Stable and Causal Inference for Discriminative Self-supervised Deep Visual Representations

Yuewei Yang, Hai Li, Yiran Chen

Duke University

Durham, USA

yuewei.yang@duke.edu, hai.li@duke.edu, yiran.chen@duke.edu

Abstract

In recent years, discriminative self-supervised methods have made significant strides in advancing various visual tasks. The central idea of learning a data encoder that is robust to data distortions/augmentations is straightforward yet highly effective. Although many studies have demonstrated the empirical success of various learning methods, the resulting learned representations can exhibit instability and hinder downstream performance. In this study, we analyze discriminative self-supervised methods from a causal perspective to explain these unstable behaviors and propose solutions to overcome them. Our approach draws inspiration from prior works that empirically demonstrate the ability of discriminative self-supervised methods to demix ground truth causal sources to some extent. Unlike previous work on causality-empowered representation learning, we do not apply our solutions during the training process but rather during the inference process to improve time efficiency. Through experiments on both controlled image datasets and realistic image datasets, we show that our proposed solutions, which involve tempering a linear transformation with controlled synthetic data, are effective in addressing these issues.

1. Introduction

Learning generalized representation with unlabeled data is a challenging task in various fields, but Self-Supervised Learning (**SSL**) has recently demonstrated remarkable success in learning semantic invariant representations without labels [40, 41, 53]. There are two main types of self-supervised learning (SSL) based on the pretext task used: generative and discriminative SSL, with generative SSL reconstructing altered or distorted data to its original input [9, 28, 31, 59, 65, 71]and early discriminative SSL predicting easily designed labels and task-specific representations that are not very generalizable [25, 57, 75]. More recent
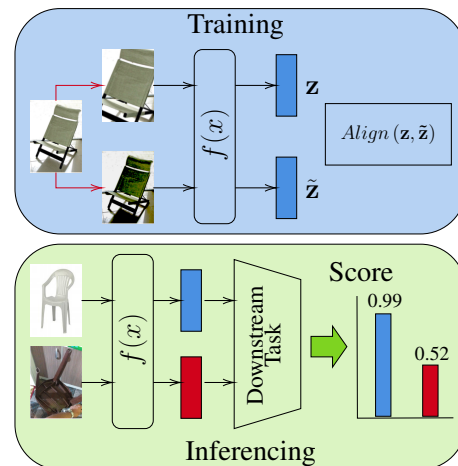

Figure 1: During the training of discriminative SSL, aligning positive representations will be robust to the changes applied as augmentations (red arrows). However, during the inference stage, one small change in the data variable (such as *view angle*) will result in an unexpected degradation on downstream performance.

discriminative SSL trains the model to identify similarities and differences between pairs of augmented examples [7, 10, 11, 26, 29, 74]. The success of SSL in deep image models has resulted in progress in other data modalities [53, 52, 54, 61, 62] and attention-based models like transformers [12, 8, 49, 72]. Recent discriminative SSL aims to learn content and semantic invariant representations that are robust to data augmentations, but the learned representations can be *unstable* when one subtle factor of the data is changed to a value that is not accessible through all augmentations. To avoid the high cost of incorporating all possible subtle changes during training, insights are needed to uncover the root cause of instability and find a solution to prevent performance deterioration during inference. Figure 1 summarizes this deterioration effect.

Causality [60] is a vital tool to investigate the causal

relationships between variables in observational data, and can uncover the underlying causal factors that explain unexpected model behavior due to changes in the environment. Independent Component Analysis (**ICA**) is often used to disentangle sources in unsupervised training [36, 37, 39, 43, 45], and causal analysis has been applied in follow-up works [76, 67] to examine the empirical success of SSL under an ICA framework. However, while these works identify factors that contribute to SSL's success, they do not address the problem's unstable mode, which can cause a severe performance drop when underlying factors shift slightly to an unseen environment. Some works [56, 27, 48] attempt to incorporate causality during the training process to identify and alleviate the impact of such shifts, but this approach is time-costly and only marginally improves performance compared to non-causal SSL methods. A more time-efficient and accessible solution would be to simply reverse the unstable shift during inference.

We aim to address the issue of unstable behavior during the **inference stage** by building upon previous theories of successful InfoNCE-facilitated contrastive SSL and extending it to **all recent SSL methods** with additional assumptions and constraints. Drawing inspiration from the relationship between the ground truth positive pairs distribution and learned positive pairs distribution, we demonstrate that the approximated transformation between the ground truth representations and learned representations is orthogonal to the augmentations applied during training. However, a change in the data factor/variable that violates the conditions for successful SSL can cause a corresponding shift in the inferred representation, resulting in a decline in downstream performance. This change of data factor/variable can be a change in the background, texture, or view angles etc. To overcome this issue, we propose learning targeted transformations that regularize the violating shift and restore performance on the unseen data shift. This approach effectively avoids the undesirable behavior and improves performance on previously unseen data shifts.

To summarize, our contributions are following:

- Through the use of a comparable data generation process in prior research, we show that **All** current SSL techniques benefit from the alignment of positive pairs.

- Through our alternative derivation of the transformation matrix between the ground truth representation and the learned representation, we have shown that the augmentations applied during training are orthogonal to the resulting matrix.

- By interpreting a change in the data variable causally, we propose two solutions **focusing on inference** to modify the negative shift in representation space caused by such a change.

- We validate the proposed solutions by conducting experiments on both controlled and realistic datasets, providing evidence for their efficacy during the inference stage without retraining the pretrained models.

2. Related Work

Discriminative SSL learns invariant representations from positive pairs of unlabeled data samples, but previous attempts to use trivial labels like colors[75], rotations[25], and patch positions[57] offer minimal benefits for complex downstream tasks due to their easy augmentations. Recent discriminative SSL apply random augmentations to generate two views of an image sample and train an encoder to extract representations for maximizing similarity between the paired augmentations. SimCLR and MoCo[10, 29] are pioneer SSL works that maximize the cosine similariy between positive pairs and minimize the cosine similarity between negative pairs via optimizing the InfoNCE loss[58]. Immense resources are used to enforce a large number of negative samples since a larger number can tighten the upper bound of the mutual information between positive pairs[58]. Later advancement of discriminative SSL excludes the notion of negative pairs by only aligning positive pairs and preventing *representation collapse* through various regularizations. BYOL[26] predicts an Exponential-Moving-Averaged (EMA) representation of one view with a projected representation of another view. SimSiam[11] maximizes similarity between a projected representation and a detached representation of two positive samples. Unlike previous work focusing optimization on an instance level, Barlow Twins[74] encourages high similarity in corresponding feature dimensions and discourages redundancy across different feature dimensions between two views of a data sample. Detailed formulations are exhibited in A.

