Implicit Neural Clustering

Thomas Kreutz Max Mühlhäuser Alejandro Sanchez Guinea
Telecooperation Lab, Technical University of Darmstadt
{<kreutz, sanchez>@tk, <max>@informatik}.tu-darmstadt.de

Abstract

Recent advances in generative models have revolutionized the controllable synthesis of realistic data for computer vision tasks. Despite their recent impact on tasks like classification and representation learning, their potential in clustering, a related and fundamental unsupervised learning technique, remains largely unexplored. Traditionally, clustering involves partitioning a dataset based on handcrafted or learned feature representations paired with a similarity metric. Our paper turns this paradigm on its head by proposing a different perspective on clustering from the view of implicit generative modeling. We propose the concept of Implicit Neural Clustering, in which clusters are generated implicitly through a generative model that is controllable by disentangled factors of variation. Specifically, we address the challenge of implicit multi-partition clustering, where a dataset may exhibit multiple plausible clusterings representing different class categories. Our contribution is threefold: We introduce a rigorous mathematical definition of Implicit Neural Clustering, propose a straightforward sampling strategy to perform implicit multi-partition clustering, and provide preliminary empirical evidence for the effectiveness of our approach on synthetic data.

1. Introduction

Recent advances in generative models for controllable image synthesis like Generative Adversarial Networks (GANs) (e.g., [3, 14]) and Diffusion models (e.g., [5, 24]) reached a point where synthetic images are so realistic that thy can improve classification performance (e.g., [1, 8]) or help self-supervised representation learning (SSL) methods learn better general purpose embeddings (e.g., [4, 13, 25]).

A closely related task to classification and SSL is clustering, which traditionally partitions a dataset with learned (e.g., SSL) or handcrafted features and a similarity metric, allowing unsupervised classification and data analysis. These approaches almost exclusively follow the prevailing idea that there is only a single correct clustering corresponding to given class labels. However, real-world data often has multiple factors of variation that one could group them by, leading to multi-partition clustering [6, 7, 9, 23, 26, 30, 32]. For instance, animals could be clustered either by species, color, size, or origin. Given that these factors of variation can be used to control the synthesis of realistic images (e.g., [7, 10, 19, 20, 22, 27, 28, 31]), it is beneficial to know (a) what these factors of variation in a dataset are and (b) how we can effectively obtain new synthetic samples from these different clusterings to exploit them for downstream tasks like classification or SSL.

Suppose we have access to a generative model controllable through disentangled representations, which can be composed and the model generalizes to out-of-distribution (OOD) combinations of factors of variation. *Would we be able to generate arbitrary different clusters and objects with combinations of factors of variation implicitly?* Regarding this question, previous works on disentangled representation learning have shown that factors of variation are embedded in single or multiple dimensions of a disentangled feature space [2, 7, 17, 18, 29]. Further, these factors are invariant to others when they are used to control (e.g., via latent traversal) the synthesis of new images [29], generative models can learn to synthesize OOD combinations [19], factors are composable [7], and they can be obtained unsupervised [17, 22, 31] or with supervision [12, 18, 29].

In light of all these recent advances in implicit generative modeling, disentangled representation learning, and realistic image synthesis, instead of a clustering algorithm *explicitly clustering* a dataset into disjoint subsets, this paper takes a step back and looks at clustering the other way around to reveal a totally different perspective: *What if instead of explicitly clustering the elements of a given dataset we would have an implicit model that generates the clusters, i.e., an implicit neural representation of clustering?*

To answer this question, we propose *Implicit Neural Clustering*, a novel clustering approach that uses a generative model conditioned on a disentangled latent space to produce the data of each cluster implicitly. Under the assumption that we can transform elements of a dataset into a disentangled representation comprised of its underlying factors of variation, we can effectively model multi-partition

clusterings of the data and implicitly generate the data for any cluster partitioning. *Implicit Neural Clustering* yields a novel way how to obtain realistic synthetic datasets based on the real factors of variation in real data. Our contribution is threefold: The concept of *Implicit Neural Clustering*, a corresponding mathematically rigorous definition, and, based on that, a sampling strategy showing an application of this concept for *implicit* multi-partition clustering.

