MANGO: Learning Disentangled Image Transformation Manifolds with Grouped Operators

Brighton Ancelin[1], Yenho Chen[1], Alex Saad-Falcon[1], Peimeng Guan[2], Chiraag Kaushik[2],
Nakul Singh[1], and Belen Martin-Urcelay[2]
[1]*ML@GT* and [2]*Dept. of Electrical and Computer Engineering*
Georgia Institute of Technology, Atlanta, GA
{ bancelin3, yenho, asf3, pguan6, ckaushik7, nsingh360, burcelay3}@gatech.edu

***Abstract*—Learning semantically meaningful image transformations (i.e. rotation, thickness, blur) directly from examples can be a challenging task. Recently, the Manifold Autoencoder (MAE) [1] proposed using a set of Lie group operators to learn image transformations directly from examples. However, this approach has limitations, as the learned operators are not guaranteed to be disentangled and the training routine is prohibitively expensive when scaling up the model. To address these limitations, we propose MANGO (transformation Manifolds with Grouped Operators) for learning disentangled operators that describe image transformations in distinct latent subspaces. Moreover, our approach allows practitioners the ability to define which transformations they aim to model, thus improving the semantic meaning of the learned operators. Through our experiments, we demonstrate that MANGO enables composition of image transformations and introduces a one-phase training routine that leads to a $100\times$ speedup over prior works.**

***Index Terms*—Image transformations, disentangled representation, autoencoder, generative model.**

I. Introduction

In many scientific domains, learning semantically meaningful transformations in high-dimensional image datasets can improve our understanding of complex patterns within the underlying system. For example, in medicine, learning image transformations between MRI scans of sick and healthy patients can provide insights into disease mechanisms. To ensure that these learned transformations are interpretable, the resulting representations must be low-dimensional and disentangled.

Although there exists numerous methods to learn a low-dimensional representation of high-dimensional image manifolds [2]–[7], few works aim to model semantically meaningful transformations between different points on this manifold. Many latent variable models incorrectly assume Euclidean transformations in the latent space, which can be inappropriate when modeling certain transformations that require forming a closed path, such as image rotations.

Recently, a new line of research [5], [8], [9] has imposed structure by constraining image transformation onto a low-dimensional manifold that can be learned. In particular, the Manifold Autoencoder (MAE) [1] learns a dictionary of Lie group operators, referred to as *transport operators*, which can be linearly combined to model possible transformations between latent points. To encourage the learning of semantically meaningful and disentangled operators, the MAE optimizes an objective function that promotes sparsity of the weights and shrinkage of the operators.

Despite demonstrating improved extrapolation along transformation paths, there are two critical limitations of the MAE that prevent it from efficiently learning a disentangled latent space. First, the MAE objective does not guarantee that latent transformations are orthogonal which often results in the learning of overlapping transport operators for distinct transformations. Second, training MAE requires an expensive three-phase procedure which becomes prohibitively expensive when scaling up the model.

To address these limitations of MAE, we propose **MANGO**, an approach for learning disentangled image transformation **Man**ifolds using **G**rouped **O**perators. We introduce the concept of disentangled operators and enforce disentanglement by constraining the action of each operator to a distinct latent subspace. Furthermore, our method allows practitioners to specify which semantically meaningful transformations they aim to model. Additionally, we show that our disentangled formulation enables composable latent transformations that can generate realistic images, even for out-of-distribution transformations. Finally, we present a one-phase training strategy that significantly improves computational efficiency, demonstrating a $100\times$ speedup over previous methods.

II. Background and Related Work

A. Disentangled Representation Learning

Disentangled representation learning seeks to identify independent latent factors of variation that best describe the dataset of interest. By doing so, disentangled models improve the interpretability of the latent variables which encourages the discovery of semantically meaningful structures. Various techniques in deep learning literature have been developed to promote the learning of disentangled latent spaces. For example, β-VAE [10] applies strong regularization to the KL divergence term in the evidence lower bound to encourage the latent factor distribution to align with an isotropic Gaussian prior. Similarly, FactorVAE [9] promotes disentanglement by

This work was supported in part by the James S. McDonnell Foundation grant number 220020399 and the Georgia Institute of Technology. The student authors are listed in alphabetical order by last name.

