This passage initiates by addressing the question depicted in Figure 1. Despite reviewing the referenced papers – (Wang and Isola, 2020; Zimmermann et al., 2021; Von Kügelgen et al., 2021; Xiao et al., 2020; Daunhawer et al., 2022; Eastwood et al., 2023), I acknowledge lingering uncertainties regarding the nuances of contrastive representation learning. In an effort to clarify my understanding, I have composed this report based on my current knowledge and seek your valuable critical comments and feedback.

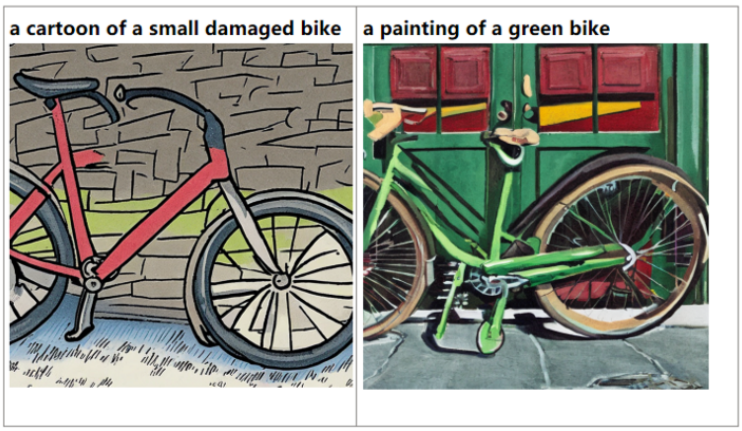


Figure 1: Question - Are these two samples (either in text-modality and image-modality) can be considered a positive pair as the learning objective for disentangling the content variable?

Recent approaches utilize data augmentations as weak supervision to distinguish retained information, termed "content", from discarded information, termed "style" (Eastwood et al., 2023). However, the agreement on real-world data generation process, particularly the identifiability of true causal relations in data generative mechanisms, currently lacks consensus. This lack of agreement introduces uncertainty around terminologies like "content" and "style", and their distinct definitions hinge on how we conceptualize them and how we set the learning objective. Nevertheless, this issue seems to be primarily a matter of nomenclature. The resolution of this uncertainty is anticipated to come with consensus on the identifiability of true mechanisms of data generative.

Although current researches demonstrates that the content variable causes the style variables, a question arises to me regarding the interpretation of "content". Specifically, as shown in Figure 2, does "content" in these works refer to (a) the latent variable governing the class name of a sample, causing other latent variables governing properties like color, position, background, etc., or (b) an intrinsic latent variable that is a common cause of other properties, including the class name? The distinction between these two interpretations is substantial, and it appears to be a pivotal for understanding disentaglement by contrastive representation learning – The literature says that representations of two positve samples should be mapped to nearby features, and thus be mostly invariant to not needed noise factors (Wang and Isola, 2020). However, from a causal perspective, I posit that the invariance to specific noise factors (i.e., latent variables) is not determined by whether the noise factors are "needed" for a task, but rather by the underlying causal relationship between latent variables, which constrains how one constructs the learning objective.

Now, let's consider the two assumed situations: If the former data-generation process (a) is correct, for a classification task, one would only need to impose changes to retain class-level information, as other factors are considered noise for this task. In this scenario, positive pairs would simply need to share the same class name, while data with different class names should be considered negative counterparts—a strategy similar to what is done in CLAP. However, the effectiveness of such representations in achieving supreme robustness and accuracy in classification tasks may be limited. While this approach might work well for images containing only one class of large-scaled objects, as seen in synthetic datasets like 3DIdent and its variations, it may not be suitable for real-world images with complex backgrounds, multiple objects of different classes, and sometimes small-scaled target objects. Looking at supervised classification methods, although they do not follow a contrastive-learning protocol, their learning objectives also emphasize class-level information by

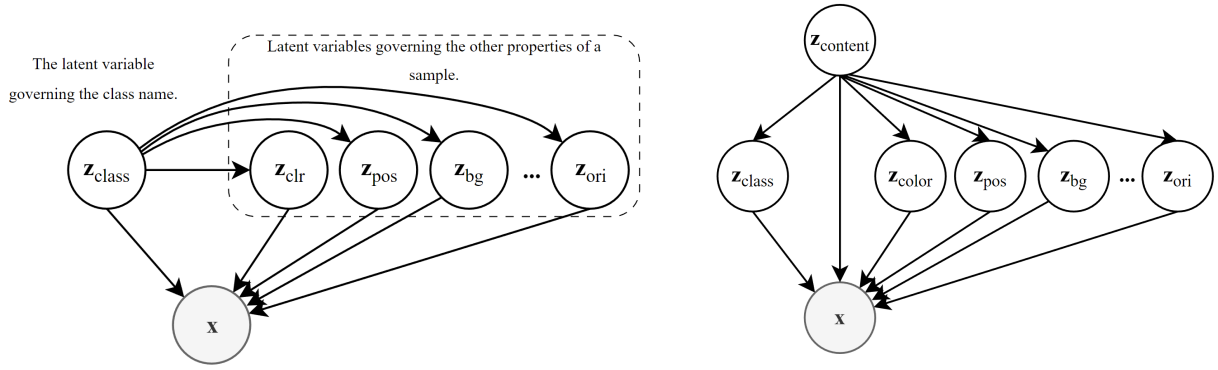


Figure 2: Which data-generative process is more appropriate? In the left, model (a), the latent variable governing class name of a sample is considered the content variable, which act as the common cause of other properties of a sample; In the right, model (b), Without modeling $\mathbf{z}_{\text{content}}$, one could not remove the dependency of the latent variable governing object name ($\mathbf{z}_{\text{class}}$) and the latent variables governing other properties (e.g., $\mathbf{z}_{\text{bg}}$ which governing the background).

