Contrastive Learning can Identify the Underlying Generative Factors of the Data

Roland S. Zimmermann[1,2*] **Steffen Schneider**[1,2*] **Yash Sharma**[1,2*]
Matthias Bethge[1†] **Wieland Brendel**[1†]

[1]University of Tübingen [2]IMPRS for Intelligent Systems

Abstract

Contrastive learning has recently seen tremendous success in unsupervised learning, but the understanding of the source of their effective generalization to a large variety of downstream tasks has been limited. We rigorously show that feedforward models trained on a common contrastive loss can implicitly invert the underlying generative model of the observed data up to affine transformations. While we detail the set of assumptions which need to be met to prove this result, our empirical results suggest our findings are robust to considerable model mismatch. We demonstrate contrastive learning performs comparably to the state-of-the-art in disentanglement on benchmark datasets, a notable observation due to the unique lack of an explicit generative objective. This highlights a deep connection between contrastive learning, generative modeling, and nonlinear independent component analysis, providing a theoretical foundation to derive more effective contrastive losses while simultaneously furthering our understanding of the learned representations.

1 Introduction and Related Work

Contrastive learning has been tremendously successful in unsupervised representation learning for image and sequential data [1–13]. In essence, contrastive methods aim to learn representations where related samples are aligned (positive pairs, e.g. augmentations of the same image), while unrelated samples are separated (negative pairs) [9]. This intuitively leads to invariance to irrelevant details or transformations (by decreasing the distance between positive pairs), while preserving a sufficient amount of information about the input needed for solving downstream tasks (by increasing the distance between negative pairs) [14].

In this work, we move beyond intuition by showing that the encoder learned with a contrastive loss can recover the true generative factors of variation (up to affine transformations) if the true generative model of the data matches the assumptions made by the model. This theory bridges the gap between contrastive learning, and the fields of nonlinear ICA and generative modeling. We verify our theoretical findings with controlled experiments and provide evidence that our theory holds true in practice, even if the assumptions on the ground-truth generative model are partially violated.

Contrastive Learning The common motivation behind contrastive learning (CL) is based on the InfoMax principle [15], which is typically instantiated as maximizing the mutual information (MI) between two views [5, 7, 16]. However, this interpretation is known to be inconsistent with actual behavior in practice, as optimizing a tighter bound on MI can empirically lead to worse representations [17]. Analysis based on the assumption of latent classes provides theoretical insights [18], but unfortunately has a rather large gap with empirical practice, as the result that representation quality

[*]Equal contribution. [†]Joint supervision. Correspondence: first.last@bethgelab.org

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

suffers with a large number of negative samples is inconsistent with empirical observations [2, 5, 8, 9]. More recently, Wang and Isola [19] analyze and characterize the behavior of CL from the perspective of *alignment* and *uniformity* properties, with experiments on standard representation learning tasks demonstrating strong correlation between both metrics and downstream task performance. We build on these results to make a further connection with the seminal cross entropy function, leveraging this to provide identifiability results for a practically successful instantiation of contrastive learning.

Nonlinear ICA Nonlinear Independent Components Analysis (ICA) attempts to find the underlying components for multidimensional data. Said components correspond to a well-defined generative model g, which is assumed to be invertible [20, 21]. In other words, nonlinear ICA solves a demixing problem, i.e., given observed data $\mathbf{x} = g(\mathbf{s})$, it finds the inverse model g^{-1} that allows the original sources $\mathbf{s}$ to be recovered. Hyvärinen et al. [22] show that the nonlinear demixing problem can be solved as long as the independent components are conditionally mutually independent with respect to some auxiliary variable. The authors further provide practical estimation algorithms for solving the nonlinear ICA problem [23–25]. Khemakhem et al. [26] show that this framework generalizes to a broad family of deep latent-variable models. While contrastive learning has previously been used for estimation of specified probabilistic models [22, 24, 25], our work instead leverages identifiability as a tool for understanding instantiations found to be practically successful for representing high-dimensional, complex sensory input [9]. In a similar vein, [27] provide identifiability conditions satisfied by practically successful models for representation learning, showing different representation functions, learned on the same data distribution, live within linear transformations of each other. We instead analyze the relation with respect to the data generating process, showing the inverse of that process lives within a linear transformation of the contrastive optimum.