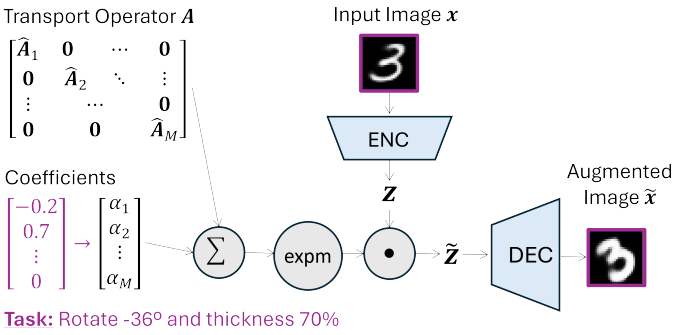

Fig. 1. Block diagram of MANGO. The model simultaneously learns a transport operator $\boldsymbol{A}$, where each block diagonal component represents a semantically meaningful transformation, and an autoencoder that reconstructs the transformed image.

directly penalizing the total correlation of the latent prior. Alternatively, Generative Adversarial Networks (GANs), which aim to learn generative models capable of producing samples indistinguishable from real data by a discriminative classifier, provide a different strategy to disentanglement. In particular, InfoGAN [11] achieves disentanglement by maximizing the mutual information between distinct latent subspaces. Other line of work learns latent representations in an unsupervised manner [12]–[14], where [13] minimizes the mutual information between the content embedding and domain embedding to encourage independence.

In general, however, these methods do not guarantee that the learned latent factors will correspond to semantically meaningful transformations, since they lack mechanisms to incorporate additional practitioner-provided information. Furthermore, these models assume a Euclidean structure in the latent space, which may be unsuitable for certain image transformations that follow closed paths, such as rotations where transformations are more appropriately modeled on $SO(\cdot)$ manifolds.

B. Transport Operators

Transport operators [15] offer a framework for modeling continuous transformations between high-dimensional data points in their original space, $\boldsymbol{x} \in \mathbb{R}^D$, by defining transformations to follow the flow of a linear dynamical systems, $\dot{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x}$. Given an initial point $\boldsymbol{x}_0$ and a linear operator $\boldsymbol{A} \in \mathbb{R}^{D\times D}$, the trajectory of the transformation is given by $\boldsymbol{x}_t = \exp(t\boldsymbol{A})\boldsymbol{x}_0$ for all time $t \in \mathbb{R}$, where expm is the matrix exponential. Many works [16]–[20] further decompose $\boldsymbol{A}$ into a linear combination of M transport operators $\{\boldsymbol{A}_m\}_{m=1}^M$ such that, $\boldsymbol{A} = \sum_{m=1}^M c_m \boldsymbol{A}_m$, where $c_m \in \mathbb{R}$ is a coefficient that determines the contribution of each operator in a particular transformation. The set of all possible transport operators define a Lie group [21] and can efficiently represent the manifolds surface.

To encourage the learning of statistically independent transport operators, [15] proposes to promote sparsity in the coefficients and shrinkage over the transport operators through regularized optimization of $\boldsymbol{A}, \boldsymbol{c}$ over the following objective,

$$\frac{1}{2}\left\|\tilde{\boldsymbol{x}} - \exp\left(\sum_{m=1}^M c_m \boldsymbol{A}_m\right)\boldsymbol{x}\right\|_2^2 + \frac{\gamma}{2}\sum_{m=1}^M \|\boldsymbol{A}_m\|_{\mathrm{F}}^2 + \zeta\|\boldsymbol{c}\|_1 \tag{1}$$

where γ and ζ are penalty weights for the operator shrinkage and the amount of coefficient sparsity respectively. Since concurrently learning both transport operators and the coefficients can lead to training instability, prior works instead alternate between updating $\boldsymbol{A}$ and $\boldsymbol{c}$.

