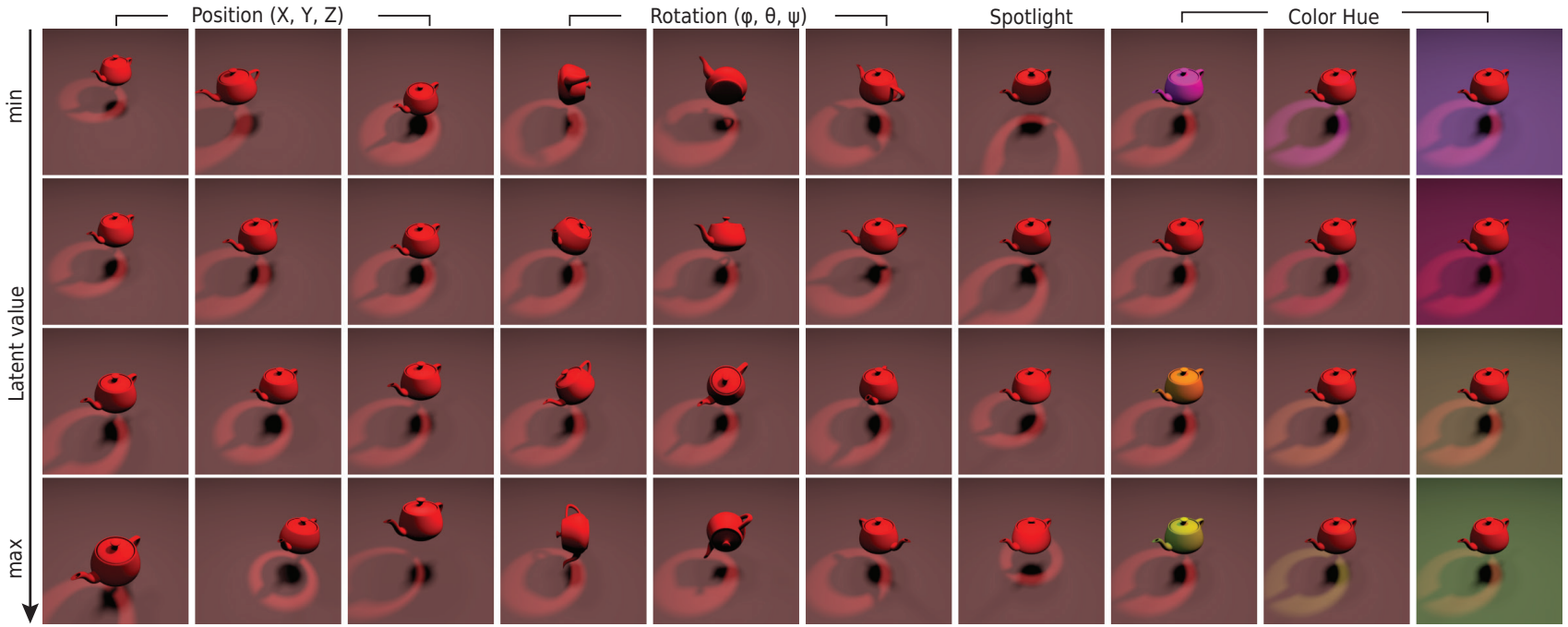
## Definition: Block-Identifiability

The true content $\mathbf{c} = \mathbf{f}^{-1}(\mathbf{x})_{1:n_c}$ is *block-identified* by $\mathbf{g} : \mathcal{X} \to \mathcal{Z}$ if the inferred content $\hat{\mathbf{c}} = \mathbf{g}(\mathbf{x})_{1:n_c}$ contains *all* and *only* information about $\mathbf{c}$, i.e., if $\exists$ *invertible* $\mathbf{h} : \mathbb{R}^{n_c} \to \mathbb{R}^{n_c}$ s.t. $\hat{\mathbf{c}} = \mathbf{h}(\mathbf{c})$.