2 Theory

We will show a connection between contrastive learning and identifiability in the form of nonlinear ICA. For this, we introduce a feature encoder f that maps observations $\mathbf{x}$ to representations. We consider the popular *InfoNCE* loss, which assumes ℓ_2 normalized representations, to perform CL

$$\mathcal{L}_{\mathsf{Contr}}(f;\tau,M) \quad := \underset{\substack{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathsf{pos}} \\ \{\mathbf{x}_i^-\}_{i=1}^M \overset{\text{i.i.d.}}{\sim} p_{\mathsf{data}}}}{\mathbb{E}} \left[-\log \frac{e^{f(\mathbf{x})^\mathsf{T} f(\tilde{\mathbf{x}})/\tau}}{e^{f(\mathbf{x})^\mathsf{T} f(\tilde{\mathbf{x}})/\tau} + \sum_{i=1}^{M} e^{f(\mathbf{x}_i^-)^\mathsf{T} f(\tilde{\mathbf{x}})/\tau}} \right]. \tag{1}$$

Here $M \in \mathbb{Z}_+$ is a fixed number of negative samples, p_{data} is the distribution of all observations and p_{pos} is the distribution of positive pairs. This loss is based on the InfoMax principle [15] and has been shown to be effective by many recent representation learning methods [1, 2, 5–9, 12]. Note that the theoretical results of this paper also hold for a loss function whose denominator only consists of the second summand (i.e. the SimCLR loss [9]).

In the spirit of literature on nonlinear ICA [22–26, 28–30], we assume that the observations $\mathbf{x} \in \mathcal{X}$ are generated by an invertible (injective) generative process $g : \mathcal{Z} \to \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^N$ is the space of observations and $\mathcal{Z}$ denotes the space of latent factors. Influenced by the feature normalization in InfoNCE, we further assume that $\mathcal{Z}$ is the unit hypersphere $\mathbb{S}^{N-1}$ (see Appx. A.1). Additionally, we assume that the ground-truth marginal distribution of the latents of the generative process is uniform and that the conditional distribution is a von Mises-Fisher (vMF) distribution:

$$p(\mathbf{z}) = |\mathcal{Z}|^{-1}, \qquad p(\mathbf{z}|\tilde{\mathbf{z}}) = C_p^{-1} e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \quad \text{with} \quad C_p := \int e^{\kappa \mathbf{z}^\top \tilde{\mathbf{z}}} \, \mathrm{d}\tilde{\mathbf{z}} = \text{const.}, \quad \mathbf{x} = g(\mathbf{z}) \tag{2}$$

Given these assumptions, we will show that if f optimizes the contrastive loss $\mathcal{L}_{\mathsf{contr}}$, then f solves the demixing problem, i.e., inverts g. Detailed proofs are given in Appx. A.3. Note that one can derive similar properties beyond spherical spaces, e.g. for $\mathbb{R}^N$ or convex bodies (Appx. A.4).

2.1 Relation between contrastive learning and cross-entropy minimization

From the perspective of nonlinear ICA, we are interested in understanding how the representations $f(\mathbf{x})$ which minimize the contrastive loss (1) are related to the ground-truth source signals $\mathbf{z}$. To study this relationship, we focus on the mapping $h = f \circ g$ between the recovered source signals $h(\mathbf{z})$ and the true source signals $\mathbf{z}$. A core insight is a connection between the contrastive loss and the cross-entropy between the ground-truth latent distribution and a certain model distribution. For this, we expand the theoretical results obtained by Wang and Isola [19]:

Theorem 1 ($\mathcal{L}_{\mathsf{contr}}$ 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*

$$\lim_{M\to\infty} \mathcal{L}_{\mathsf{contr}}(f;\tau,M) - \log M + \log|\mathcal{Z}| = \mathop{\mathbb{E}}_{\mathbf{z}\sim p(\mathbf{z})}\left[H(p(\cdot|\mathbf{z}), q_h(\cdot|\mathbf{z}))\right] \tag{3}$$

where H is the cross-entropy between the ground-truth conditional distribution over positive pairs p and the conditional distribution over the recovered latent space q_h, and $C_h(\tilde{\mathbf{z}}) \in \mathbb{R}^+$ is the partition function of q_h (see Appx. A.2):

$$q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h(\tilde{\mathbf{z}})^{-1} e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau} \quad \textit{with} \quad C_h(\mathbf{z}) := \int e^{h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau}\,\mathrm{d}\mathbf{z}. \tag{4}$$

This result makes analyzing the problem of CL simpler, as it reduces it to deducing properties from the well-understood cross-entropy objective. Interestingly, for uniform ground truth marginal distributions on bounded spaces, $C_h^{-1}(\mathbf{z})$ also coincides with the pushforward of $p(\mathbf{z})$ through h (see Sec. A.5), which eventually becomes a uniform distribution at the optimum.