C. Manifold Autoencoder (MAE)

Learning transport operators in the high-dimensional data space can be challenging numerically and computationally expensive as a result of the matrix exponential. To address this limitation, the manifold autoencoder (MAE) [1], [22] proposes to learn transport operators in a low-dimension latent space of an autoencoder. This is accomplished through a three-phase training routine. First, an autoencoder is trained with the basic reconstruction loss. Second, the algorithm fixes the autoencoder weights and trains the transport operators (in the latent space) using pairs of neighboring points with loss function (1) (modified so that data $\boldsymbol{x}, \tilde{\boldsymbol{x}} \in \mathbb{R}^D$ are replaced by latent representations $\boldsymbol{z}, \tilde{\boldsymbol{z}} \in \mathbb{R}^L$). Third, the algorithm prunes irrelevant operators and simultaneously fine-tunes the autoencoder weights alongside the remaining transport operators. Although MAE improves the computational efficiency of transport operators by learning transformations in the latent space, it still relies on an expensive inner optimization procedure as a result of the ℓ_1 term in Equation (1). Additionally, the learned latent transformations are not guaranteed to be disentangled since the inferred active support set between different transformation may use overlapping coefficients.

III. Methodology

This Section describes the main components of MANGO. Figure 1 illustrates our approach.

A. Learning a Low-Dimensional Latent Manifold

For a given dataset $\{\boldsymbol{x}_i\}_{i=1}^N$, let $\mathcal{H} := \{h_{m,\alpha} : \mathbb{R}^D \to \mathbb{R}^D\}_{m=1}^M$ be a set of continuous semantic transformations on the data, where each h_m is parameterized by a scalar $\alpha \in [-1, 1]$. We train an autoencoder (with encoder and decoder denoted f and g, respectively) to learn a low-dimensional latent representation $\boldsymbol{z}_i \in \mathbb{R}^L$ of the $\boldsymbol{x}_i \in \mathbb{R}^D$ such that the transformations of $\mathcal{H}$ are represented in the latent space (approximately) by a structured class of manifolds.

B. Enforcing Disentangled Operators with Group Structure

Building on prior works [23], [24], we propose the following definition of a disentangled representation for operators,

Definition 1. *(Disentangled Operators) A set of operators $\{\boldsymbol{A}_m\}_{m=1}^M$ are disentangled if every pair of operators $\boldsymbol{A}_i$ and $\boldsymbol{A}_j$, where $i \neq j$, satisfy $\langle \boldsymbol{A}_i, \boldsymbol{A}_j \rangle = 0$.*

Algorithm 1 MANGO Algorithm

Input: Randomly initialized operators $\{\boldsymbol{A}_m\}$ and network weights $\boldsymbol{\theta}$
Output: Learned transport operators $\{\boldsymbol{A}_m\}$ and autoencoder weights $\boldsymbol{\theta}$
for $t = 1, 2, \ldots$ **do**
  Sample data batch $\mathcal{B} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_B\}$
  Encode batch to obtain $\mathcal{L} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_B\}$
  for $m = 1$ **to** M **do**
    $\alpha \sim \text{Unif}([-1, 1])$
    Apply $h_{m,\alpha}$ to $\mathcal{B}$ to obtain $\tilde{\mathcal{B}}_m = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_B\}$
    Encode $\tilde{\mathcal{B}}_m$ to obtain $\tilde{\mathcal{L}}_m = \{\tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_B\}$
    $\boldsymbol{A}_m \leftarrow \boldsymbol{A}_m - \eta_1 \sum_{i=1}^{B} \frac{\partial E_i}{\partial \boldsymbol{A}_m}$
  end for
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta_2 \sum_{i=1}^{B} \frac{\partial E_i}{\partial \boldsymbol{\theta}}$
end for

For each transformation (indexed $m = 1, \ldots, M$), we learn the transport operator $\boldsymbol{A}_m$ by attempting to minimize

$$T_m = \|\tilde{\boldsymbol{z}} - \text{expm}(\alpha \boldsymbol{A}_m)\boldsymbol{z}\|_2^2 + \gamma \|\boldsymbol{A}_m\|_F^2$$

where $\boldsymbol{z} := f(\boldsymbol{x})$ and $\tilde{\boldsymbol{z}} := f(\tilde{\boldsymbol{x}})$ are the latent representations corresponding to the original sample $\boldsymbol{x}$ and the transformed sample $\tilde{\boldsymbol{x}} := h_{m,\alpha}(\boldsymbol{x})$, respectively. This formulation encourages the operator to define a continuous transformation path in the latent space parameterized by α. We further constrain $\boldsymbol{A}_m$ to be a block-diagonal matrix with a different support for each m, so that in aggregate the $\boldsymbol{A}_m$ are *disentangled*. This has the effect that the action of the m^{th} transport map on $\boldsymbol{z}$ is localized to a small, identifiable subset of coordinates.