Mutual Information is a different perspective on the behaviour of discriminative SSL. Referred to InfoMax principle[51], the MI between different transformations of a data sample is maximized via optimizing the InfoNCE loss[2, 32, 44]. Though showing theoretical relation between optimizing InfoNCE and maximizing MI between positive pairs, the underlying factors instructing the behaviour of different SSL methods are not explored. Nonlinear ICA[38], on the other hand, captures complex data structures of SSL methods by disentangling underlying factors via minimizing the mutual information between learned representations and the original data input[17, 73, 20]. Other works associate the nonlinear ICA objectives with the contrastive SSL so the MI between positive pairs are maximized and negative pairs are minimized.

Other researchers have explored **causality and causal inference** as a means of understanding the success of discriminative SSL. Prior work has focused on partitioning the InfoNCE loss to *alignment* between the positive pairs and

uniformity between aggregations of all positive clusters[70]. By formulating a data generation process, [76] empirically explains that networks optimized via InfoNCE infer an orthogonal transformation of the ground truth latent representations. Furthermore [67] validates that that augmentations used in both generative and discriminative SSL isolate the *content* factor from the *style* factor. Our theory and work draw great inspirations from these two work. However, instead of solely focusing on InfoNCE-driven SSL and two factors, this work extends the framework to all recent discriminative SSL methods and identifies reasons for unstable circumstances analytically. We also propose methods to nullify the negative effects of unstable representations during inference.

Domain Adaptation is a strategy to bridge the gap between the model performance on a source domain and that on a target domain [4, 68]. Feature adaptation methods try to learn a new feature representation that is more invariant to the domain shift[66, 19, 15, 50], while instance adaptation methods try to reweight the importance of the labeled source examples to better align with the target domain[55, 6, 34]. Recent studies also implement a contrastive framework to learn a shared latent space between the source and target domains by maximizing the agreement between representations of corresponding samples while minimizing the agreement between representations of non-corresponding samples [69, 63, 42]. Unlike these works focusing on adapting to a target domain for better performance, we investigate the underlying causal factors for the performance gap and based on the findings we propose easy solutions to connect the gap.

3. Theory

In this section, we build on previous data generation process [76, 67] (3.1) and show that current SSL techniques benefit from the alignment of positive pairs (3.2). During a deeper dive into the generation process, we show the relation between the ground truth representations and the inferred representations, and this relation is an linear transformation matrix that is orthogonal to the augmentations applied during training (3.3). Finally, we disclose the causal reason for the unstable behaviour caused by a change in the data variable during the inference stage and propose analytical solutions to address this unstable issue (3.4).

3.1. Problem formulation

Data generation is assumed to be a generation function that takes ground truth latent representation as the input to generate an observation data/image.

Formally, we assume that the marginal distribution of sampling ground truth representations $\mathbf{z} \subseteq \mathcal{Z} \in \mathbb{R}^{d_1}$ w.r.t. a *class* is uniform on a unit sphere $\mathbb{S}^{d_1-1}$. An injective generation function $g(.) : \mathbb{R}^{d_1} \to \mathbb{R}^d$ takes a ground truth rep-
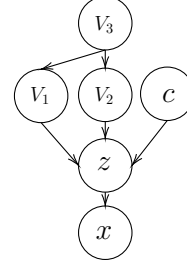

Figure 2: A causal graph for data generation process.

resentation and generate an observation sample $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{x} = g(\mathbf{z})$. Different variables, denoted as a set $V = \{V_i\}$ and *class/content* variable c, constitute the values of each dimension in the ground truth representation. These variables can be direct causes or confounding factors, $\mathbf{z} = (c, \{V_i\})$. Specific examples include *view angles*, *object size*, *background colors*, etc. A simple causal graph depicts this relationship is shown in Figure 2. So the general generation process is as following:

$$p(\mathbf{z} = (c, V_*)) \sim \frac{1}{|\mathcal{Z}|} \qquad \mathbf{x} = g(\mathbf{z}) \qquad (1)$$

To sample a positive example w.r.t the same *class*, we assume the conditional distribution is a von Mises-Fisher (vMF) distribution [21]:

$$p(\tilde{\mathbf{z}} = (c, \tilde{V}_*)|\mathbf{z} = (c, V_*)) = C_p^{-1} e^{\kappa_1 \mathbf{z}^\top \tilde{\mathbf{z}}} \qquad (2)$$

where $C_p = \int e^{\kappa_1 \mathbf{z}^\top \tilde{\mathbf{z}}} d\tilde{\mathbf{z}}$ is a normalization constant and κ is a concentration parameter.

Representation learning is a process that a feature encoder, $f(.) : \mathbb{R}^d \to \mathbb{R}^{d_2}$, extracts representations from two positive observations $f(\tilde{\mathbf{x}}) = f \circ g(\tilde{\mathbf{z}})$, $f(\mathbf{x}) = f \circ g(\mathbf{z})$ and a distribution associated with the encoder f through $h = f \circ g$ is:

$$q_h(\tilde{\mathbf{z}}|\mathbf{z}) = C_q^{-1}(\mathbf{z}) e^{\kappa_2 h(\tilde{\mathbf{z}})^\top h(\mathbf{z})} \qquad (3)$$

with $C_q(\mathbf{z}) = \int e^{\kappa_2 h(\tilde{\mathbf{z}})^\top h(\mathbf{z})} d\tilde{\mathbf{z}}$ be the normalization term.[1] Optimizing any discriminative SSL objective will maximize the similarity between these positive pairs. An example of a well constructed objective is the InfoNCE loss:

$$\mathcal{L}_{infoNCE} = \mathbb{E}[-log \frac{e^{f(\tilde{\mathbf{x}})^\top f(\mathbf{x})/\tau}}{e^{f(\tilde{\mathbf{x}})^\top f(\mathbf{x})/\tau} + \sum_{i=1}^{K} e^{f(\mathbf{x}_i^-)^\top f(\mathbf{x})/\tau}}] \qquad (4)$$

where $\tilde{\mathbf{x}}$ is a positive example w.r.t. $\mathbf{x}$ in the obsevation space and $\{\mathbf{x}_i^-\}_1^K$ are K samples from distributions of all observations. The global minimum of (4) is reached when

---

[1]The mapping of representations on a hypersphere may be different to Barlow Twins methodology, but as shown in [74], normalize representations on a unit sphere also works under Barlow Twins.

the cosine similarity between positive pairs is maximized and the cosine similarity between all negative pairs is minimized. In the following section, we will show that discriminative SSL including InfoNCE-driven and non InfoNCE-driven (EMA and Siamese with a predictor) follows strict rules of alignment to maximize the similarity.

