DISCOV: A Time Series Representations Disentanglement via Contrastive for Non-Intrusive Load Monitoring (NILM)

Khalid Oublal[1,2]*, **Saïd Ladjal**[1], **David Benhaiem**[2], **Emmanuel Le-borgne**[2], **François Roueff**[1]
[1]Institute Polytechnique de Paris, Telecom Paris LTCI/S2A, [2]OneTech TotalEnergies, DS&AI

Editors: Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

Abstract

Improving the generalization capabilities of current machine learning models and improving interpretability are major goals of learning disentangled representations of time series. Notwithstanding this, methods for disentangling time series have mainly focused on identifying independent factors of variation in the data. This overlooks that the causal factors underlying real-world data are often not *statistically independent*. In this paper, we investigate the problem of learning disentangled representations for the electricity consumption of customers' appliances in the context of Non-Intrusive Load Monitoring (NILM) (or energy disaggregation), which allows users to understand and optimise their consumption in order to reduce their carbon footprint. Our goal is to disentangle the role of each attribute in total aggregated consumption. In contrast to existing methods that assume attribute independence, we recognise correlations between attributes in real-world time series. To meet this challenge, we use weakly supervised contrastive disentangling, facilitating the generalisation of the representation across various correlated scenarios and new households. We show that Disentangling the latent space using Contrastive On Variational inference (`DISCOV`) can enhance the downstream task. Furthermore, we find that existing metrics to measure disentanglement are inadequate for the specificity of time series data. To bridge such a gap, an alignment time metric has been introduced to assess the quality of disentanglement. We argue that ongoing efforts in the domain of NILM need to rely on causal scenarios rather than solely on statistical independence. Code is available at `https://oublalkhalid.github.io/DISCOV/`.

1 Introduction

Disentangled representation learning is crucial in various fields like computer vision, speech processing, and natural language processing [6]. It aims to improve model performance by learning latent disentangled representations and enhancing generalizability, robustness, and explainability. These representations have latent units that respond to single attribute changes while remaining invariant to others. Existing approaches assume independent attributes, but in real-world time series data, latent attributes are often causally related. This necessitates a new framework for causal disentanglement. For instance, the consumption profile of some appliances such as "Dishwasher" causes variations in "Washing machine" showing the inadequacy of existing methods in capturing these non-independent attributes [49, 47]. One of the most common frameworks for disentangled representation learning is Variational Autoencoders (VAE) [18], a deep generative model trained to disentangle the underlying explanatory attributes.

*Correspondence to:`khalid.oublal@polytechnique.edu`. This work was honored with the Gatsby Garant award, UniReps, NeurIPS 2023.

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

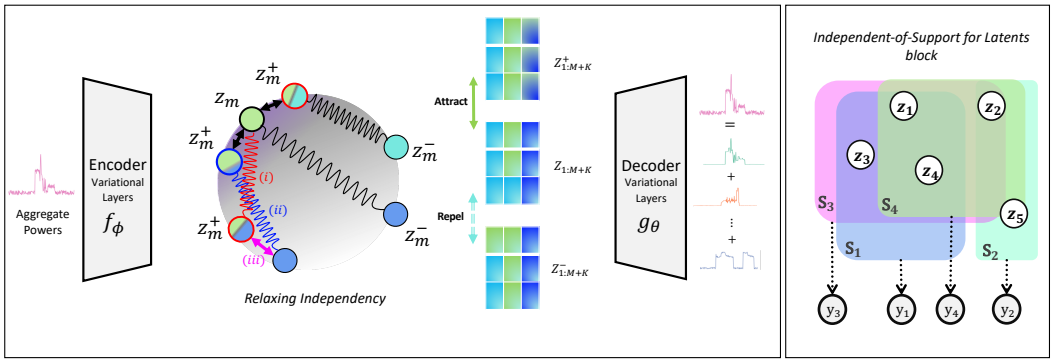

Figure 1: **Relaxing Independency via Contrastive.** Latent attributes are causally correlated, allows positive pairs ($\mathbf{z} := f_\phi(\mathbf{x})$, $\mathbf{z}^+ := f_\phi(\mathbf{x}^+)$) to decrease their distance, while negative pairs ($\mathbf{z} := f_\phi(\mathbf{x})$, $\mathbf{z}^- = f_\phi(\mathbf{x}^-)$) increase it, and allows cases where unlikely combinations occur, although forcing statically independence does not prohibit these cases.

Definition 1.1 (**Disentangled Representation**, [21, 30])**.** Disentangled representation should separate the distinct, independent, and informative generative factors of variation in the data. Single latent variables are sensitive to changes in single underlying generative factors while being relatively invariant to changes in other factors.