C. Improving Training Efficiency with One-Phase Approach

Combining these two components, we obtain an overall loss value for each pair of transformed points in the input space $(\boldsymbol{x}, \tilde{\boldsymbol{x}})$ and their corresponding latent representations $(\boldsymbol{z}, \tilde{\boldsymbol{z}})$:

$$E = \|\boldsymbol{x} - g(f(\boldsymbol{x}))\|_2^2 + \|\tilde{\boldsymbol{x}} - g(f(\tilde{\boldsymbol{x}}))\|_2^2 + \lambda T_m, \quad (2)$$

where the the first two terms encourage accurate reconstruction of the original and transformed inputs, while the final term ensures that the block-diagonal transport operator $\boldsymbol{A}_m$ transforms one latent representation into the corresponding transformed latent representation. The training methodology is summarized in Algorithm 1.

IV. Results

The MNIST handwritten digits dataset [25] is a widely used benchmark for evaluating latent structures [1], [22] because it allows for straightforward assessment of semantically meaningful transformations through standard image operations (e.g., rotations, thickness, blurriness). We use MNIST to assess MANGO's ability to learn disentangled latent operators. As a result of enforcing disentanglement, we demonstrate that we can linearly combine the learned operators to perform multiple semantic effects simultaneously. Additionally, we compare the computational complexity of MANGO's one-phase training procedure with the MAE's three-phase training procedure, demonstrating a substantial improvement in training runtime.

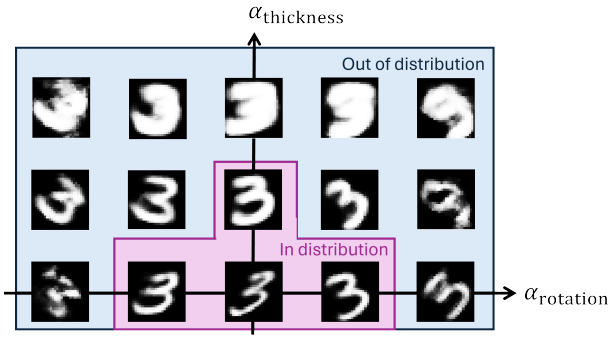

Fig. 2. Augmentations with various combinations of rotations and thickness changes.

A. Disentangled Operators are Composable

We empirically show that MANGO leverages the disentangled manifold structure to learn semantically meaningful operators on the MNIST dataset. Figure 2 shows augmented images generated by applying two distinct transport operators. One operator corresponds to a rotation transformation, while the other increases the thickness of the digits. Note that the reconstruction performance remains comparable to the baseline autoencoders. MANGO provides the additional benefit of interpretability, without a substantial trade-off in image quality.

Additionally, the operators learned by MANGO generalize beyond the training dataset in two ways. First, they are able to transport images further than the transformations observed during training, showing robustness in extrapolation. Second, MANGO is able to linearly combine the learned operators to achieve complex transformations. For instance, as illustrated in Figures 2 and 5, the model successfully generates augmented images where the digits are both rotated and thickened simultaneously. This demonstrates the model's ability to compose transformations in a meaningful and interpretable manner, in contrast with vanilla AEs where images lose their identity.

Figure 6 includes quantitative metrics for our MNIST experiments. We measure 1) image reconstruction error to evaluate the quality of the autoencoder and 2) transformed image reconstruction error to assess the quality of the latent transformation function. We report MSE and LPIPS [26]. For the image transformation metric, we obtain latent traversals from a vanilla autoencoder (AE) by fitting linear transformations on latent points and applying these transformations iteratively to a reference image's embedding. Although AE slightly outperforms MANGO in reconstructing available images, it suffers when generating image transformations. In fact, MANGO demonstrates a 60% improvement in transformed reconstructions.