3.2. Alignment in discriminative SSL

In this section we will combine theories stated in [76, 67, 70] so that the general factors of successful discriminative SSL can be summarized. Additionally, instead of focusing on just InfoNCE SSL variation and content block-identifiability [35, 47], we extend the combined theory to demonstrate that all discriminative SSL benefit from *alignment* between positive representations.

Theorem 3.1 *With a data generation process described in 3.1, all discriminative SSL objectives have an alignment loss function between positive pairs from the network:*

$$\mathcal{L}_{align} = \|f(\mathbf{x}) - f(\tilde{\mathbf{x}})\|_2^2 \tag{5}$$

This is a weaker statement than **Theorem 4.4** in [67] since we only focus on the alignment term. For analysis on the regularization term on the network entropy, see B.

Proof. For InfoNCE-driven SSL (SimCLR and MoCo), as derived in [68, 76], the InfoNCE loss converges to an *alignment* term and a *uniformity* term as the number of negative samples approaches infinity. (See B for full details).

For EMA-based SSL methods (BYOL), a predictor p is associated with the online network so the SSL objective becomes $\mathcal{L} = \mathbb{E}_{(\mathbf{x}, \tilde{\mathbf{x}})} \|p \circ f(\mathbf{x}, \theta) - f(\tilde{\mathbf{x}}, \xi)\|_2^2$, where $\xi^t = \alpha \xi^{t-1} + (1 - \alpha)\theta^t$ is target network parameter and θ is the online network parameter. Denote $p' = p \circ f$. By adding and subtracting $p'(\tilde{\mathbf{x}}, \theta)$ we derive:

$$
\begin{aligned}
\mathcal{L} &= \mathop{\mathbb{E}}_{(\mathbf{x}, \tilde{\mathbf{x}})} \|p'(\mathbf{x}, \theta) - p'(\tilde{\mathbf{x}}, \theta) + p'(\tilde{\mathbf{x}}, \theta) - f(\tilde{\mathbf{x}}, \xi)\|_2^2 \\
&= \mathop{\mathbb{E}}_{(\mathbf{x}, \tilde{\mathbf{x}})} \|p'(\mathbf{x}, \theta) - p'(\tilde{\mathbf{x}}, \theta)\|_2^2 + \mathop{\mathbb{E}}_{\tilde{\mathbf{x}}} \|p'(\tilde{\mathbf{x}}, \theta) - f(\tilde{\mathbf{x}}, \xi)\|_2^2 \\
&\quad - 2 \mathop{\mathbb{E}}_{(\mathbf{x}, \tilde{\mathbf{x}})} [(p'(\tilde{\mathbf{x}}, \theta) - p'(\mathbf{x}, \theta))^\top (p'(\tilde{\mathbf{x}}, \theta) - f(\tilde{\mathbf{x}}, \xi))]
\end{aligned}
\tag{6}
$$

Since $f(\mathbf{x}, \theta)$ and $f(\tilde{\mathbf{x}}, \xi)$ maps in the same space $\mathbb{R}^{d_2}$, p can be considered as a bijective linear transformation within $\mathbb{R}^{b_2}$. In fact, the performance difference between a linear predictor and non-linear predictor is subtle. The global minimizer of (6) must align network output w.r.t to $(\mathbf{x}, \tilde{\mathbf{x}})$ with the first term in (6) and align outputs from different networks with second term in (6). Hence completes the proof.

For Siamese network with a predictor (SimSiam), a similar approach can refomulate the objective as it is a special case when $f(\tilde{\mathbf{x}}, \xi) = f(\tilde{\mathbf{x}}, \theta)$ and a `stop-gradient` is

applied on the $f(\tilde{\mathbf{x}}, \theta)$. Hence by substuting $\mathrm{sg}(f(\tilde{\mathbf{x}}, \theta))$ with $f(\tilde{\mathbf{x}}, \xi)$ in (6), we complete the proof by:

$$
\begin{aligned}
\mathcal{L} &= \mathop{\mathbb{E}}_{(\mathbf{x}, \tilde{\mathbf{x}})} \|p'(\mathbf{x}, \theta) - p'(\tilde{\mathbf{x}}, \theta)\|_2^2 \\
&\quad + \mathop{\mathbb{E}}_{\tilde{\mathbf{x}}} \|p'(\tilde{\mathbf{x}}, \theta) - \mathrm{sg}(f(\tilde{\mathbf{x}}, \theta))\|_2^2 \\
&- 2 \mathop{\mathbb{E}}_{(\mathbf{x}, \tilde{\mathbf{x}})} [(p'(\tilde{\mathbf{x}}, \theta) - p'(\mathbf{x}, \theta))^\top (p'(\tilde{\mathbf{x}}, \theta) - \mathrm{sg}(f(\tilde{\mathbf{x}}, \theta)))]
\end{aligned}
\tag{7}
$$

Note that the third term in (6) and (7) can be formulated by the differential entropy of $H(p'(.))$ hence prevent representation collapse.

