# 1 Prerequisites on Learning Disentanglement

## 1.1 Theoretical Impediments from Causality

For us, generative factors represent causal elements. Learning a disentangled representation is then understood as causal inference from image data. In accordance with the causal literature, we can make statements about the type of knowledge, that can be gained by the type of data provided. It turns out that from "raw" image data, raw data meaning images $x$ sampled from $p(x)$, without further assumptions, it is impossible learn a disentangled representation $z$. We start with a short intro to causal learning (1.1.1), the conclusions for disentangling (1.1.2),

### 1.1.1 Causal Learning

Learning to infer causality is harder than statistical learning, as in addition to the probability distribution also the causal structure has to be inferred.

We will start with an example: How to learn the connection between a barometer and the weather from data? If the barometer is working well, there exists a clear correlation between the precipitation and the needle position. A highly capable machine learning algorithm that learns only with access to an image dataset showing the barometer and the weather will be able to capture the correlation between needle position and weather condition, but never understand the causal direction, since this is not in the data. Imagine how a human would go about solving this problem: Having a mechanistic model of the world he could reason about the precise causal mechanism relating weather to humidity to needle position. For example a model of influences (humidity -> barometer) What if he has no prior knowledge? A child-level simple solution is to force the needle to move with a finger. The weather will not change. Hence causality has to go other way or via a third latent variable influencing both.

Pearl [1] distinguishes between three types of questions that can be answered by different types of knowledge:

1. Association.

2. Intervention.

3. Counterfactual. Particularly problematic to learn from data: data consist of facts, not counterfacts.

### 1.1.2 Disentangling requires Interventions or Model Assumptions

This means that "purely" unsupervised disentangling, *i.e.* estimating $\hat{\mathbf{z}}_\mathbf{i}$ from samples $x \sim p(x)$, is impossible. A rigorous argument for this can be found in [2]. Current machine learning operates mostly on the level of association, estimating (complex) correlations from raw data. As we have seen, this purely data-driven approach can only go so far. In contrast, humans seem to have the ability to interact with their environment and have innate assumptions on coherence, causality, physics etc., which introduce inductive priors. To bring *i)* interventions and *ii)* model assumptions to computer vision, we *i)* apply changes to an image, which are assumed to change only one factor and *ii)* model the causal process of the image generation in the theme of analysis-by-synthesis.

### 1.1.3 Image Transformation as Intervention

$p(x|do(a), b)$

### 1.1.4 Analysis-by-Synthesis as Model Assumption

# 2 Bibliography

[1] Judea Pearl and Dana Mackenzie. *The Book of Why*. Hachette Book Group, 2018. 1

[2] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. *arXiv*, 2018. 2