B. Grouped Operators Improve Training Time

The disentangled group structure of MANGO allows for simpler backpropagation computations, leading to a faster training process. During training for our MNIST experiments,

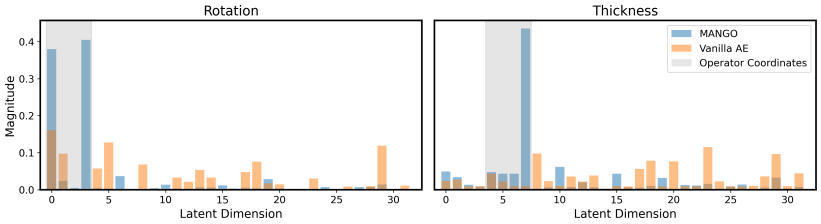

Fig. 3. MANGO achieves a neatly disentangled latent space. The figure shows the magnitude of each coordinate in the first principal component for both models. MANGO exhibits stronger concentration and alignment with learned operator coordinates.

Latent dimension L	16	32	64	128
MAE	6.90	20.52	72.54	319.50
MANGO	0.18	0.18	0.20	0.20

Dictionary Size M	2	4	6	8
MAE	8.98	12.17	14.50	19.92
MANGO	0.16	0.16	0.18	0.18

Fig. 4. Training runtimes (in seconds) per batch (of size 64) for fixed dictionary size $M = 8$ and for fixed latent dimension $L = 32$.

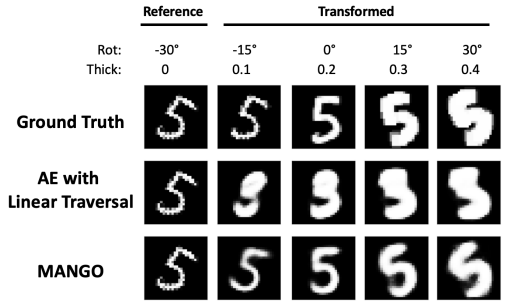

Fig. 5. Comparison of image transformations. MANGO transformations retain image identity unlike the AE.

Models	α	Image MSE	Image LPIPS	Transformed MSE	Transformed LPIPS	Disent. MIG
AE	rotate	**0.011**	**0.043**	0.072	0.119	0.005
	thick	**0.007**	**0.020**	0.093	0.082	0.034
	rotate + thick	**0.013**	**0.042**	0.076	0.085	-
MANGO	rotate	0.015	0.051	**0.027**	**0.057**	**0.031**
	thick	0.011	0.030	**0.022**	**0.039**	**0.11**
	rotate + thick	0.018	0.053	**0.040**	**0.067**	

Fig. 6. Quantiative metrics for the MNIST experiments. We compute scores that quantify image reconstruction, image transformations, and disentanglement. Lower is better for MSE and LPIPS while higher is better for MIG.

we observe that MANGO takes 12 minutes to converge while the baseline MAE requires 138 hours to converge. The overall algorithm takes $\mathbf{0.14\%}$ of the time to fully converge.

The autoencoders in both approaches are fully connected neural networks with hidden layer sizes (256, 64, $\boldsymbol{L}$ (latent space), 64, 256), leaky ReLU hidden layer activations, and a sigmoid final layer activation. The first table in Figure 4 compares the runtimes of batch processing for different latent dimensions $\boldsymbol{L}$; while MANGO requires similar computation time for all $\boldsymbol{L}$, MAE quickly scales on order roughly $\boldsymbol{L^2}$. This is due to the order $\boldsymbol{L}$ (with small coefficient) scaling of the block diagonal MANGO transport matrices, whereas the dense transport matrices of MAE scale on order $\boldsymbol{L^2}$. As a result MANGO trains up to $\mathbf{1500}\times$ faster than MAE for a latent space of size 128. The second table in Figure 4 compares the runtimes for different dictionary sizes $\boldsymbol{M}$; again, MANGO requires similar computation time for all $\boldsymbol{M}$ while MAE scales poorly. For reasonable $\boldsymbol{L}$ and $\boldsymbol{M}$, the manifold-relevant computations are nearly negligible compared to the neural network computations, and as such we observe the nearly constant runtimes for varying $\boldsymbol{L}$ and $\boldsymbol{M}$ on MANGO. All experiments were run on an AMD Ryzen 3900 12 core processor and NVIDIA GeForce RTX 2070.