2. From Explicit to Implicit Neural Clustering

2.1. Explicit (Multi-Partition) Clustering

Explicit clustering is defined by a partition function C_{sim} that partitions an input data set $\mathcal{D}$ under an arbitrary notion of similarity sim into k clusters that either maps any $x \in \mathcal{D}$ to a hard $C : \mathcal{D} \mapsto \mathbb{N}$ (e.g., k-means) or soft cluster assignment $C : \mathcal{D} \mapsto \mathbb{R}^k$ (e.g., maximum likelihood). Hard clustering can be defined as applying C_{sim} on each $x \in \mathcal{D}$, which yields k disjoint subsets $\mathcal{D}_k^{sim} \subseteq \mathcal{D}$:

$$\mathcal{D} := \bigcup_k \mathcal{D}_k^{sim} = \bigcup_k \{x \mid x \in \mathcal{D} \wedge C_{sim}(x) = k\} \quad (1)$$

with $\bigcup_k \mathcal{D}_k^{sim} = \mathcal{D}$ and $\bigcap_k \mathcal{D}_k^{sim} = \emptyset$. A different $sim' \neq sim$ formally defines any arbitrary clustering different from sim over $\mathcal{D}$, i.e., multi-partition clustering [9].

In the multi-partition clustering context, explicit clustering w.r.t. sim yields only one out of many possible clusterings of the data. However, high-dimensional data such as images typically encompass multiple interesting factors of variation that one could cluster over [7]. As shown in Figure 1, images of objects in a scene could be clustered based on shape, color, or style, where each of these factors of variation reflects a different sim. For multi-partition clustering, sim either corresponds to clustering over different sub-dimensions of the feature representation leading to different clustering partitions [6, 7, 9, 23, 26, 30, 32], or with representation learning, we could also train a different feature extractor for each possible sim. Different from the prevailing concept in representation learning that a representation only disentangles with respect to one factor of variation, in the context of disentangled representation learning, each factor is disentangled in a single (or across multiple) dimension(s) [7, 29] in the disentangled representation. The latter allows multi-partition clustering w.r.t. to sim (any factor of variation) by clustering in isolation on these sub-dimensions [7, 9].

2.2. Definition of Disentangled Representations

In contrast to explicit clustering $\mathcal{D}$ under various sim, *Implicit Neural Clustering* can be derived from a *disentangled* representation $\mathcal{F}$ of $\mathcal{D}$. As an initial intuition, if any $x \in \mathcal{D}$ could be decomposed into its factors of variation, we can impose specific changes to any $x \in \mathcal{D}$ by modifying the desired parts of the factor in the representation.

We quickly recall essential parts of the symmetry group-based definition for disentangled representations by Higgins et al. [11]. Let G be a symmetry group acting on W, W be a set of world states (ground truth factors of variation), O observations (e.g., pixel space), and Z the internal agent representation of W. A generative process $b : W \to O$ leads from world to observation states, and an inference process $h : O \to Z$ leads from observation to an agent's internal representation of W. Following this definition of disentangled representations, we have a dataset $\mathcal{D} = \{o_1, ..., o_N\}$ of observations $o_i \in O$. We now define the inference process $h : O \to Z$ as a parameterized feature extractor [1] $h_\varphi : \mathcal{O} \to \mathcal{F}$ with parameters φ that yields a disentangled representation $\mathcal{F}$ of any $o \in \mathcal{D}$, where $\mathcal{F}$ decomposes into its factors of variation $\mathcal{F} = \mathcal{F}_1 \times ... \times \mathcal{F}_M$. In addition, we have access to a parameterized generator $G_\theta : F \to O$ with parameters θ that transforms samples from the disentangled representation space $\mathcal{F}$ to the observation space O.

