CLeaR 2023 Datasets Track 1–7, 2023

3DIdentBox: A Toolbox for Identifiability Benchmarking

Alice Bizeul ALICE.BIZEUL@INF.ETHZ.CH
Department of Computer Science, ETH Zurich
ETH AI Center, ETH Zurich

Imant Daunhawer IMANT.DAUNHAWER@INF.ETHZ.CH
Department of Computer Science, ETH Zurich

Emanuele Palumbo EMANUELE.PALUMBO@INF.ETHZ.CH
Department of Computer Science, ETH Zurich
ETH AI Center, ETH Zurich

Bernhard Schölkopf BERNHARD.SCHOELKOPF@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen

Alexander Marx* ALEXANDER.MARX@AI.ETHZ.CH
Department of Computer Science, ETH Zurich
ETH AI Center, ETH Zurich

Julia E. Vogt* JULIA.VOGT@INF.ETHZ.CH
Department of Computer Science, ETH Zurich

Abstract

In this paper, we present 3DIDENTBOX, a collection of 12 synthetic multi-view datasets, that offers a versatile toolbox for identifiability benchmarking. As a natural extension of prior work (Zimmermann et al., 2021; von Kügelgen et al., 2021) to the multi-view setting, 3DIDENTBOX features high-dimensional image pairs of a 3D-scene with corresponding ground-truth generative factors that vary across datasets and views. Moreover, the included data-generating code offers flexibility for generating custom multi-view datasets through simple modifications of the underlying generative causal model, which is useful for creating interventions and distribution shifts. Altogether, 3DIDENTBOX provides useful resources for identifiability benchmarking and for causal representation learning more generally.

Keywords: Multi-view Learning, Identifiability Benchmarking, Causal Representation Learning

1. Introduction

Causal representation learning aims to discover the causal factors that generate observed data as well as their relationship (Schölkopf et al., 2021). The underlying assumption is that understanding the causal factors can lead to a better understanding on the data generative process and improve the performance of downstream tasks such as classification, regression, or decision-making. Therefore, the goal is to learn robust and interpretable causal representations of the data, which can fully capture the underlying causal structure, or at least exploit properties induced by the causal structure to improve, e.g., the fairness of a model (Karimi et al., 2022), or its out of distribution performance (Schölkopf et al., 2021).

Although there exist a rich set of real-world datasets to evaluate the downstream performance, such as out-of-distribution generalization (Koh et al., 2021; Gulrajani and Lopez-Paz, 2021), or

*Joint authorship.

fairness of a model (Karimi et al., 2022), it is difficult to impossible to draw definite conclusions about the identification of ground truth causal factors in real-world settings. Synthetic data generation with full control over the generating factors is therefore crucial to systematically analyze which ground truth causal factors can be identified by a model. For example, in out-of-distribution generalization, modified versions of MNIST (LeCun et al., 1998) such as rotated, or colored MNIST (Gulrajani and Lopez-Paz, 2021) or subsampled versions of dSprites (Matthey et al., 2017) are used (Xu et al., 2021; Chen et al., 2021). However, the adaptation of these datasets to new tasks is tedious as they do not allow full control over the generative factors.

In this work, we present 3DIDENTBOX, an ensemble of datasets comprising image pairs and associated ground-truth generative factors. It gains inspiration from real-world multi-view datasets by offering pairs of high-resolution images (224×224px) depicting a common scene: a colored teapot in front of a colored background illuminated by a colored spotlight. Similarly to the real-world scenario where two recording devices would capture a scene with some level of variability, each image renders the teapot with a specific texture. 3DIDENTBOX is an extension of the 3DIdent (Zimmermann et al., 2021) and Causal3DIdent (von Kügelgen et al., 2021) datasets. While both 3DIdent and Causal3DIdent already offer full control over the generative factors of a similar scene, our work extends these contributions to the multi-view setting by keeping some of the generative factors specific to one view only. To offer more setting diversity, 3DIDENTBOX encompasses 12 datasets with distinct generative assumptions detailed in Section 2. Furthermore, we provide full access to the data generator, which allows other researchers to modify the causal relationships among the latent variables and generate new images (e.g., including a distribution shift) based on their modified structural causal model.