For Barlow Twins, the diagonal of the cross-correlation matrix C_{ii} is the cosine similarity between positive pairs. Hence completes the proof by:

$$
\sum_i^{d_2} (1 - C_{ii})^2 = \sum_i (1 - (f(\mathbf{x})^\top f(\tilde{\mathbf{x}}))_i / (d_2 - 1))^2 \tag{8}
$$

$$
= \sum_i (\|f(\mathbf{x})_i - f(\tilde{\mathbf{x}})_i\|_2^2 / (2 * (d_2 - 1)))^2 \tag{9}
$$

3.3. Transformation of the ground-truth factors is orthogonal to the applied augmentations

By demonstrating that all discriminative SSL have a alignment loss term, the transformation between the ground truth representations and the inferred representations can be derived as in [76]. But different to [76], our derivation of minimization of cross entropy is assumed to be a lower bound since we only include the alignment term and the uniformity term is always positive [14, 5]. However, with all SSL objectives there are additional terms to maximize the output entropy of the model (some described in 3.2). So optimizing SSL objectives as a complete loss function will minimize the cross entropy $\mathbb{E}[H(p(.|\mathbf{z}), q_h(.|\mathbf{z}))]$.

Theorem 3.2 *By considering the generation conditional distribution as $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1} e^{\kappa_1 \mathbf{z}^\top \tilde{\mathbf{z}}}$, the inferred conditional distribution $q_h(\tilde{\mathbf{z}}|\mathbf{z})$ can match $p(\tilde{\mathbf{z}}|\mathbf{z})$ by minimizing $\|h(\mathbf{z}) - h(\tilde{\mathbf{z}})\|_2^2$ and $\forall \mathbf{z}, \tilde{\mathbf{z}} : \kappa \mathbf{z}^\top \tilde{\mathbf{z}} = h(\mathbf{z})^\top \tilde{\mathbf{z}}$ with $h = f \circ g$ or $h = p \circ f \circ g$ maps onto a hypersphere with radius $\sqrt{\kappa_1 / \kappa_2}$.*

The proof exactly follows **Proposition 1** in [76] just with minor modification on the concentration term κ and h so that **Theorem** 3.2 can apply to non-InfoNCE SSL. A global minimizer of the alignment term $\|h(\mathbf{z}) - h(\tilde{\mathbf{z}})\|_2^2$ will also minimize the cross entropy of between $p(\tilde{\mathbf{z}}|\mathbf{z})$ in (2) and $q_h(\tilde{\mathbf{z}}|\mathbf{z})$ in (3). This indicates that minimizers of SSL objective alignment maintain the dot product. Then we can use **Proposition 2** in [76] directly to show that h is an orthogonal linear transformation.

Theorem 3.3 *Assume the data generation process (cf. 3.1), a model parameterized by $h = f \circ g$ or $h = p \circ f \circ g$ ($h : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$) that minimizes the alignment term in all discriminative SSL objectives: $\|h(\mathbf{z}) - h(\tilde{\mathbf{z}})\|_2^2$ as (15), h is an orthogonal transformation: $h(\tilde{\mathbf{z}}) = \mathbf{A}\tilde{\mathbf{z}}$ where $\mathbf{A}$ is an orthogonal matrix.*

The proof follows that a function h minimizes the alignment will also minimize the cross entropy between the ground truth conditional distribution $p(\tilde{\mathbf{z}}|\mathbf{z})$ and the inferred conditional distribution $q_h(\tilde{\mathbf{z}}|\mathbf{z})$. Therefore if h is isometric w.r.t the dot product as indicated in **Theorem** 3.2 then $h(\tilde{\mathbf{z}}) = \mathbf{A}\tilde{\mathbf{z}}$ according to **Proposition 2** in [76].

Theorem 3.4 *Assume augmentations applied during the training can be represented as a change in the ground truth representations, i.e. $\tilde{\mathbf{x}} = g(\tilde{\mathbf{z}}(c, \tilde{V}*))$ and change in data variables induces a shift in the ground truth representations $\delta\mathbf{z} = (c, \tilde{V}_*) - (c, V_*)$, then $\delta\mathbf{z}$ is in orthogonal to $\mathbf{A}$ i.e. $\mathbf{A}\delta\mathbf{z} = 0$ if a discriminative SSL objective is optimized.*

Proof. With augmentations normally applied during SSL such as color distortions, rotations, random cropping, and etc., one can view the alteration as a change of a variable in the data variable (color, view angles, sizes). This change of variable under the generation framework described in 3.1 will result in a change in $V = \{V_i\}$ since c is not changed. Regardless of structure of the causal graph shown in Figure 2, the change in V can be reflected in the ground truth representation space as $\tilde{\mathbf{z}} = \mathbf{z} + \delta\mathbf{z}$. A global minimizer of any discriminative SSL objective will minimize the alignment term $h(\mathbf{z}) = h(\tilde{\mathbf{z}}) = h(\mathbf{z}+\delta\mathbf{z})$. According to **Theorem** 3.3 we derive:

$$h(\mathbf{z}) = \mathbf{A}\mathbf{z} = \mathbf{A}(\mathbf{z} + \delta\mathbf{z}) \tag{10}$$

$$\mathbf{A}\delta\mathbf{z} = 0 \tag{11}$$

with $\mathbf{A}$ being a linear orthogonal matrix. This indicates that the transformation A is learned to annul the change in data variables or the effect of augmentations applied.

3.4. Reason and solutions for unstable change in data variables

In section 3.3 we show that generalized representation is robust to augmentations since $\mathbf{A}$, the linear transformation between the ground truth representation and inferred representations, is orthogonal to augmentation applied during the training. And if a change in the data variable reflects a shift, $\delta\mathbf{z}$, in the ground truth representations and the newly inferred representation is stable, meaning $\delta\mathbf{z}$ will be absorbed by the transformation matrix $\mathbf{A}$, then the shift in the ground truth representation corresponds to an augmentation that is applied during the training. However, when $\delta\mathbf{z}$ appears in the range of $\mathbf{A}$ and $\mathbf{A}\delta\mathbf{z} \neq 0$, the resultant inferred representation can be unstable and lead to performance drop

on downstream models (denote as $D(\mathbf{z})$). We quantify this deterioration on D by:

$$m(D(h(\mathbf{z}))_{stable}) - m(D(h(\tilde{\mathbf{z}})_{unstable}) \tag{12}$$

where $m(.)$ is a metric on an outcome of the downstream task. In example of D being a linear classifier, $m(.)$ can be the probability of predicting the target class (**prediction score**) or the proportion of correct predictions (**accuracy**). In order to overcome this deterioration, we propose two methods, namely **Robust Dimensions** and **Stable Inference Mapping**:

1. **Robust Dimensions**: Under a stable condition, $D(f(\mathbf{x}))^2 = D(\mathbf{Az})$, each dimension in the inferred representation $f(\mathbf{x})$ is a linear combination of dimensions of the ground truth representation. The dimensions contributing most to $D()$ should be robust to unstable shift $\delta\mathbf{z}_{unstable}$. In other words, most robust dimensions of $f(\mathbf{x})$ should be also robust in $f(\tilde{\mathbf{x}})$ where $\tilde{\mathbf{x}} = g(\mathbf{z} + \delta\mathbf{z}_{unstable})$ as some dimensions of $\mathbf{A}\delta\mathbf{z}_{unstable}$ will be zero. Hence identifying most important dimensions in $f(\mathbf{x})_{stable}$ and pass through the same dimensions of $f(\tilde{\mathbf{x}})_{unstable}$ should alleviate the deterioration by making $m(D(h(\mathbf{z}))_{stable})_{\{dim\}} = m(D(h(\tilde{\mathbf{z}})_{unstable})_{\{dim\}}$ where $\{dim\}$ is a set of most robust dimensions. In example of D being a linear classifier, the contribution of each dimension can be calculated by $W_c^\top f(\mathbf{x})$ where $W_c^\top$ is the Jacobian of the linear classifier w.r.t target class c.

2. **Stable Inference Mapping**: Since $f(\mathbf{x})_{stable} - f(\tilde{\mathbf{x}})_{unstable} = -\mathbf{A}\delta\mathbf{z}$, we can learn another linear transformation $\mathbf{F}$ to absorb $\delta\mathbf{z}$. Especially, we want to learn $\mathbf{F}f(\tilde{\mathbf{z}})_{unstable} = \mathbf{FAz} + \mathbf{FA}\delta\mathbf{z}$ such that the additional $\mathbf{F}$ will not only set $\mathbf{A}\delta z$ to 0, but also make stable representation more robust to the unstable shift by assuring $\mathbf{FA}$ orthogonal to $\delta\mathbf{z}_{unstable}$ in addition to the augmentations applied during training. Formally, we model a linear layer $l(f(\tilde{\mathbf{x}})_{unstable}) = f(\mathbf{x})_{stable} - f(\tilde{\mathbf{x}})_{unstable}$, hence during the inference $f(\mathbf{x}) + l(f(\mathbf{x}))$ is used for downstream task.

Relation to Causal Inference Since we only have observations in the image space, we consider the augmentations or changes of a data variable as *interventions* on the ground truth representations: $\tilde{\mathbf{x}} = g(\tilde{\mathbf{z}}|(c, do(V_i = v_i))$. With an access to ground truth representation, we can evaluate treatment-control effect i.e. $Pr(f(g(\tilde{\mathbf{z}}|(c, do(V_i = v_{stable})))) - Pr(f(g(\tilde{\mathbf{z}}|(c, do(V_i = v_{unstable}))))$. Without any access to the ground truth representations, we can evaluate the average treatment effect [33] by $\mathbb{E}[D(f(\mathbf{x})_{stable}) -$

---

[2]Note that in general f contains a projector. However we exclude the notion of the projector to simplify the problem and [10, 29, 26] show that a projector is not necessary in SSL.

$D(f(\tilde{\mathbf{x}})_{unstable})]$ via synthesizing manual data samples $\tilde{\mathbf{x}}_i = g(\tilde{\mathbf{z}}|(c, do(V_i = v_i))$.

4. Experiments and Discussions

In this section we evaluate our solutions to unstable shift in data variable on two datasets: Causal3DIdent and ImageNet pretrained SSL and corresponding linear classifiers as the downstream task.
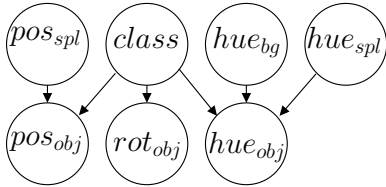
4.1. Causal3DIdent

Figure 3: Causal graph for V in Causal3DIdent.

[76] develops the dataset that 7 objects in the dataset that each 224×224 image is associated with an object class and a 10-dimensional latent representation. These 10 dimensions correspond to 10 data variables. A causal graph imposed on these variables is shown in Figure 3. As the dependency shown in Figure 3, an object is placed at $pos_{obj} = (x, y, z)$ with $rot_{obj} = (\phi, \theta, \psi)$ and $hue_o bj$ under a spot light at an angle pos_{spl} with color hue_{spl} on a background with color hue_{bg}. Detailed information in C.

To simulate unstable change of variable that is not accessible during the training, a range of some data variables is hold out and hence this portion of data is treated as potential unstable samples that causes the performance drop on the linear classifier. Since all data variables are truncated in range [-1,1], we hold out the edge value(s) to further portray 'unexpected' values during the inference time. We select 8 dimensions to intervene, namely: z in object position[3], all 3 object rotation angles, spot light position, and all 3 hue variables. Both training and testing data are sampled and for detailed sampling procedures refer to C.

SSL Experiment Setup ResNet18[30] is the backbone of the feature encoder f. Same augmentations in [10] are applied during training. An Adam optimizer with a learning rate at 0.0001 and a weight decay at 0.00001 is optimized for all discriminative SSL. Hyperparameters for each SSL is presented in C. The dimension of the inferred representation space is set to 128. The network is trained for 20 epochs on intervened data. Then a linear classifier is trained on the frozen representations of the network with a same optimizer for 10 epochs.

---

[3]Only z is altered because most of deep vision models are translation invariant.

To illustrate the simulation results in the deterioration of the downstream performance, for each testing data that has the same data variable distribution as the training data, we search in the ground truth representation space for 5 nearest representations in the hold-out test distribution when n dimension of the testing data is shifted to a random value in the hold-out range of corresponding data variable. Within the corresponding 5 images representing a shift in n dimensions that is not seen during the training, we record the lowest performance to fulfill the deterioration scenario. In Figure 4 the score and accuracy are averaged over all combinations of $\binom{8}{n}$. When $n = 0$, this indicates the performance of SSL on the testing data with the same data variable distribution as the training data. The performance is exceptional since hold-out values of some data variables may be covered by the augmentations applied (z covered by random cop, and hue colors covered by color distortions), $\mathbf{A}$ will be orthogonal to changes in these data variables. In fact in C we show that the performance difference between seen distribution and unseen distribution in testing data is comparatively small. As illustrated in Figure 4, even when only one variable is changed to a unseen value, there is a large drop in both accuracy and prediction score (20% in accuracy and 30% in prediction score for SimCLR). With more tangled changes in data variables, the unstable representations results in poorer downstream performance. Since the selection of the data variables may be too complex due to the dependency, we also validate the same issue on selecting only children nodes in Figure 3. And also we visualize the latent shift between stable and unstable examples. See C for more results.