Next, we show that the minimizers h^* of the cross-entropy (4) are isometries in the sense that $\kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h^*(\mathbf{z})^\top h^*(\tilde{\mathbf{z}})$ for all $\mathbf{z}$ and $\tilde{\mathbf{z}}$. In other words, they preserve the dot product between $\mathbf{z}$ and $\tilde{\mathbf{z}}$.

Proposition 1 (Minimizers of the cross-entropy maintain the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, $\tau > 0$ and consider the ground-truth conditional distribution of the form $p(\tilde{\mathbf{z}}|\mathbf{z}) = C_p^{-1}\exp(\kappa\tilde{\mathbf{z}}^\top\mathbf{z})$. Let h map onto a hypersphere with radius $\sqrt{\tau/\kappa}$.*[2] *Consider the model conditional distribution*

$$q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z}) = C_h^{-1}(\mathbf{z})\exp(h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau) \quad \textit{with} \quad C_h(\mathbf{z}) := \int_{\mathcal{Z}} \exp(h(\tilde{\mathbf{z}})^\top h(\mathbf{z})/\tau)\,\mathrm{d}\tilde{\mathbf{z}}, \tag{5}$$

where the hypothesis class for h is assumed to be sufficiently flexible such that $p(\tilde{\mathbf{z}}|\mathbf{z})$ and $q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z})$ can match. If h^* is a minimizer of the cross-entropy $\mathbb{E}_{p(\tilde{\mathbf{z}}|\mathbf{z})}[-\log q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z})]$, then $p(\tilde{\mathbf{z}}|\mathbf{z}) = q_\mathrm{h}(\tilde{\mathbf{z}}|\mathbf{z})$ and $\forall \mathbf{z}, \tilde{\mathbf{z}} : \kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$.

We show in the Appendix (Lemma 3) that this implies that the empirical marginal distribution (pushforward of p through h, denoted as $p_{\#h}$, which coincides with $C_h(\mathbf{z})^{-1}$) is also a uniform distribution, i.e., $p_{\#h}(\mathbf{z}) = \text{const}$.

2.2 Contrastive learning identifies the ground-truth factors

From the strong property of isometry, we can now deduce a key property of the minimizers h^*:

Proposition 2 (Extension of the Mazur-Ulam theorem to hyperspheres and the dot product)**.** *Let $\mathcal{Z} = \mathbb{S}^{N-1}$. If $h : \mathcal{Z} \to \mathcal{Z}$ maintains the dot product up to a constant factor, i.e., $\forall \mathbf{z}, \tilde{\mathbf{z}} \in \mathcal{Z} : \kappa\mathbf{z}^\top\tilde{\mathbf{z}} = h(\mathbf{z})^\top h(\tilde{\mathbf{z}})$, then h is an affine transformation.*

In the last step, we can combine the previous propositions to derive our main result: the minimizers of $\mathcal{L}_{\mathsf{contr}}$ solve the demixing problem of nonlinear ICA up to linear transformations, i.e., they identify the original sources $\mathbf{z}$ for observations $g(\mathbf{z})$ up to linear transformations.

Theorem 2. *Let $\mathcal{Z} = \mathbb{S}^{N-1}$, the ground-truth marginal be uniform, and the conditional a vMF distribution (cf. Eq. 2). If the mixing function g is differentiable and invertible, and if f is differentiable and minimizes the CL loss (1), then for fixed $\tau > 0$ and $M \to \infty$, $h = f \circ g$ is affine, i.e., it recovers the latent sources up to affine transformations.*

Note that we do not assume knowledge of the ground-truth generative model g; we only make assumptions about the distribution of latents and that f minimizes the contrastive loss. In Appx. Theorem 6, we show a similar result for $\mathcal{Z} = \mathbb{R}^N$ and a wider class of ground-truth conditional densities: if f minimizes the cross-entropy, then h is again an affine transformation. Having an exact match between the assumed model distribution q_h and the ground-truth conditional is unlikely to happen in practice. While a theoretical account for these cases is beyond the scope of this work, we provide empirical evidence that h is still an affine transformation even if there is a severe mismatch.

[2] Note that in practice this can be implemented as a learnable rescaling operation of the network f.