## Main Result

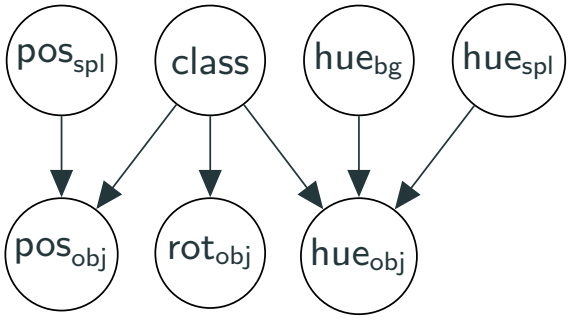
**CL with InfoNCE (e.g., SimCLR) asymptotically isolates content.**

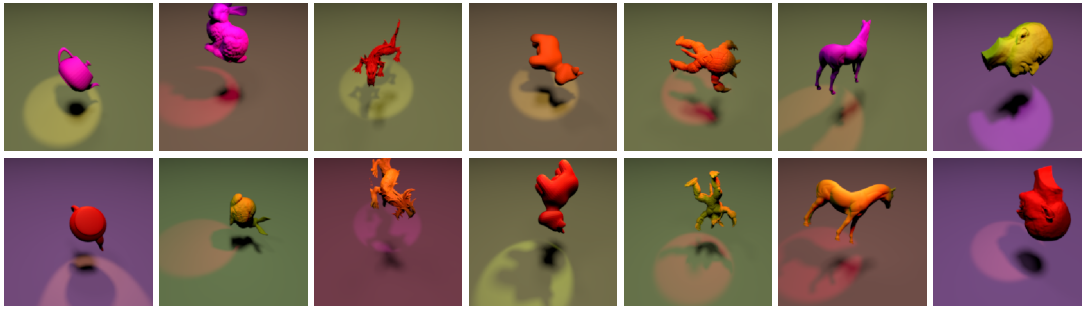
# 3DIdent (Zimmermann et al., 2021)