Robust Dimensions For each pair of testing data $\mathbf{x}_{stable}$ with seen data variable distribution and the selected data $\mathbf{x}_{unstable}$ among the 5 nearest neighbours when changing n dimensions to unseen distribution values, we apply the Jacobian of the linear classifier w.r.t the target class $W_c^\top$ on the stable representation $f(\mathbf{x})_{stable}$ to identify the top k most important dimensions and pass the same dimensions of $f(\tilde{\mathbf{x}})_{unstable}$ to D to evaluate the performance. As shown in Figure 5, the accuracy only deteriorate slightly when top 90% most important features of $f(\tilde{\mathbf{x}})_{unstable}$ are selected for the downstream task. This is true even when all 8 variables are shifted. This suggests that $\mathbf{A}$ is orthogonal to changes in ground truth representation $\mathbf{z}$ in most dimensions. This high percentage of dimensions may be due to f optimizes the SSL objectives to a high level and the augmentations applied covers some of the hold-out variable values. Interestingly, there are cases where passing the top $k\%$ (around 40%) important dimensions results in higher performance than stable representations (unitervened in Figure 5). However, as expected, including more less important dimensions where the unseen shift in the data variable results in non-zero adjustment ($\mathbf{A}\delta\mathbf{z} \neq 0$) initiates the
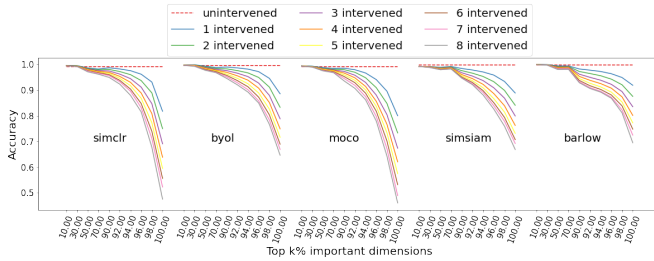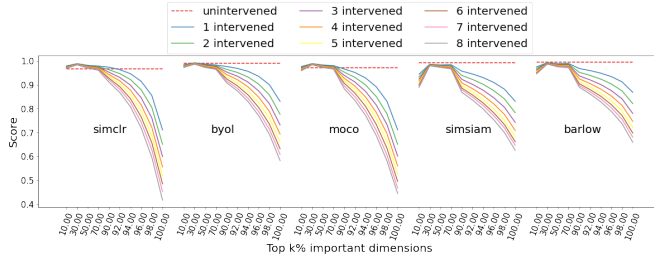
(a) Deterioration in Accuracy
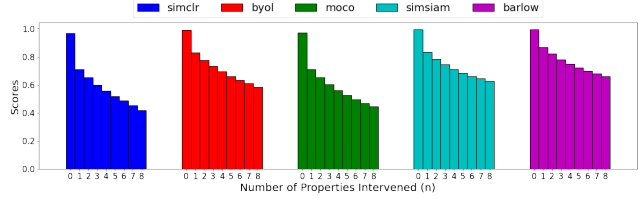
(b) Deterioration in Score

Figure 4: Deterioration of unstable changing in data variables on all SSL. The more unexpected changes occur in the data variables, the more severe the deterioration.

deterioration phenomenon we observe.

(a) Improvement on Accuracy

(b) Improvement on Score

Figure 5: By identifying and passing top k % important dimensions, the deterioration in both accuracy and prediction score is significantly alleviated.

Stable Inference Mapping As shown in 3.4, a linear transformation $\mathbf{F} : \mathbb{R}^{d_2} \rightarrow \mathbb{R}^{d_2}$ is trained to cancel the effect $\mathbf{A}\delta\mathbf{z}$ and also further improve the robustness of transformation on the ground truth representation $\mathbf{A}\mathbf{z}$. In this experiment, we match each training data with a random member in the 5 nearest representations when only **one** dimension is changed to a unstable value. A linear layer is trained with the same optimizer for 10 epochs on the training pairs. In Table 1, all accuracy for unstable examples increases significantly (except SimSiam) after learning $\mathbf{F}$. However, since the f is very close to the global minimizer of the alignment term, the improvement on stable examples cannot be observed.

	$\mathbf{x}_{stable}$		$\tilde{\mathbf{x}}_{unstable}$	
	w/o $\mathbf{F}$	w/ $\mathbf{F}$	w/o $\mathbf{F}$	w/ $\mathbf{F}$
SimCLR	0.996	0.998	0.833	0.889
MoCo	0.992	0.992	0.800	0.849
BYOL	0.996	0.998	0.886	0.920
SimSiam	0.998	0.999	0.919	0.928
Barlow Twins	0.991	0.995	0.818	0.862

Table 1: The effect of $\mathbf{F}$ on both stable and unstable samples. The accuracy is average over 3 random seeds.
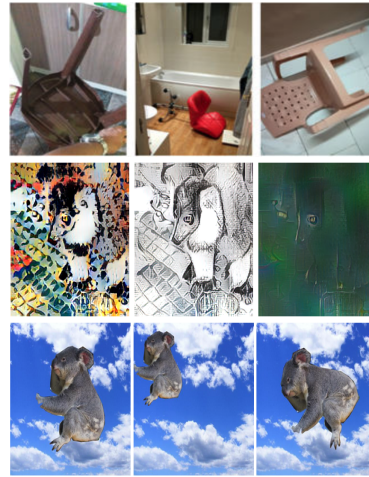

Figure 6: Examples of ObjectNet dataset (1st row), Stylized ImageNet dataset (2nd row), Synthetic dataset (3rd row).