Table 1: Robustness of CL to a mismatch between model assumptions and the ground truth averaged over 5 runs. Note that only the first two rows correspond to settings that match our assumptions, while the other show results for violated assumptions (see column *M.*).

Generative process g			Model f			R^2 [%]		
Space	$p(\cdot)$	$p(\cdot\|\cdot)$	Space	$q_{\mathrm{h}}(\cdot\|\cdot)$	M.	Linear Score	Supervised Score	Unsupervised Score
$\mathbb{S}_1^9$	Uniform	vMF(κ=1)	$\mathbb{S}_1^9$	vMF(κ=1)	✓	77.58 ± 3.06	99.82 ± 0.03	99.42 ± 0.05
$[0,1]^{10}$	Uniform	GenNorm(β=2, λ=0.05)	$\mathbb{R}^{10}$	GenNorm(β=2)	✓	87.71 ± 4.64	99.80 ± 0.03	99.52 ± 0.02
$\mathbb{S}_1^9$	Uniform	vMF(κ=10)	$\mathbb{S}_1^9$	vMF(κ=1)	✗	77.58 ± 3.06	99.81 ± 0.03	99.86 ± 0.01
$\mathbb{S}_1^9$	Uniform	GenNorm(β=1, λ=0.05)	$\mathbb{S}_1^9$	vMF(κ=1)	✗	77.98 ± 2.97	99.81 ± 0.03	99.88 ± 0.03
$\mathbb{S}_1^9$	Uniform	GenNorm(β=2, λ=0.05)	$\mathbb{S}_1^9$	vMF(κ=1)	✗	77.58 ± 3.06	99.80 ± 0.04	99.86 ± 0.00
$[0,1]^{10}$	Uniform	GenNorm(β=1, λ=0.05)	$\mathbb{R}^{10}$	GenNorm(β=2)	✗	87.71 ± 4.64	99.81 ± 0.03	99.53 ± 0.02
$[0,1]^{10}$	Uniform	GenNorm(β=1, λ=0.05)	$\mathbb{R}^{10}$	GenNorm(β=3)	✗	87.71 ± 4.64	99.81 ± 0.03	99.70 ± 0.02
$[0,1]^{10}$	Uniform	GenNorm(β=2, λ=0.05)	$\mathbb{R}^{10}$	GenNorm(β=3)	✗	87.71 ± 4.64	99.83 ± 0.03	99.69 ± 0.02
$\mathbb{S}_1^9$	GenNorm(β=2, λ=1)	GenNorm(β=1, λ=0.05)	$\mathbb{S}_1^9$	vMF(κ=1)	✗	78.34 ± 3.83	99.78 ± 0.04	98.94 ± 0.03
$\mathbb{S}_1^9$	GenNorm(β=2, λ=1)	GenNorm(β=2, λ=0.05)	$\mathbb{S}_1^9$	vMF(κ=1)	✗	78.05 ± 3.17	99.78 ± 0.05	98.94 ± 0.02
$\mathbb{R}^{10}$	GenNorm(β=1, λ=1)	GenNorm(β=2, λ=1)	$\mathbb{R}^{10}$	GenNorm(β=2)	✗	73.28 ± 2.35	99.76 ± 0.02	97.57 ± 0.37
$\mathbb{R}^{10}$	GenNorm(β=2, λ=1)	GenNorm(β=2, λ=1)	$\mathbb{R}^{10}$	GenNorm(β=2)	✗	75.43 ± 2.68	99.80 ± 0.04	98.70 ± 0.04

3 Experiments

Validation of theoretical claim We now validate our theoretical claims under both perfectly matching and violated conditions regarding the ground truth marginal distribution, the mixing function, and the ground truth conditional. We consider source signals of dimensionality $n = 10$. We sample pairs of source signals in two steps: First, we sample from $p(\mathbf{z})$, which, if chosen to be uniform, matches our assumptions, if not (e.g. Normal distribution), violates our assumptions. Second, we generate the positive pair by sampling from a conditional $p(\tilde{\mathbf{z}}|\mathbf{z})$. We also consider spaces beyond the hypersphere, such as unbounded $\mathbb{R}^N$ and the bounded hypercube (which is a convex body). We generate the observations with a multi-layer perceptron (MLP) following the settings used by [24, 25]. Specifically, we use leaky ReLU units and control the condition number of the weight matrix to ensure that the MLP is invertible. For more details about the experimental setup and used loss functions, see Sec. A.6.