2.3. Implicit Neural Clustering

Figure 1 provides an overview of *Implicit Neural Clustering* with its main differences to explicit (multi-partition) clustering. Following the definition of implicit probabilistic models, *Implicit Neural Clustering* can be defined as a special sampling procedure from a disentangled compositional latent space. Different from implicit models where a parameterized generator $G_\theta(\cdot)$ (e.g., GAN) transforms samples from an analytic distribution (e.g., isotropic Gaussian) to synthetic examples [16], *Implicit Neural Clustering* transforms samples from a disentangled distribution $\mathcal{F}$ into synthetic clusters $\mathcal{D}_k^{sim}$. More specifically, for each cluster $\mathcal{D}_k^{sim}$, there exists an implicit cluster that can be obtained by sampling from $\mathcal{F}$ while fixing one respective factor of variation $\mathcal{F}_k$. Let $G = G_1 \times ... \times G_M$ be the group actions that act on $\mathcal{F}$, and $\cdot : G \times \mathcal{F} \to \mathcal{F}$ the action that changes $\mathcal{F}$ to the respective factor. Given that each factor of variation consists of several *atomic* attributes (i.e., class labels like shape, color, or animal species), we precisely define fixing a factor of variation as follows: Each G_i consists of *atomic* partitions $G_i = \{G_{i,1}, G_{i,2}, ...\}$, that can be parameterized by a single value $f \in \mathbb{R}$ or a parameterized distribution $P(f|\phi)$. We further define a function $\overset{G_{ij}}{=}$ that yields true if an atomic factor of variation G_{ij} is present in $z \in \mathcal{F}$.

When clustering with respect to a factor of variation $\mathcal{F}_i$, let $sim \equiv \mathcal{F}_i$. Based on Equation 1, explicit clustering $C_{\mathcal{F}_i}$ splits $\mathcal{D}$ into $|G_i|$ disjoint subsets. In the *implicit* case, together with the generator G_θ, the feature extractor h_φ, we can (a) generate each cluster $\mathcal{D}_{ij}$ implicitly, and (b) generate

[1] We change Z to $\mathcal{F}$ for notation and readability reasons.

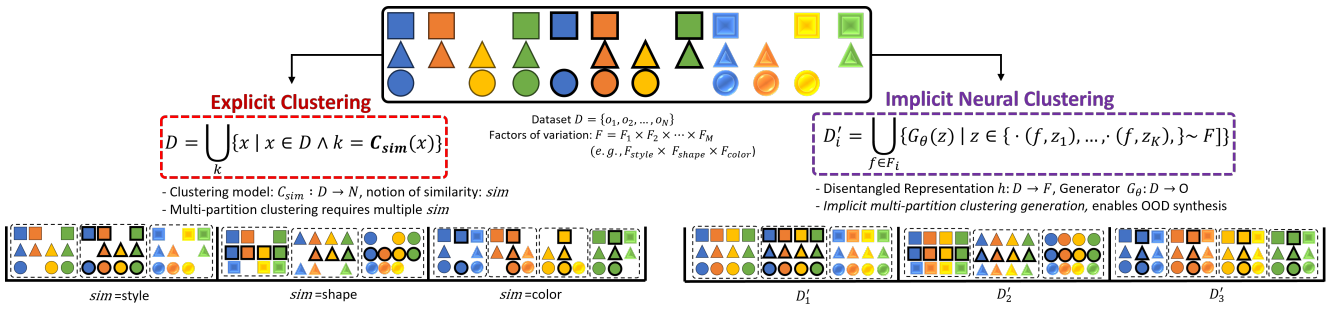

Figure 1. We visualize the difference between multi-partition clustering in the traditional sense and *Implicit Neural Clustering*. Under different notions of similarity, explicit clustering can cluster the dataset correctly in three different partitionings. However, since not all possible combinations between factors of variation are observed in the data, certain combinations are not present in the final clusters because we only *explicitly cluster* the real data. In contrast, *Implicit Neural Clustering* leads to *implicit clusters* that in addition include realistic examples not observed in the underlying dataset, and can from definition also synthesize out-of-distribution cross-combinations.

a synthetic version of the original dataset D as follows.

$$\mathcal{D} \approx \mathcal{D}' := \bigcup_{f \in G_i} \mathcal{D}_{ij} = \bigcup_{f \in G_i} \{G_\theta(h(o)) \mid o \in \mathcal{D} \wedge \overset{f}{=} (h(o))\} \tag{2}$$