We provide the code used to recreate the 3DIDENTBOX datasets and to extend the generation to custom settings in our Github repository.* The datasets *without* and *with* causal dependencies between generative factors can be downloaded on Zenodo.†‡ Combined, the available datasets and code base constitutes a toolbox, named 3DIDENTBOX, for the benchmarking of identifiability results. A subset of the functionality of 3DIDENTBOX has already been featured in (Daunhawer et al., 2023) to evaluate the identification of latent factors in multimodal contrastive learning.

2. Dataset Description

The 3DIDENTBOX datasets provide image pairs depicting two views of a colored teapot on a colored background illuminated by a colored spotlight. Each scene is fully controlled by nine so-called generative factors. While three factors are systematically shared between images within the same pairs (i.e., content factors, c), the remaining factors are either stochastic between views (i.e., style factors, s) or specific to one view only (i.e., view-specific factors, m). The image rendering process is deterministic,

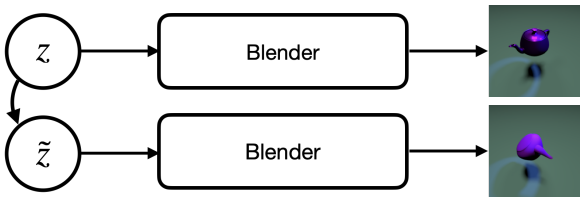

Figure 1: Overview of the data generative process. Generative factors, referred to as $z = [c, s, m], \tilde{z} = [\tilde{c}, \tilde{s}, \tilde{m}]$, are inputs to the *Blender* image rendering software.

*https://github.com/alicebizeul/3DIdentBox.
†3DIDENTBOX part 1 :https://doi.org/10.5281/zenodo.7699779
‡3DIDENTBOX part 2: https://doi.org/10.5281/zenodo.7699773

relies fully on the aforementioned generative factors and is supported by the Blender Online Community (2018) software. An overview of the image generation process is presented in Figure 1, and additional examples of images from the 3DIDENTBOX datasets are provided in Figure 2.

The following sections aim at describing the provided datasets in detail, in particular, their generative factors and the image rendering procedure.

2.1. Generative Factors

Table 1 summarizes the list of factors defining each observation. Information sources are partitioned into three blocks: object & spotlight rotation angles, object position, and scene color. Each block comprises three generative factors. Pairs of generative factors, $\{z, \tilde{z}\}$, are structured according to a set of rules: content information, $\{c, \tilde{c}\}$, is shared across views, style information, $\{s, \tilde{s}\}$, describes factors which are stochastically shared (i.e., $\tilde{s}$ is a perturbed version of s), and view-specific information, $\{m, \tilde{m}\}$, is partitioned such that each factor is specific to a view. The causal graphical model governing the generation of each scene according to the content, style, and view-specific partition is depicted in Figure 3. Daunhawer et al. (2023) provides a more detailed description of this generative model. Table 2 reports the distributions used to structure content, style, and view-specific pairs of factors. Two scenarios are considered. The former does not consider any causal dependencies between generative factors whilst the latter considers causal dependencies between content factors, style factors, and between content and style information.

Each dataset features a unique combination of generative rules for each information block. As a consequence, 3DIDENTBOX encapsulates a total of 12 datasets (6 without and 6 with inter-factor causal dependencies). Figure 3 shows a specific combination of information block and generative rules for one of the provided datasets: scene hues, object position, and object & spotlight rotation are content, style, and view-specific information respectively. The causal dependencies between factors are defined by the following structural causal model:

$$\begin{aligned} c_1 &:= c_2 + \epsilon_{c_2} & \tilde{s_1} &:= s_1 + \epsilon_{s_1} \\ s_1 &:= s_2 + \epsilon_{s_2} & \tilde{s_2} &:= s_2 + \epsilon_{s_2} \\ s_3 &:= c_3 + \epsilon_{c_3} & \tilde{s_3} &:= s_3 + \epsilon_{s_3} \end{aligned} \tag{1}$$

where ϵ_{i_j} follow a normal distribution truncated to the $[-1 - i_j, 1 - i_j]$ interval with i_j, the realisation of the corresponding variable. The remaining variables, $c_2, c_3, s_2, \tilde{c_1}, \tilde{c_2}, \tilde{c_3}$ are drawn from a uniform distribution in the $[-1, 1]$ interval as mentioned in Table 2. View-specific variables are drawn from δ distributions. In this specific configuration, object color, spotlight color, background color, object x-coordinate, object y-coordinate, and object z-coordinate are $c_1, c_2, c_3, s_1, s_2, s_3$ respectively.