4.2. ImageNet

To validate our findings on a larger scale and more realistic settings, we apply proposed solutions on ImageNet[16] with various altered and synthetic datasets as unstable shift in the data variable. See Figure 6 and D for more information on **ObjectNet, Stylized-ImageNet, Synthetic dataset**.

ObjectNet[3] is a collection of objects that are intentionally placed at an unusual view angle and backgrounds, so that the bias learned by the model with usual data variables is more prominent when testing on ObjectNet. With focus-

ing on 113 classes overlapping with ImageNet classes, we evaluate our second proposed solution explicitly on Object-Net since the shift presented in ObjectNet is very unstable and it is very challenging to overcome the negative effect of the shift.

Stylized ImageNet[24] change the style of original ImageNet image to a random artistic style. With this drastic shift in the data variable, the analysis of the first solution is more insightful on the robustness of dimensions of $\mathbf{A}$ on a very different shift in data variable.

Synthetic Data follows synthetic procedure in [18] where object is masked on a background at a location with a rotation angle. We explore the benefit of this synthesized dataset at three modes: *background, location, rotation* where the target variable is randomly sampled with the other two variables fixed. Additionally, we explore *texture* as an independent variable by masking 'texturized' objects in [23]. We set total number of updating steps per epoch as 100 with batch size 256. This means we select total number of 512000 synthesized images every epoch.

Experiment Setup ResNet 50 pretrained on ImageNet and a linear classifier fintuned via SimCLR, BYOL, and SimSiam are tested with both proposed solutions. For Stable Inference Mapping, a linear layer is optimized with an Adam optimizer for 10 epochs on the synthetic dataset.

Robust Dimensions For each of ImageNet validation data sample, we stylized the image to a random artistic fashion. We observe the dramatic performance difference between the ImageNet stable images and Stylized unstable ImageNet images $(\mathbf{x}_{stable}, \mathbf{x}_{unstable})$. The result is shown in Figure 7. For SimCLR, passing the top 10% important dimensions can close a small performance gap between stable representations and unstable representations. Nonetheless, all SSL seem to be sensible to the strong style change as they shorten the difference between ImageNet to an modest extent. This is as expected since the Stylized ImageNet changes multiple variables to an extreme value.

Stable Inference Mapping As described in **Synthetic Data**, we explore the benefit of our second proposed method to evaluate on a dataset that most of samples are unstable according to learned $\mathbf{A}$. At each training step, with other variables randomly fixed, 10 images with random target variable values are generated. The pair with maximum $m(D(\mathbf{x})) - m(D(\tilde{\mathbf{x}}))$ is selected to train the linear transformation. In Table 2, inferring $\mathbf{F}$ via controlling location produces least improvement. This is expected since the network is robust to translation by design. While background, rotation, and texture improves the performance considerably with the consideration on the training time. However, in D we show that training the model longer using **Stable Inference Mapping** yield less favorable results since the improvement is less significant and starts to saturate at around 30 epochs.
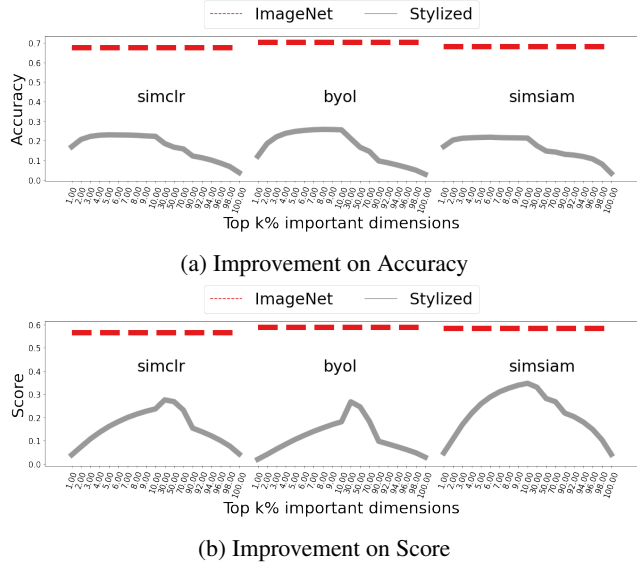

(a) Improvement on Accuracy

(b) Improvement on Score

Figure 7: By identifying and passing top k % important dimensions, the deterioration in both accuracy and prediction score is alleviated.

	SimCLR	BYOL	SimSiam
w/o F	10.12	14.04	11.15
background	12.34	15.43	12.45
location	10.35	14.13	11.17
rotation	12.04	15.88	12.38
texture	12.79	15.06	13.11

Table 2: Comparison of inferring different $\mathbf{F}$ on ObjectNet with a target data variable random trials.

5. Limitations

Though we identify the root cause of unstable inference for descriminative SSL by constructing a causal framework inspired by the prior work, the proposed solutions are constrained and limited to be applied on realistic applications. **Robust Dimensions** involves establishing a correspondence between stable and unstable instances on a one-to-one basis, enabling the identification of dimensions contributing to stability. On the other hand, **Stable Inference Mapping** necessitates a collection of unstable instances with a specific alteration in a particular group of data variables. Within the Causal3dIdent dataset, both solutions can be assessed using the same unstable instances. In more realistic datasets, achieving a one-to-one correspondence is feasible, and manipulation of one group of data variables can be accomplished using synthetic data. However, any assessments with involving artificially generated images might introduce some level of uncertainty. In a realistic setup, since training samples are not directly observ-

able during the inference stage, simple interventions on inference samples may not effectively separate the unstable variables from the stable ones. Consequently, the potential benefits of the proposed solutions in realistic datasets are undermined.

6. Conclusions

In conclusion, this paper has proposed a novel approach to address the issue of unstable behavior during the inference stage in SSL methods. By building on the previous theories of successful InfoNCE-facilitated contrastive SSL and extending it to recent SSL methods, we have demonstrated that a change in the data factor can result in a shift in the inferred representation, leading to a decline in downstream performance. We have proposed learning targeted transformations that regularize the violating shift and restore performance on the unseen data shift. Our experiments on both controlled and realistic datasets have shown the efficacy of our proposed solutions. These contributions provide a better understanding of SSL methods and offer a promising solution to the problem of unstable behavior during the inference stage. We hope that our work will inspire further research in this area and lead to improved SSL methods that are more robust to changes in data factors.

