# ProDAS: Probabilistic Dataset of Abstract Shapes

Jens Müller, Lynton Ardizzone, Ullrich Köthe

December 8, 2023

**Abstract**

We introduce a novel and comprehensive dataset, named ProDAS, which enables the generation of diverse objects with varying shape, size, rotation, and texture/color through a latent factor model. ProDAS offers complete access and control over the data generation process, serving as an ideal environment for investigating disentanglement, causal discovery, out-of-distribution detection, and numerous other research questions. We provide pre-defined functions for the important cases of creating distinct and interconnected distributions, allowing the investigation of distribution shifts and other intriguing applications. The library can be found at https://github.com/XarwinM/ProDAS.

## 1 ProDAS – Probabilistic Dataset of Abstract Shapes

*Probabilistic Dataset of Abstract Shapes (ProDAS)* is a versatile library that provides a customizable latent factor model applicable to any rendering function. This is schematically illustrated in Figure 1. The library consists of two parts: Firsly, a customizable latent factor model. For instance, the library enables defining a distribution over object types (e.g. squares or triangles), their color and texture as well as the background. Samples drawn from this distribution can be processed through a renderer to generate the final images. As this distribution is predefined, we can evaluate the likelihood of the rendered images. The library offers also support managing multiple different distributions at the same time, for instance in-distribution and out-of-distributions, or different environments as in Figure 2 or Figure 3.

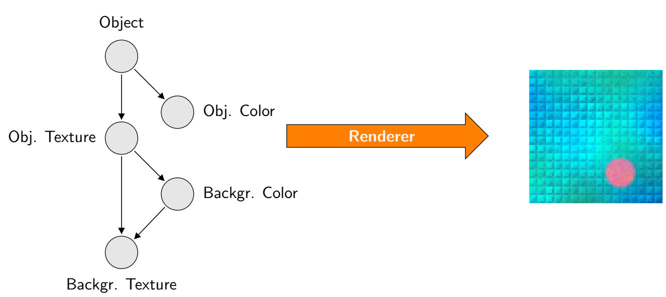



Figure 1: An illustration of ProDAS. A distribution over a latent variable model is defined. From this distribution, an instance is sampled, comprising attributes such as object color and texture. This instance is then processed through the renderer, resulting in data accessible to models. Consequently, ProDAS facilitates the sampling of high-dimensional complex data for which the ground truth is known.

As the second component of the library, ProDAS provides "Dsprites++" as a default rendering frontend, supporting colors, textures, and more. By default ProDAS offers different shapes similar to Dsprites [MHHL17], supporting also colors and textures. The sensible default that ProDAS provides encompasses the following variables:

- object shape $o_{\text{shape}} \in \{\text{circle}, \text{square}, \text{triangle}\}$
- object size $o_{\text{size}} \in \mathbb{R}_{>0}$

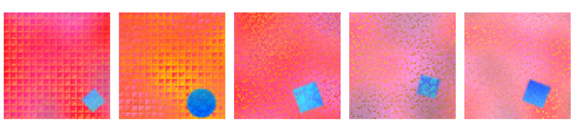
- object position $o_{\text{position}} \in [a, b]^2$
- object rotation $o_{\text{rotation}} \in [0, 360]$
- object and background color (e.g., in RGB)
- 9 different foreground and background textures

## 2 Target Applications

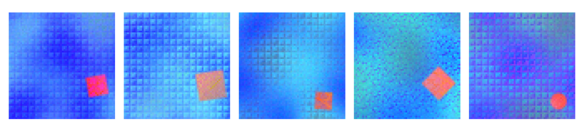
ProDAS offers the capability to alter both the latent model and rendering functions, enabling the creation of numerous intriguing applications and scenarios. In the following section, we introduce four specific scenarios with the default rendering function. For more challenging applications,the factor model can be applied to a different rendering function such as VirtualKitti [GWCV16], Carla [DRC+17], etc.

**Causal Discovery in Latent Space** Given complete access to the latent factors that generate the data, we can presume a latent causal model, thereby delving into the task of causal discovery. In the realm of causal discovery, there is often the assumption that the relevant causal variables are predetermined. However, in our scenario, we consider the task of identifying the underlying causal graph from variables devoid of intrinsic meaning, such as pixels. For instance, the task might involve uncovering the latent causal model as depicted in Figure 1. Additionally, this setting aligns with the broader objective of the *disentanglement* task where the latent factors are often assumed to be jointly independent.