Disentanglement via VAE can be achieved by a regularization term of the Kullback-Leibler divergence between the posterior of the latent attributes and a standard Multivariate Gaussian prior [18], which enforces the learned latent attribute to be as independent as possible. It is expected to recover the latent variables if the observation in the real world is generated by countable independent attributes. To further enhance the independence, various extensions of VAE consider minimizing the mutual information among latent attributes [24]. [24] further encourage independence by reducing the total correlation among attributes. Our focus in this work is a more general case, where the data does not have specificities like domain frequency, or amplitude to analysis. Household energy consumption disaggregation, also known as Non-Intrusive Load Monitoring (NILM), is a key application. Given only the main consumption of a household, the energy disaggregation algorithm identifies which appliances are operating. Such a capability is extremely vital given the growing interest in reducing carbon footprints through user energy behavior, which poses a challenge to conventional algorithms. Many households rely on past bills to adjust future energy use, underscoring the importance of energy disaggregation algorithms. Further, the growing need for electricity flexibility originating from the ever-increasing fraction of renewable energy in the electricity mix requires having a better understanding of consumer uses. Recent work [9, 50, 39] hold promising results, yet persistent challenges in generalisability and robustness stem from the correlations occurring within time series a challenge that spans beyond the domain of time series in general. In this work, we tackle the energy disaggregation problem from the perspective of disentanglement.

Our work is distinguished by instead of assuming independent factors we will only assume that the support of the distribution factorizes. We explore how to design an efficient and disentangling representation under correlated attributes using weak supervised contrastive learning. We perform an ablation investigation to understand the impact of considering statistical independence versus the case where we avoid it by giving the latent space a support factorization through weakly supervised contrastive learning. This addresses latent space misalignment between attributes, maintains generalizability, and preserves disentanglement through the *Pairwise similarity* over $\mathbf{z}$ setting it apart from methods relying on *independence*. More clearly, we break the concept of independence, allowing any combination of individual attributes, to be possible, even if some combinations are unlikely. Our experiments on three datasets of increasingly difficult correlation settings, show that `DISCOV` improves robustness to attribute correlation and improves disentanglement (as measured by SAP, DCI, RMIG,TDS) over state of the art (c.f. §4.3).

[1] Furthermore, we introduce an in-depth variational-based self-attention for extracting high semantic representations from time series. An ablation study shows that VAE learns complex representations;

[2] added attention improves further (in-depth model with L layer $L = \{4, 8, 16, 32\}$ c.f. §5). This approach retains dimension reduction while avoiding temporal locality.

[3] Additionally, our proposed Time Disentanglement Metric (TDS) aligns more effectively with decoder output compared to existing metrics. These findings strongly recommend its application for time series representation in diverse fields, including finance and medical data like ECG.

2 Problem Formulation and Preliminaries

We consider a c-variate time series observed at times $t = 1, \ldots, \tau$. We denote by $\mathbf{x} \in \mathbb{R}^{c \times \tau}$ the $c \times \tau$ resulting matrix with rows denoted by $x_1, \ldots, x_c$, where each row can be seen as a univariate time series. In the electric load application, we have $c = 3$, and x_1 is the sampled active power which is the one billed to customers, x_2 is the sampled reactive power, and x_3 is the sampled apparent power. The goal of non-intrusive load monitoring (NILM) is to use $\mathbf{x}$ to express x_1 as

$$x_1 = \sum_{m=1}^{M} y_m + \xi \,, \tag{1}$$

where, for each $m = 1, \ldots, M$, $y_m \in \mathbb{R}^\tau$ represents the contribution of the m-th electric device among the M ones identified in the household, and $\xi \in \mathbb{R}^\tau$ denotes a residual noise. We further denote by $\mathbf{y}$ the $M \times \tau$ matrix with row-wise stacked devices' contributions.

The NILM mapping $\mathbf{x} \mapsto \{\mathbf{y}_1 \ldots \mathbf{y}_M\}$, where $\mathbf{x} = \sum_i \mathbf{y}_i$ is generally learnt from a training data set $\mathcal{Z} = \{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$. VAEs rely on two main ingredients: 1) a generative model (p_θ) based on a latent variable, and a decoder g_θ; 2) a variational family (q_ϕ), which approximates the conditional density of the latent variable given the observed variable based on an encoder f_ϕ.