To test for identifiability up to affine transformations, we fit a linear regression between the ground-truth and recovered sources and report the coefficient of determination (R^2). In each setting, we report three scores. First, we ensure that the problem requires nonlinear demixing by considering a linear model, which amounts to the score of a linear fit between observations and sources (Linear Score). Second, we ensure that the problem is solvable within our model class by training our model f with supervision, minimizing the mean-squared error between $f(g(\mathbf{z}))$ and $\mathbf{z}$ (Supervised Score; upper bound). Third, we fit our model without supervision using a contrastive loss (Unsupervised Score). The results in Table 1 show that CL recovers a score close to the empirical upper bound, and mismatch in assumptions on the marginal and conditional only lead to a slight drop in performance.

Extensions to image data We evaluate CL on disentanglement with complex pixel inputs (Appx. A.6.2, 31, 32), benchmarking applicability beyond the theoretical conditions. For better control over the objective, we split the CL loss into uniformity and alignment terms (Appx. A.6.1). Appx. Table 2 shows that scores increase consistently with an increased relative weight on the uniformity term, denoting the importance of uniformity in preventing collapse due to optimization of the alignment loss. For Laplace transitions of the latents (LAP), the BetaVAE score [33], which corresponds to fitting a logistic classifier to the absolute differences in the model latents, reaches 100%, providing evidence to the robustness of CL in identifying the sources up to affine transformations. Finally, we note the surprising performance on KITTI Masks (Appx. Table 5), which contains natural shapes and continuous natural transitions, hinting at the scalability benefits of CL.

4 Conclusion

This work shows that contrastive learning can uncover the true generative factors of variation underlying the observational data. We verify this claim theoretically and give evidence that our theory also holds under much milder assumptions on the generative model. Our work connects CL, generative modeling and nonlinear ICA, thereby laying a strong theoretical foundation for one of the most successful self-supervised learning techniques.

References

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

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

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

[4] Olivier J Hénaff, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

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

[6] 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.

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

[8] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

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

[11] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. *CoRR*, abs/1910.05453, 2020.

[12] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*, 2020.

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

[14] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning, 2020.

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

[16] Mike Wu, Chengxu Zhuang, Daniel Yamins, and Noah Goodman. On the importance of views in unsupervised representation learning. 2020.

[17] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

[18] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637, 2019.

[19] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. *arXiv preprint arXiv:2005.10242*, 2020.

[20] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

[21] Christian Jutten, Massoud Babaie-Zadeh, and Juha Karhunen. Nonlinear mixtures. *Handbook of Blind Source Separation, Independent Component Analysis and Applications*, pages 549–592, 2010.

[22] Aapo Hyvärinen, Hiroaki Sasaki, and Richard E Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. *arXiv preprint arXiv:1805.08651*, 2018.

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

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

[25] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Proceedings of Machine Learning Research*, 2017.

[26] Ilyes Khemakhem, Diederik P Kingma, and Aapo Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

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

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

[29] Stefan Harmeling, Andreas Ziehe, Motoaki Kawanabe, and Klaus-Robert Müller. Kernel-based nonlinear blind source separation. *Neural Computation*, 15(5):1089–1124, 2003.

[30] Henning Sprekeler, Tiziano Zito, and Laurenz Wiskott. An extension of slow feature analysis for nonlinear blind source separation. *The Journal of Machine Learning Research*, 15(1):921–947, 2014.

[31] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020.

[32] 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. *arXiv preprint arXiv:xxx*, 2020.

[33] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations (ICLR)*, 2(5):6, 2017.

[34] Yen-Chi Chen. A tutorial on kernel density estimation and recent advances, 2017.

[35] I. 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.

[36] Michael Ruzhansky and Mitsuru Sugimoto. On global inversion of homogeneous maps. *Bulletin of Mathematical Sciences*, 5(1):13–18, 2015.

[37] John M Lee. Smooth manifolds. In *Introduction to Smooth Manifolds*, pages 606–607. Springer, 2013.

[38] Stanisław Mazur and Stanisław Ulam. Sur les transformations isométriques d’espaces vectoriels normés. *CR Acad. Sci. Paris*, 194(946-948):116, 1932.

[39] Bogdan Nica. The mazur-ulam theorem, 2013.

[40] Piotr Mankiewicz. Extension of isometries in normed linear spaces. *Bulletin de l'Academie polonaise des sciences: Serie des sciences mathematiques, astronomiques et physiques*, 20(5): 367–+, 1972.

[41] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

[42] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

[43] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. *arXiv preprint arXiv:1905.04804*, 2019.

[44] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[45] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.