In this way, $\mathcal{D}'$ *implicitly* models $\mathcal{D}$ with respect to a clustering under a factor of variation $\mathcal{F}_i$. On the basis of Equation 2, we derive *Implicit Neural Clustering* based on a very strong assumption. If we assume the disentangled representation space $\mathcal{F}$ to be composable, we can modify any $z \in \mathcal{F}$ by acting with the atomic group action G_{ij}, to change the factor of variation and cluster membership under factor $\mathcal{F}_i$ from any previous D_{il} to $D_{ij}, l \neq j$. Together with a sampling procedure $(\cdot) \sim \mathcal{F}$ for each factor of variation $\mathcal{F}_i$, *Implicit Neural Clustering* is defined as follows.

$$D' = \bigcup_{f \in G_i} \{G_\theta(z) \mid z \in \{\cdot(f, z_1), ..., \cdot(f, z_K)\} \sim \mathcal{F}\} \tag{3}$$

where we fix a factor of variation, sample random representations from $z \in F$, take the respective group action f of an atomic factor of variation[2], and modify each z accordingly with $\cdot(f, z)$. Under the respective definition of disentangled representations, the resulting synthetic dataset D' is now partitioned into a clustering w.r.t. a certain factor of variation. As a result, up to the capabilities of the encoder h_φ and generator G_θ, *Implicit Neural Clustering* can synthesize *any* cluster that could be obtained by a explicit clustering for synthetic dataset generation.

Sampling Procedure. Algorithm 1 outlines the sampling procedure used for implicit clustering based on some factor of variation F_i. For this procedure, we must first obtain

Algorithm 1: Algorithm of the proposed method

Input: Group actions G_i for factor of variation F_i, Generator G_θ, number of samples K
Output: Implicit clustering $\mathcal{D}'$ with respect to F_i

1 $D' \leftarrow \emptyset$
2 **for** *each* $f \in G_i$ **do**
3 $\quad D_{ij} \leftarrow \emptyset$
4 $\quad$ **for** k *in* $1..K$ **do**
5 $\quad\quad f \sim G_{ij}$
6 $\quad\quad z = (z_1, z_2, ..., z_d) \sim \{G_1, G_2, ..., G_M\}$
7 $\quad\quad z' = \cdot(f, z)$
8 $\quad\quad D_{ij} \leftarrow D_{ij} \cup \{G_\theta(o')\}$
9 $\quad D' \leftarrow D' \cup D_{ij}$

the atomic group actions G_{ij}. We assume factors of variation to be disentangled dimension-wise, i.e., in only one dimension l of the representation $(z_1, z_2, ..., z_d) \in \mathbb{R}^d$. In practice, the definition that F decomposes into only its factors is too restrictive. Therefore, we relax this definition so that $\mathcal{F}$ decomposes into dimensions with and without factors of variation, $z = (z_1, z_2, ..., z_d) \in \mathbb{R}^d$, $d > M$, and we identify which z_l corresponds to any G_{ij}. To this end, we first encode the full dataset and then partition each dimension using kernel density estimation (KDE) together with local minima on the resulting density estimates[3]. Afterward, for each partition in each dimension we identify unique co-occurences with the known ground truth factors of variation. Finally, we turn each G_{ij} into a parameterized probability distribution $P(f|\phi)$ (e.g., Uniform or Normal), which allows sampling for more variety.

With the atomic group actions, the specific sampling procedure is defined in lines 5-8 in Algorithm 1. We

[2]In practice, we would parameterize f with a probability distribution and sample the respective modification for more variety, but a single value, like the mean over all possible values, would also work.

[3]Any partition algorithm could be used. KDE has the advantage over, e.g., k-means that we do not specify the number of partitions in advance.