treating samples with the same class name as positive samples. However, these methods often face out-of-distribution (OOD) issues, suggesting that the generative process (a) may not be appropriate, and positive pairs sharing the same class name might not be sufficient to obtain disentangled representations, regardless of the changes imposed on the data.

> **Discussion 1**.The varied versions of 3DIdent datasets (i.e., 3DIdent, Causal3DIdent, Multimodal3DIdent) may be insufficient to verify if the intrinsic content variable is disentangled, as they lack inner-class variations of objects. In these datasets, all the samples sharing the same class name should be considered as the same object, as long as we one agree that traditional augmentation techniques (color distortion, rotation, crop, etc.) does not alter the object identity. Yet, defining "the same object" in data is not straightforward, as some works regard different views augmented from a same sample by "texture randomization" as a positive pair (Xiao et al., 2020) – This perspective contrasts with the intuition that topology and shape determine a coarse-grained class of objects, while additional texture information defines object identity (or a fine-grained class of objects).

On the contrary, if the latter data-generation process (b) is more appropriate, changes need to be imposed on data without altering the fact that the target objects in a positive pair should be the same instances, i.e., maintaining object identity. In this context, the answer to the question illustrated in Figure 1 would be that two images with different bicycles should not be considered a positive pair, nor should the two text prompts. This perspective explains the muted performance of CLAP: considering a positive pair like "a dog in a sketch" and "a photo of a yellow dog," this prompt pair cannot constrain the dog as the same one in the two prompts. Consequently, CLAP only attempts to disentangle the $\mathbf{z}_{class}$ variable governing the class name by aligning the samples with the same class names, which is unachievable due to the unresolved dependence on $\mathbf{z}_{class}$ variable and other latent variables without modeling the true content variable.

Xiao et al. (2020) argue that current methods introduce inductive bias by encouraging neural networks to be less sensitive to information regarding augmentation, which may help or hurt. However, I believe that this "hurt" is attributed to the insufficient disentanglement of the latent content variable due to limited changes. Upon the (asymptotical) disentanglement of the intrinsic content, the resulting representations should distributed on hypersphere ($\mathcal{S}^{n-1}$) in a $n$-dimensional space. Therefore, the representations could be applied to instance identification with clustering by spherical distance on this hypersphere, and with the linear separablility of the representations, other downstream tasks (relating to one or several properties of a sample, such as classification, action detection, etc.) can be realized with linear combinations of disentangled representations. This can be regarded as the projection of hypersphere along one or several bases/axis to a lower-dimensional hypersphere ($\mathcal{S}^{n-m}$).

To clarify the goal of disentangled representation learning, it becomes imperative to ensure that samples of a

Figure 3: An old American person stand in front of a house with white wall, with his black dog by his side (generated with SDv2.1)

positive pair share the same indentity, and enough changes are imposed between the negative counterparts. Yet, current methods in contrastive representation learning literature faces challenges on one or both of the two aspects for real-world data:

- Different augmented views of an image (e.g., SimCLR) are considered suitable positive pairs for isolating the content. However, the combination of the traditional data augmentation methods are not sufficient to impose enough changes on data to cast aside all the style information.
- CLAP employs only augmented text pairs, making it easy to impose (some aspects of) style changes on data due to the semantic and logical nature of text. However, class names in a prompt provide only the class-level constraint on positive samples, making it challenging to determine the identity of an instance solely using text data.
- Contrastive language-image training (e.g., CLIP) uses Image-text pairs, where the both parts of a image-text pair can be considered two different views of one sample, if the caption of the image is detailed, e.g. in Figure 3. However, using only image-text pairs may not be capable for disentangling intrinsic content, as text data lacks the informativeness needed to precisely constrain the same object in its image counterpart. For instance, even if one adds numerous attributive adjectives to an object in the textual modality (e.g., "dog"), the text cannot be constrained to represent only the exact same dog as shown in the image.

Therefore, there might be a need to develop a method that combines the logic and semantic nature of textual modality with the informative nature of image modality. Text data is inherently more "particulate" (in a property-wise manner) than image data, while image data is more precise than text data in describing "the exact same object(s)/event(s)" due to its greater informativeness than text data (per image vs. per text prompt, not per memory byte).

# References

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E Vogt. Identifiability results for multimodal contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2022.

Cian Eastwood, Julius von Kügelgen, Linus Ericsson, Diane Bouchacourt, Pascal Vincent, Mark Ibrahim, and Bernhard Schölkopf. Self-supervised disentanglement by leveraging structure in data augmentations. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023.

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. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

Tete Xiao, Xiaolong Wang, Alexei A Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2020.

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.