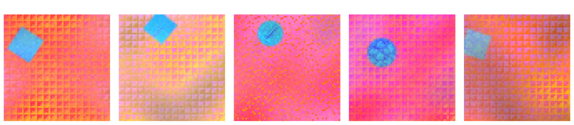
**Out-of-distribution, but why?** In Figure 2 we showcase three different out-of-distribution (OOD) scenarios with respect to the *color*, *position*, or *shape* of the objects. In this case, we have created a scenario where one could evaluate an OOD detection algorithm under different conditions. With these scenarios, we enable OOD detection algorithms to show their capabilities in the field of explainability: The samples in Figure 2 are OOD for different reasons. Some samples are OOD due to low-level features such as color and some are OOD due to more high-level features such as position. Therefore we can assess a model's ability to understand differences between OOD-ness.



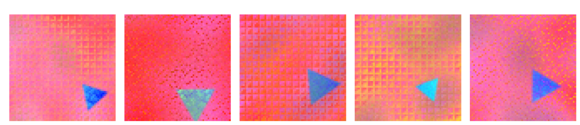
(a) Training Data.



(b) OOD due to Color.



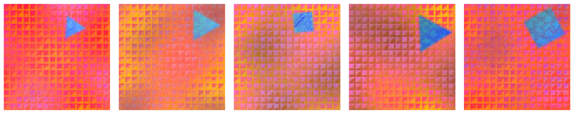
(c) OOD due to Position.



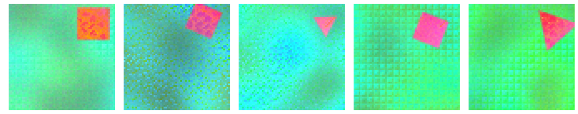
(d) OOD due to shape

Figure 2: ProDAS enables the support for multiple distributions concurrently. In this figure, we implemented one in-distribution scenario and three distinct out-of-distribution (OOD) situations.

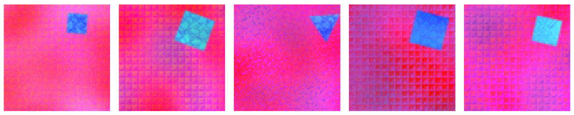
**Distribution Shift** Furthermore, we can explore scenarios involving distribution shifts. In Figure 3, we defined four domains, each sharing identical characteristics except for their varying appearance across domains. This example showcases the potential for exploring distribution shifts, which can manifest in different forms.



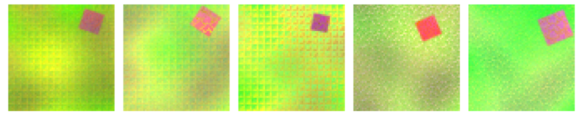
(a) Domain 1



(b) Domain 2



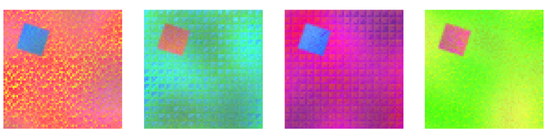
(c) Domain 3



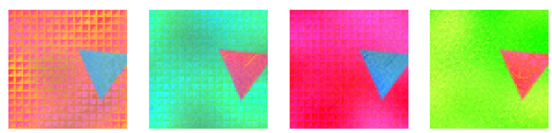
(d) Domain 4

Figure 3: ProDAS provides support for multiple distributions simultaneously. In this figure, we implemented various domains to simulate a distribution shift setting.

**Multi-View Learning** In a multi-view setting, practitioners have access to various representations (also called multiple views) of the same instance. For example, a book translated into multiple languages provides multiple views of the same content. Likewise, within ProDAS, we can establish a multi-view setting. For instance in Figure 4, we observe the same objects in multiple views.



(a) Multiple views of one instance



(b) Multiple views of one instance

Figure 4: This figure shows two instances in four views.

**Many More** In addition to the demonstrated application, ProDAS offers a wide range of potential applications. These encompass tasks such as disentanglement, domain transfer, domain adaptation, few-shot learning, density estimation, among many others.

## Acknowledgments

JM and UK were supported by Informatics for Life funded by the Klaus Tschira Foundation. We thank Martin Rohbeck for fruitful discussions.

# References

[DRC+17] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017.

[GWCV16] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

[MHHL17] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.