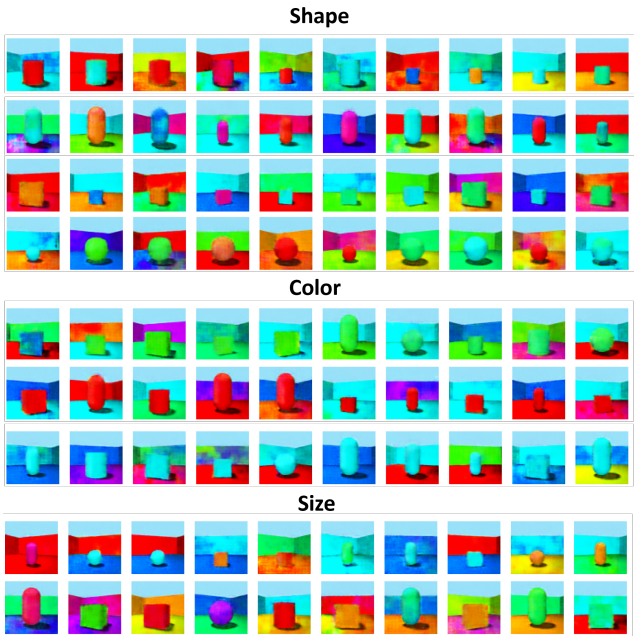

Figure 2. *Implicit Neural Clustering* of Shapes3D. Our approach can implicitly cluster the dataset into shape, different object colors, and size. Each row represents random samples for an atomic factor of variation that we were able to disentangle. Each row is the result of applying atomic group actions we extracted from the disentangled representation space to random samples.

first sample a random value from the partition distribution $f \sim P(f|G_{ij})$. Next, we sample random values from all partitions to obtain a random latent $z \in \mathcal{F}$. We then act with f on z, i.e. $\cdot(f, z)$, which modifies z accordingly. For *Implicit Neural Clustering*, this process is repeated K times for each cluster F_{ij}, and *implicit* multi-partition clustering is achieved by repeating the Algorithm 1 with different F_i.

3. Experiments

We evaluate our approach on the popular 3DShapes Dataset [15], which includes 6 ground truth factors of variation. For a strong basis, we train a weakly-supervised ADA-GVAE [18] that is known to work well with this dataset. The encoder of the VAE serves as the encoder h_φ, and the decoder as the generator G_θ. We evaluate the finding of atomic group actions, generating implicit clusters with these atomic group actions, and compositionality in a qualitative manner. In this evaluation, we have access to the ground truth factors of variation. In an unsupervied setting, one could potentially fall back to pseudo-labels with a zero-shot classifier instead (e.g., using CLIP [21]).

Implicit Neural Clustering of Shapes3D. Figure 2 shows three coherent synthetic multi-partition clusters with respect to shape, color, and size that we have *implicitly* clustered with our approach. They are the result of applying the atomic group actions we were able to extract from the disentangled representations to random samples. We specifically show that we can modify arbitary samples with the atomic group actions in Figure 3, where we show that randomly generated samples are modified to the desired atomic factor of variation. However, even though we find atomic group actions for all ground truth labels, not all of them are invariant to the other factors of variation, showing a limitation in the disentanglement learned by the ADA-GVAE for some factors of variation. In summary, our experiment on synthetic data shows that when the assumptions we make for our approach hold for a desired factor of variation, we can implicitly cluster a dataset with *Implicit Neural Clustering*.

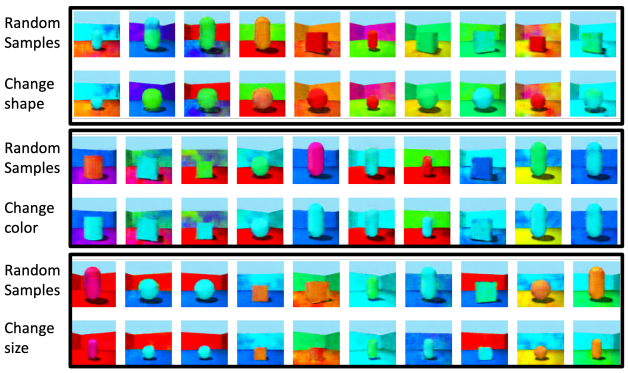

Figure 3. For random generated samples, we can specifically modify a certain factor of variation by applying the respective atomic group action, e.g., change the object shape,color, or size.

4. Conclusion

We propose the concept of *Implicit Neural Clustering* together with a mathematically rigorous definition and sampling procedure for implicit (multi-partition) clustering. We provide preliminary empirical evidence that our approach can be an effective and efficient way for controllable synthetic data generation when known factors of variation are disentangled. Regarding future work, we plan to evaluate our approach on more datasets, including real-world data, for stronger empirical evidence and with an empirical evaluation on downstream tasks such as classification and self-supervised representation learning. We strongly believe that our approach can serve as a foundation to generate any kind of clusterings imposed by factors of variation in real datasets, enabling effective synthesis of extended training datasets with varying class labels, more balanced class distributions, and data variations not observed in the original data (including out of distribution) in the future.