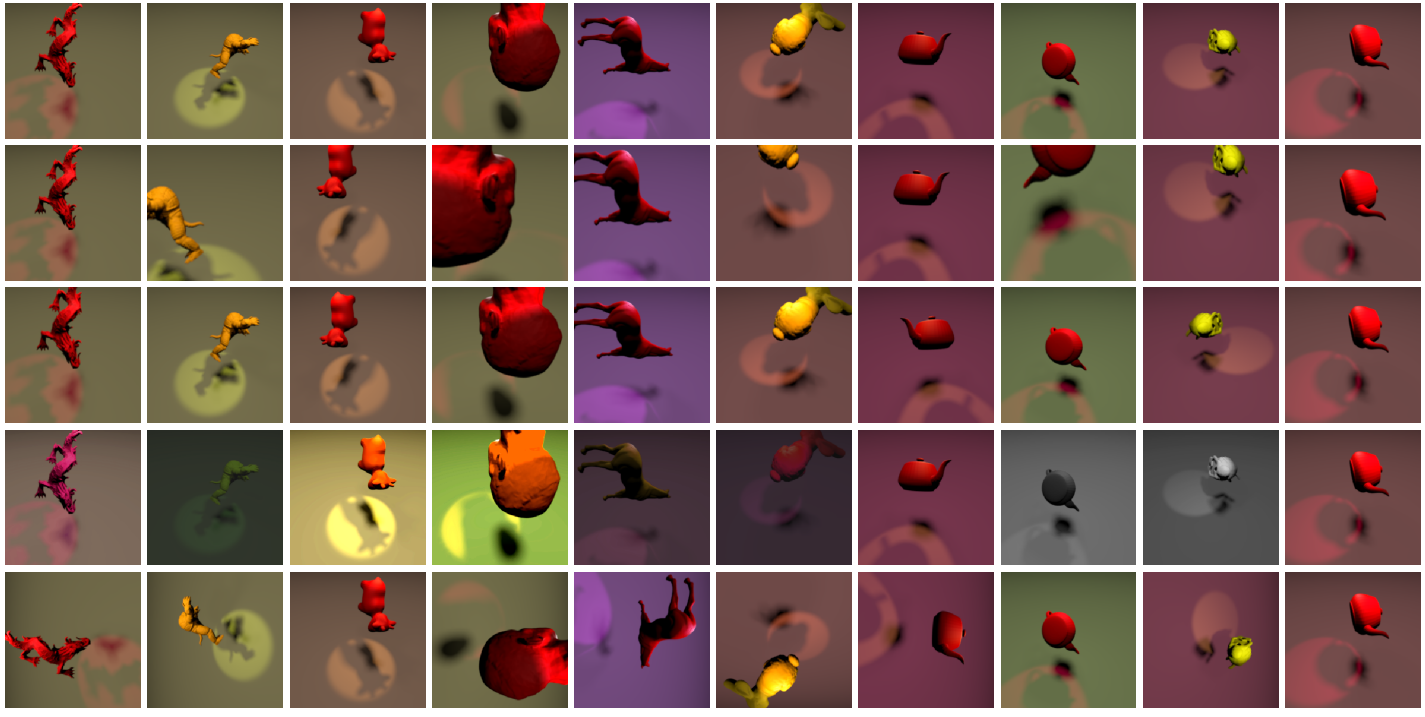
# Causal3DIdent

## Dataset:

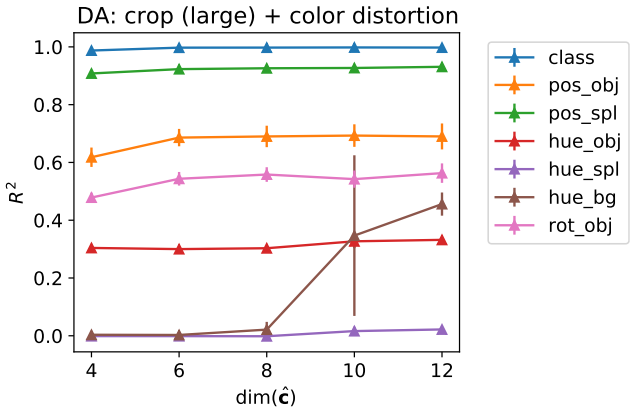


## Augmentations:

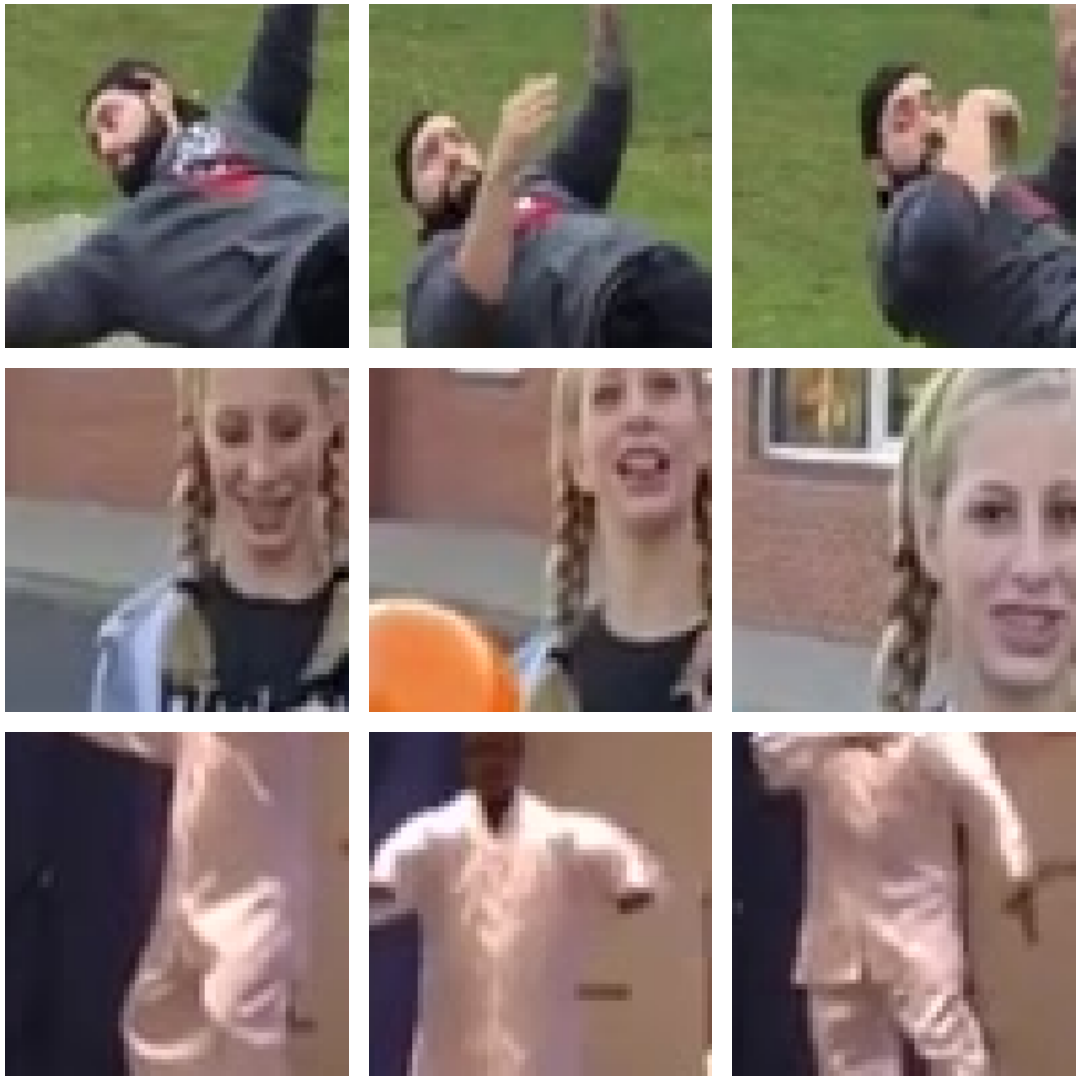


# SimCLR Results




| Views generated by | Class | Positions | | Hues | | | Rotations |
|---|---|---|---|---|---|---|---|
| | | object | spotlight | object | spotlight | background | |
| DA: colour distortion | 0.42 ± 0.01 | **0.61** ± 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 | **1.00** ± 0.00 | **0.59** ± 0.33 | **0.91** ± 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 | **0.87** ± 0.00 | 0.09 ± 0.02 | **1.00** ± 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 | **1.00** ± 0.00 | 0.00 ± 0.00 |
| LT: change positions | **1.00** ± 0.00 | 0.16 ± 0.23 | 0.00 ± 0.01 | 0.46 ± 0.02 | 0.00 ± 0.00 | **0.97** ± 0.00 | 0.29 ± 0.01 |
| DA: crop (large) + colour distortion | **0.97** ± 0.00 | **0.59** ± 0.07 | **0.59** ± 0.05 | 0.28 ± 0.00 | 0.01 ± 0.01 | 0.01 ± 0.00 | **0.74** ± 0.03 |
| DA: crop (small) + colour distortion | **1.00** ± 0.00 | **0.69** ± 0.04 | **0.93** ± 0.00 | 0.30 ± 0.01 | 0.00 ± 0.00 | 0.02 ± 0.03 | **0.56** ± 0.03 |
| LT: change positions + hues | **1.00** ± 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 | **0.83** ± 0.01 | 0.30 ± 0.12 | **0.99** ± 0.00 | 0.05 ± 0.03 |
| LT: change rotations | **1.00** ± 0.00 | **0.53** ± 0.33 | **0.90** ± 0.00 | 0.41 ± 0.00 | 0.00 ± 0.00 | **0.97** ± 0.00 | 0.28 ± 0.00 |
| DA: rotation + colour distortion | **0.59** ± 0.01 | **0.58** ± 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 | **1.00** ± 0.00 | **0.57** ± 0.34 | **0.91** ± 0.00 | 0.30 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.28 ± 0.00 |