2.2. Image Rendering

Blender is a sophisticated 3D creation software that empowers users to render intricate visual scenes with complete control over all featured elements. The software is used to generate 3DIDENTBOX images at a 224x224x3 resolution. The rendering serves as a complex invertible mapping that generates the images from the set of 9 factors presented in section 2.1. Similar to a scene captured by two distinct recording devices, the object texture is specific to each view. In the first view, the

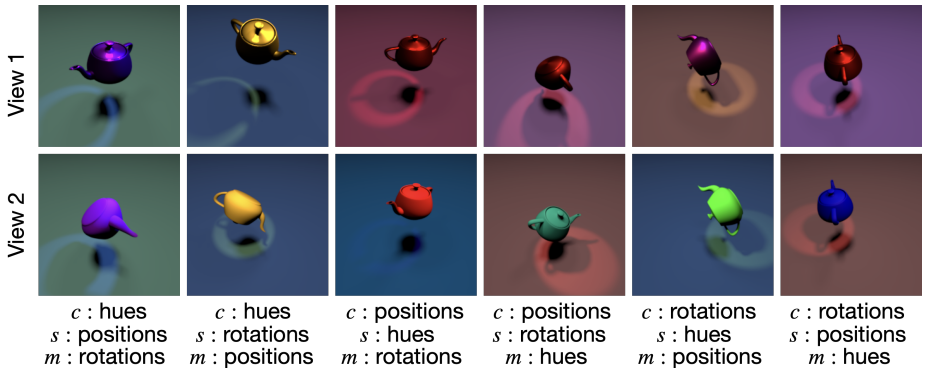

Figure 2: Example of image pairs for each 3DIDENTBOX dataset with inter- and intra- block causal dependencies. Each pair is characterised by a unique information block (content-style-view-specific) and information type (scene hues-object position- object/spotlight rotation) combination, described below each pair.

Block	Factor Description	Raw Support	Blender Support	Visual Range
Rotation	Object α-rotation angle	$[-1, 1]$	$[-\pi, \pi]$	$[0°, 360°]$
Rotation	Object β-rotation angle	$[-1, 1]$	$[-\pi, \pi]$	$[0°, 360°]$
Rotation	Spotlight rotation angle	$[-1, 1]$	$[-\pi, \pi]$	$[0°, 360°]$
Position	Object x-coordinate	$[-1, 1]$	$[-2, 2]$	-
Position	Object y-coordinate	$[-1, 1]$	$[-2, 2]$	-
Position	Object z-coordinate	$[-1, 1]$	$[-2, 2]$	-
Hues	Object HSV color (H param.)	$[-1, 1]$	$[-\pi, \pi]$	$[0°, 360°]$
Hues	Spotlight HSV color (H param.)	$[-1, 1]$	$[-\pi, \pi]$	$[0°, 360°]$
Hues	Background HSV color (H param.)	$[-1, 1]$	$[-\pi, \pi]$	$[0°, 360°]$

Table 1: Description of 3DIDENTBOX generative factors. To match Blender's requirements the raw support, $[-1, 1]$, is converted to the *Blender Support*. The rotation and HSV hue ranges observed on generated images are reported in the *Visual Range* column.

Type	Symbol	View 1 Distribution	View 2 Distribution
Content	$c = [c_1, c_2, c_3]$	$c \sim [\mathcal{U}(a,b), \mathcal{U}(a,b), \mathcal{U}(a,b)]$	$\tilde{c} \sim [\delta(c_1), \delta(c_2), \delta(c_3)]$
Style	$s = [s_1, s_2, s_3]$	$s \sim [\mathcal{U}(a,b), \mathcal{U}(a,b), \mathcal{U}(a,b)]$	$\tilde{s} \sim [\mathcal{N}_t(s_1, 1), \mathcal{N}_t(s_2, 1), \mathcal{N}_t(s_3, 1)]$
View-specific	$m = [m_1, m_2, m_3]$	$m \sim [\mathcal{U}(a,b), \delta(0), \delta(0)]$	$\tilde{m} \sim [\delta(0), \delta(0), \mathcal{U}(a,b)]$