Acknowledgement

This work has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY centre; and partially funded by the Federal Ministry of Education and Research (BMBF) under grant 01|S17050 as part of the Software Campus project "UnErObVe".

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 1

[2] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164*, 2021. 1

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 1

[4] Lucy Chai, Jun-Yan Zhu, Eli Shechtman, Phillip Isola, and Richard Zhang. Ensembling with deep generative views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14997–15007, 2021. 1

[5] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[6] Marie du Roy de Chaumaray and Vincent Vandewalle. Non-parametric multi-partitions clustering. *arXiv preprint arXiv:2301.02422*, 2023. 1, 2

[7] Fabian Falck, Haoting Zhang, Matthew Willetts, George Nicholson, Christopher Yau, and Chris C Holmes. Multi-facet clustering variational autoencoders. *Advances in Neural Information Processing Systems*, 34:8676–8690, 2021. 1, 2

[8] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. *arXiv preprint arXiv:2312.04567*, 2023. 1

[9] Giuliano Galimberti and Gabriele Soffritti. Model-based methods to identify multiple cluster structures in a data set. *Computational statistics & data analysis*, 52(1):520–536, 2007. 1, 2

[10] Markos Georgopoulos, James Oldfield, Grigorios G Chrysos, and Yannis Panagakis. Cluster-guided image synthesis with unconditional models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11543–11552, 2022. 1

[11] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. 2

[12] Yordan Hristov, Alex Lascarides, and Subramanian Ramamoorthy. Interpretable latent spaces for learning from demonstration. In *Conference on Robot Learning*, pages 957–968. PMLR, 2018. 1

[13] Ali Jahanian, Xavier Puig, Yonglong Tian, and Phillip Isola. Generative models as a data source for multiview representation learning. *arXiv preprint arXiv:2106.05258*, 2021. 1

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 1

[15] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018. 4

[16] Ke Li and Jitendra Malik. Implicit maximum likelihood estimation. *arXiv preprint arXiv:1809.09087*, 2018. 2

[17] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019. 1

[18] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International conference on machine learning*, pages 6348–6359. PMLR, 2020. 1, 4

[19] Maya Okawa, Ekdeep S Lubana, Robert Dick, and Hidenori Tanaka. Compositional abilities emerge multiplicatively: Exploring diffusion models on a synthetic task. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[20] Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. *arXiv preprint arXiv:2001.10238*, 2020. 1

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[22] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. *arXiv preprint arXiv:2102.10543*, 2021. 1

[23] Fernando Rodriguez-Sanchez, Concha Bielza, and Pedro Larrañaga. Multipartition clustering of mixed data with bayesian networks. *International Journal of Intelligent Systems*, 37(3):2188–2218, 2022. 1, 2

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[25] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[26] Vincent Vandewalle. Multi-partitions subspace clustering. *Mathematics*, 8(4):597, 2020. 1, 2

[27] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In

International conference on machine learning, pages 9786–9796. PMLR, 2020. 1

[28] Shiyu Wang, Yuanqi Du, Xiaojie Guo, Bo Pan, Zhaohui Qin, and Liang Zhao. Controllable data generation by deep learning: A review. *arXiv preprint arXiv:2207.09542*, 2022. 1

[29] Xin Wang, Hong Chen, Si'ao Tang, Zihao Wu, and Wenwu Zhu. Disentangled representation learning. *arXiv preprint arXiv:2211.11695*, 2022. 1, 2

[30] Matthew Willetts, Stephen Roberts, and Chris Holmes. Disentangling to cluster: Gaussian mixture variational ladder autoencoders. *arXiv preprint arXiv:1909.11501*, 2019. 1, 2

[31] Tao Yang, Yuwang Wang, Yan Lv, and Nanning Zheng. Disdiff: Unsupervised disentanglement of diffusion probabilistic models. *arXiv preprint arXiv:2301.13721*, 2023. 1

[32] Nevin L Zhang. Hierarchical latent class models for cluster analysis. *The Journal of Machine Learning Research*, 5:697–723, 2004. 1, 2