7. Acknowledgements

References

[1] Ibrahim Ahmad and Pi-Erh Lin. A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.). *IEEE Transactions on Information Theory*, 22(3):372–375, 1976. 14

[2] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019. 2

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32, 2019. 7

[4] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006. 3

[5] Sergiy V Borodachov, Douglas P Hardin, and Edward B Saff. *Discrete energy on rectifiable sets*, volume 3. Springer, 2019. 4

[6] Yue Cao, Mingsheng Long, and Jianmin Wang. Unsupervised domain adaptation with distribution matching machines. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3

[7] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020. 1

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[9] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 1

[10] 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. 1, 2, 5, 6, 13

[11] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 1, 2

[12] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021. 1

[13] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 17

[14] Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20(1):99–148, 2007. 4

[15] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009. 3

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[17] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2

[18] Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, Sylvain Gelly, Neil Houlsby, Xiaohua Zhai, and Mario Lucic. On robustness and transferability of convolutional neural networks, 2020. 8, 16

[19] Lixin Duan, Dong Xu, and Ivor Tsang. Learning with augmented features for heterogeneous domain adaptation. *arXiv preprint arXiv:1206.4660*, 2012. 3

[20] Deniz Erdogmus, Kenneth E Hild, Yadunandana N Rao, and Jose C Principe. Minimax mutual information approach for independent component analysis. *Neural Computation*, 16(6):1235–1252, 2004. 2

[21] Ronald Aylmer Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305, 1953. 3

[22] Bernhard C. Geiger and Gernot Kubin. On the information loss in memoryless systems: The multivariate case. *CoRR*, abs/1109.4856, 2011. 14

[23] Robert Geirhos, Claudio Michaelis, Felix A. Wichmann, Patricia Rubisch, Matthias Bethge, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *7th International Conference on Learning Representations, ICLR 2019*, (c):1–22, 2019. 8

[24] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. 8

[25] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 1, 2

[26] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1, 2, 5

[27] Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable Guarantees for Self-Supervised Deep Learning with Spectral Contrastive Loss. *Advances in Neural Information Processing Systems*, 7(NeurIPS):5000–5011, 2021. 2

[28] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 1, 2, 5

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[31] Geoffrey E Hinton and Richard Zemel. Autoencoders, minimum description length and helmholtz free energy. *Advances in neural information processing systems*, 6, 1993. 1

[32] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 2

[33] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986. 5

[34] Tzu Ming Harry Hsu, Wei Yu Chen, Cheng-An Hou, Yao-Hung Hubert Tsai, Yi-Ren Yeh, and Yu-Chiang Frank Wang. Unsupervised domain adaptation with imbalanced cross-domain data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4121–4129, 2015. 3

[35] Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720, 2000. 4

[36] Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and Nonlinear ICA. *Advances in Neural Information Processing Systems*, (Nips):3772–3780, 2016. 2

[37] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017*, 2017. 2

[38] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999. 2

[39] Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics*, 89, 2020. 2

[40] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020. 1

[41] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4037–4058, 2020. 1

[42] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4893–4902, 2019. 3

[43] Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. 108(2015), 2019. 2

[44] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020. 2

[45] David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding. pages 1–51, 2020. 2

[46] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 16

[47] Quoc V Le, Will Y Zou, Serena Y Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features

for action recognition with independent subspace analysis. In *CVPR 2011*, pages 3361–3368. IEEE, 2011. 4

[48] Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting What You Already Know Helps: Provable Self-Supervised Learning. *Advances in Neural Information Processing Systems*, 1(NeurIPS):309–323, 2021. 2

[49] Chunyuan Li, Jianwei Yang, Pengchuan Zhang, Mei Gao, Bin Xiao, Xiyang Dai, Lu Yuan, and Jianfeng Gao. Efficient Self-supervised Vision Transformers for Representation Learning. 3:1–27, 2021. 1

[50] Wen Li, Lixin Duan, Dong Xu, and Ivor W Tsang. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1134–1148, 2013. 3

[51] Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988. 2

[52] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Bjoern W Schuller. Audio self-supervised learning: A survey. *Patterns*, 3(12):100616, 2022. 1

[53] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2021. 1

[54] Yixin Liu, Ming Jin, Shirui Pan, Chuan Zhou, Yu Zheng, Feng Xia, and Philip Yu. Graph self-supervised learning: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 1

[55] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer joint matching for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1410–1417, 2014. 3

[56] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation Learning via Invariant Causal Mechanisms. pages 1–21, 2020. 2

[57] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI*, pages 69–84. Springer, 2016. 1, 2

[58] Aaron Van Den Oord. Representation Learning with Contrastive Predictive Coding. 2

[59] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 1

[60] Judea Pearl. *Causality: Models, reasoning, and inference, second edition*. 2011. 1

[61] Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. Self-supervised learning for videos: A survey. *ACM Computing Surveys*, 2022. 1

[62] Saeed Shurrab and Rehab Duwairi. Self-supervised learning methods and applications in medical imaging analysis: A survey. *PeerJ Computer Science*, 8:e1045, 2022. 1

[63] Mamatha Thota and Georgios Leontidis. Contrastive domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2209–2218, 2021. 3

[64] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 15

[65] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010. 1

[66] Riccardo Volpi, Pietro Morerio, Silvio Savarese, and Vittorio Murino. Adversarial feature augmentation for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5495–5504, 2018. 3

[67] Kügelgen Von, Julius and, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style Luigi Gresele. *Yash Sharma*, 3(NeurIPS):4, 2021. 2, 3, 4, 15

[68] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 3, 4

[69] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022. 3

[70] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-13:9871–9881, 2020. 3, 4

[71] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022. 1

[72] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. Self-Supervised Learning with Swin Transformers. pages 1–8, 2021. 1

[73] Detai Xin, Tatsuya Komatsu, Shinnosuke Takamichi, and Hiroshi Saruwatari. Disentangled speaker and language representations using mutual information minimization and domain adaptation for cross-lingual tts. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6608–6612. IEEE, 2021. 2

[74] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021. 1, 2, 3, 14, 15

[75] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Octo-*

ber 11-14, 2016, Proceedings, Part III 14, pages 649–666. Springer, 2016. 1, 2

[76] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive Learning Inverts the Data Generating Process. 2021. 2, 3, 4, 5, 6, 15