Type	Symbol	View 1 Distribution	View 2 Distribution
Content	$c = [c_1, c_2, c_3]$	$c \sim [\mathcal{N}_t(c_2, 1), \mathcal{U}(a,b), \mathcal{U}(a,b)]$	$\tilde{c} \sim [\delta(c_1), \delta(c_2), \delta(c_3)]$
Style	$s = [s_1, s_2, s_3]$	$s \sim [\mathcal{N}_t(s_2, 1), \mathcal{U}(a,b), \mathcal{N}_t(c_3, 1)]$	$\tilde{s} \sim [\mathcal{N}_t(s_1, 1), \mathcal{N}_t(s_2, 1), \mathcal{N}_t(s_3, 1)]$
View-specific	$m = [m_1, m_2, m_3]$	$m \sim [\mathcal{U}(a,b), \delta(0), \delta(0)]$	$\tilde{m} \sim [\delta(0), \delta(0), \mathcal{U}(a,b)]$

Table 2: Description of the distributions used for the generation of 3DIDENTBOX generative factors pairs, (*top*) *without* inter-factor causal dependencies, (*bottom*) *with* inter-factor causal dependencies. a, b are set to $-1, 1$ respectively. $\mathcal{N}_t$ refers to a normal distribution truncated to the interval $[-1, 1]$.

teapot is rendered with a metallic texture while in the second the object is displayed with a rubber texture. Train/validation/test partitions of the 3DIDENTBOX datasets comprise 250,000/10,000/10,000 samples respectively.

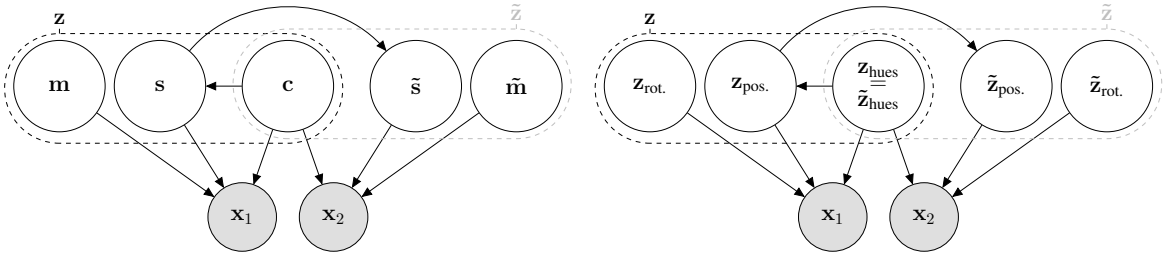

Figure 3: Graphical causal model illustrating the generative process behind 3DIDENTBOX datasets. (*left*) general formulation, (*right*) example for the case when scene hues, object positions and object & spotlight rotation follow content, style, and view-specific generation rules respectively.

3. Conclusion

In this paper, we introduced 3DIDENTBOX, a toolbox comprising a code base and a collection of 12 multi-view synthetic datasets. We believe 3DIDENTBOX offers a valuable tool for identifiability benchmarking, causal representation learning, and multi-view learning. In addition, our framework provides the possibility for users to easily adjust the causal graphical models to custom settings, allowing for the design of controlled interventions, and leading to additional benefits in the investigation of, e.g., the impact of distribution shifts or confounding. Prospective directions include the exploration of more non-linear structural causal models and the extension of our framework to the generation of videos featuring fluid transitions between latent states.

Acknowledgments

AB, EP and AM were supported by ETH AI Center doctoral/postdoctoral fellowships. ID was supported by the SNSF grant *#200021-188466*. EP was supported by the grant *#2021-911* of the Strategic Focal Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain (Swiss Federal Institutes of Technology). The datasets were generated on the ETH Zurich Leonhard cluster.

References

Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. URL http://www.blender.org.

Xinyang Chen, Sinan Wang, Jianmin Wang, and Mingsheng Long. Representation subspace distance for domain adaptation regression. In *International Conference on Machine Learning*, pages 1749–1759, 2021.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations*, 2023.

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

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.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021.

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*, 2021.