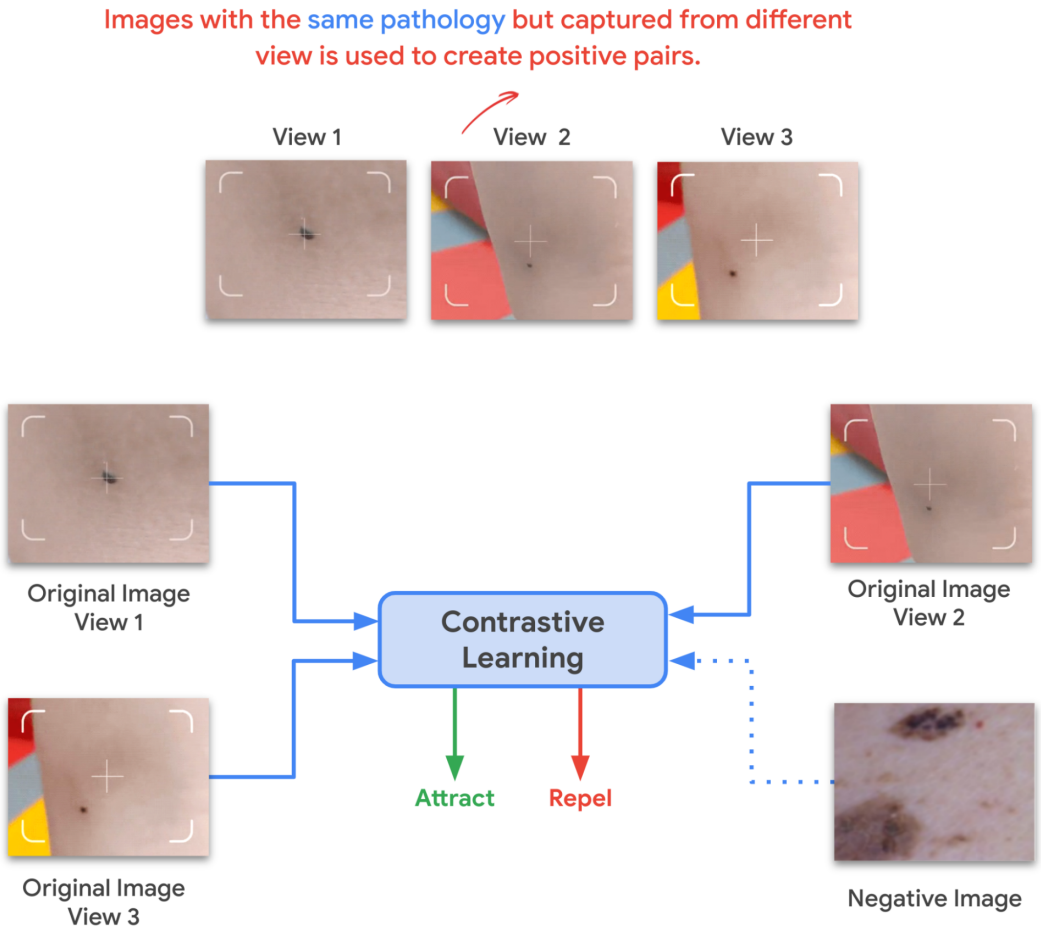
# Video?



*(Sharma et al., 2022)*

# Take-home Message

Given the practical considerations of medical applications, can we move beyond augmentations?




*(Azizi et al., 2021)*