C. MANGO Disentangles the Latent Space

To study MANGO's effect on the latent space, we apply principal component analysis (PCA) to latent representations of image augmentations. We select 10 random images from the dataset and generate 100 augmentations for each by varying rotation and thickness. These augmentations are then fed to both MANGO and the vanilla autoencoder for comparison. We apply PCA to each set of augmentations and average the results over the 10 sets. For both models, the explained variance concentrates in the first few singular components, with MANGO showing slightly better concentration. Crucially, the large coordinates of MANGO's leading eigenvector align with the operator coordinates, indicating a disentangled latent space. In contrast, the vanilla autoencoder's energy is spread across many coordinates. Figure 3 illustrates these results.

We also measure disentanglement quantitatively with the MIG score [27]. the results in Figure 6 show that MANGO achieves superior disentanglement compared to AE.

V. Future Work

MANGO provides a general framework which can in principle apply to complex image transformations. As future work, we plan on extending our experimental results to learning image transformation in the Fruits-360 dataset [28], a public dataset of images of rotated fruit. This dataset is significantly more challenging than MNIST since our models must represent 3D transformations given 2D snapshots. We train on a small range of rotation parameters ($\boldsymbol{\alpha}$), and find that the learned transport operator can in fact generate images of rotated bananas for values outside the training set. These preliminary results, shown in Figure 7, indicate that our framework could be used for more challenging transformation learning.

Fig. 7. Generated images of rotated bananas, for rotation angles not seen in the training set.

REFERENCES

[1] Marissa Connor and Christopher Rozell. Representing closed transformation paths in encoded network latent space. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3666–3675, 2020.

[2] Wei Wang, Yan Huang, Yizhou Wang, and Liang Wang. Generalized autoencoder: A neural network framework for dimensionality reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2014.

[3] Yangqin Feng, Lei Zhang, and Juan Mo. Deep manifold preserving autoencoder for classifying breast cancer histopathological images. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17(1):91–101, 2020.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[5] Yao Ni, Piotr Koniusz, Richard Hartley, and Richard Nock. Manifold learning benefits gans, 2022.

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[7] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models, 2014.

[8] Kion Fallah, Alec Helbling, Kyle A Johnsen, and Christopher J Rozell. Manifold contrastive learning with variational lie group operators. *arXiv preprint arXiv:2306.13544*, 2023.

[9] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2649–2658, 2018.

[10] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[11] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2172–2180, 2016.

[12] Roland S Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

[13] Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR, 2020.

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

[15] Benjamin Culpepper and Bruno Olshausen. Learning transport operators for image manifolds. *Advances in neural information processing systems*, 22, 2009.

[16] Noga Mudrik, Yenho Chen, Eva Yezerets, Christopher J Rozell, and Adam S Charles. Decomposed linear dynamical systems (dlds) for learning the latent components of neural dynamics. *Journal of Machine Learning Research*, 25(59):1–44, 2024.

[17] Ho Yin Chau, Frank Qiu, Yubei Chen, and Bruno Olshausen. Disentangling images with lie group transformations and sparse coding. *arXiv preprint arXiv:2012.12071*, 2020.

[18] Jascha Sohl-Dickstein, Ching Ming Wang, and Bruno A Olshausen. An unsupervised algorithm for learning lie group transformations. *arXiv preprint arXiv:1001.1027*, 2010.

[19] Yenho Chen, Noga Mudrik, Kyle A Johnsen, Sankaraleengam Alagapan, Adam S Charles, and Christopher J Rozell. Probabilistic decomposed linear dynamical systems for robust discovery of latent neural dynamics. *arXiv preprint arXiv:2408.16862*, 2024.

[20] Noga Mudrik, Eva Yezerets, Yenho Chen, Christopher Rozell, and Adam Charles. Linocs: Lookahead inference of networked operators for continuous stability. *arXiv preprint arXiv:2404.18267*, 2024.

[21] William M Boothby. *An introduction to differentiable manifolds and Riemannian geometry, Revised*, volume 120. Gulf Professional Publishing, 2003.

[22] Marissa Connor, Gregory Canal, and Christopher Rozell. Variational autoencoder with learned latent structure. In *International conference on artificial intelligence and statistics*, pages 2359–2367. PMLR, 2021.

[23] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[24] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *arXiv preprint arXiv:2211.11695*, 2022.

[25] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[26] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[27] Tianqi Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2615–2625, 2018.

[28] Horea Muresan and Mihai Oltean. Fruit recognition from images using deep learning. *Acta Universitatis Sapientiae, Informatica*, 10(1):26–42, 2018.