In a VAE, both (unknown) parameters θ and ϕ are learnt from the training data set $\mathcal{Z} = \{\mathbf{x}_n\}_{n=1}^N$. A key idea for defining the goodness of fit part of the learning criterion is to rely the **E**vidence **L**ower **B**ound (ELBO), which provides a lower bound on (and a proxy of) the log-likelihood

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})} \left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})) \,, \tag{2}$$

where we denoted the latent variable by $\mathbf{z}$, defined as a $(M + K) \times d$ matrix and p denotes its distribution. The use of ELBO goes back to traditional variational Bayes inference. An additional feature of VAE's is to define q_ϕ and p_θ through an encoder/decoder pair of neural networks (f_ϕ, g_θ). A standard choice in a VAE is to rely on Gaussian distributions and, for instance, to set $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu(\mathbf{x}, \phi), \sigma^2(\mathbf{x}, \phi))$, where $\mu(\mathbf{x}, \phi)$ and $\sigma^2(\mathbf{x}, \phi)$ are the outputs of the encoder f_ϕ.

As discussed in Section 1, various alternatives features such as β/TC/Factor/DIP-VAE [13] have been proposed, where a specific distribution $p(\mathbf{z})$ is learned. The objective is to increase the disentanglement of the latent variable $\mathbf{z}$ and align it with the corresponding attribute. However, as previously stated, they assume statistical independence among attributes, leading to the assumption: $p(\mathbf{z}) = p(\mathbf{z}_1) \ldots p(\mathbf{z}_{\mathbf{M+K}})$. As we explained in 1, appliances are not usually used independently. In [42], correlated attributes have been taken into account by replacing the factorization constraint with support factorization via Hausdorff Factorized Support (HDF). To meet this criterion, they penalize the Hausdorff pairwise estimate beteween support and its Cartesian product, based solely on the distance without any alignment on the input. In this paper, we are investigating an alternate way to achieve both alignment and disentanglement leading to a generalizable representation. To that end, we draw on support factorization, and we replace HDF with a *Pairwise Similarity* penalty. In the next section, we develop our proposed method based on weakly contrastive learning to have factorized support, and we highlight how it provides an advantage both in terms of computation and latent representation.

3 Proposed Methods: Disentangled as Factorization of Supports

Our primary aim is to disentangle the latent space by relaxing the assumption of independence. To achieve this, we introduce a concrete training criterion that promotes a factorized support structure. We begin by considering deterministic representations obtained through the encoder, where $\mathbf{z} := f_\phi(\mathbf{x})$. The criterion we apply enforces the factorized support condition on the aggregate distribution, denoted as $\bar{q}_\phi(\mathbf{z}) = \mathbb{E}_\mathbf{x}[f_\phi(\mathbf{x})]$. It is worth noting that $\bar{q}_\phi(\mathbf{z})$ shares a conceptual similarity with the aggregate posterior $q_\phi(\mathbf{z})$ used in methods like TCVAE, although we focus on points generated by f_ϕ. To align with our factorized support assumption concerning the ground truth, our goal is to promote the factorization of the support for $\bar{q}_\phi(\mathbf{z})$, ensuring that $\mathcal{Z}$ is equivalent to the Cartesian product of the support in each dimension, denoted as $\mathcal{Z}^\times$. In practical scenarios, we often work with a finite set of

data, denoted as $\{\mathbf{x}_i\}_{i=1}^N$, and can only estimate the support based on a finite set of representations, $\{f_\phi(\mathbf{x}_i)\}_{i=1}^N$. To encourage such a pairwise factorized support, we can minimize sliced/pairwise contrastive with the additional benefit of keeping computation tractable when the dimension of latent space is large. Specifically, we approximate the support as $\mathcal{Z} \approx \mathbf{Z}$ and the Cartesian product of each dimension's support as $\mathcal{Z}^\times \approx Z_{:,1} \times Z_{:,2} \times \ldots \times Z_{:,M+K}$.

3.1 Support factorization via Weakly supervised Constrastive

Let us first formalize the contrastive learning setup. Each training triplet comprises a reference sample $\mathbf{x}$ along with a positive (similar) sample $\mathbf{x}^+$ and negative (dissimilar) samples $\mathbf{x}_1^-, \ldots, \mathbf{x}_N^-$ against which it is to be contrasted. As introduced in the previous section, we assume that these samples generate corresponding latent: $\mathbf{z}$, $\mathbf{z}^+$, $\mathbf{z}_1^-, \ldots, \mathbf{z}_N^-$. The positive sample, denoted as $\mathbf{z}^+$, is generated from a closely related dataset in which appliance m is activated. In contrast, the negative samples, $\mathbf{z}_1^-, \ldots, \mathbf{z}_N^-$, are drawn from a dataset where appliance m remains inactive. This formalization of contrastive learning ensures that positive samples are semantically similar and negatives are dissimilar. Self-supervised contrastive learning is widely used in vision [22], where the loss is defined as:

$$\mathcal{L}_{\text{self-contrastive}} = -\sum_{i \in I} \log \frac{e^{(z_i \cdot z_{j(i)}/\tau)}}{\sum_{a \in \mathcal{A}(i)} e^{(z_i \cdot z_a/\tau)}}. \tag{3}$$

where, $z_i \in \mathcal{Z}$, the $\cdot$ symbol denotes the inner (dot) product, $\tau \in \mathbb{R}^+$ is a scalar temperature parameter, and $\mathcal{A}(i) \equiv I \setminus \{i\}$. The index i is called the anchor z_i, index $j(i)$ is refer to the positive z_i^+, and the other $2(N-1)$ indices ($\{k \in \mathcal{A}(i) \setminus \{j(i)\}\}$) are called the negatives $z_{k \neq i}^-$. We note that for each anchor i, there is 1 positive pair and $2N-2$ negative pairs. The denominator has a total of $2N-1$ terms (the positives and negatives). In a multi-class scenario, disentangling and aligning data encounters challenges when several samples belong to the same class, as we aim to match certain pairs of data points (e.g., $z_{i,j}$ to $z_{i,j}^+$) and drive others away (i.e. $z_{i,j}$ from $z_{k \neq i,j}$ or $z_{k \neq i,j}^+$). Using contrastive objective, we link the learned latent representation to ground-truth attributes using a limited number of pair labels. This connection is facilitated by employing positive and negative samples, as demonstrated in [53].

We adopt this by following these two rules: firstly, the loss should not rely on statically independent attributes, mirroring realistic data scenarios; secondly, it should prioritize attribute alignment to maintain sufficient information [55]. Under our relaxed hypothesis on statistical independence, [42] introduces a concrete training criterion to promote factorized support. Specifically, they focus on deterministic representations generated by the encoder, denoted as $\mathbf{z} = f_\phi(\mathbf{x})$. They apply the factorial support criterion to the combined distribution $\bar{q}_\phi(\mathbf{z}) = \mathbb{E}_{\mathbf{x} \sim \mathbf{X}}[f_\phi(\mathbf{x})]$. This distribution conceptually resembles the aggregate posterior $q_\phi(\mathbf{z})$ in models like TCVAE, but we generate points through a deterministic mapping, rather than a stochastic one. To align with our factorized support assumption regarding the ground truth, we aim to encourage the support of $\bar{q}_\phi(\mathbf{z})$ to factorize. In other words, we want the support of $\bar{q}_\phi(\mathbf{z})$ and the Cartesian product of the support of each dimension, $\mathcal{Z}(\bar{q}_\phi(z_1)) \times \cdots \times \mathcal{Z}(\bar{q}_\phi(z_{M+K}))$, to be identical. For simplicity, we use the notations $\mathcal{Z}$ and $\mathcal{Z}^\times$ to represent $\mathcal{Z}(\bar{q}_\phi(\mathbf{z}))$ and $\mathcal{Z}^\times(\bar{q}_\phi(\mathbf{z}))$ respectively. To guide the learning process, it requires a divergence or metric to measure the dissimilarity between $\mathcal{Z}$ and $\mathcal{Z}^\times$. Since supports are sets, it is natural to employ a set distance measure, such as the Hausdorff distance. In practical scenarios, we work with a finite latent space $\{f_\phi(\mathbf{x}^{(i)})\}_{i=1}^N$ and aim to encourage pairwise factorized support, with access to a batch of b inputs sequence denotes $\mathbf{X}$ yielding $b \times (M+K)d$-dimensional latent representations $\mathbf{Z} = f_\phi(\mathbf{X})$, we estimate Hausdorff distances using sample-based approximations to the support $\mathcal{Z} \approx \mathbf{Z}$ and the Cartesian product of each dimension support as, $\mathcal{Z}^\times \approx \mathbf{Z}_{:,1} \times \ldots \times \mathbf{Z}_{:,M+K}$. They propose to ensure this constraint by penalizing the pairwise Hausdorff estimate $\hat{d}_H(\mathbf{Z})$.

$$\hat{d}_H(\mathbf{Z}) = \sum_{i=1}^{M+K} \sum_{j=i+1}^{M+K} \max_{\mathbf{z} \in \mathbf{Z}_{:,i} \times \mathbf{Z}_{:,j}} \left[\min_{\mathbf{z}' \in \mathbf{Z}_{:,(i,j)}} \|\mathbf{z} - \mathbf{z}'\|^2 \right]. \tag{4}$$

Rather than optimizing $\hat{d}_H(\mathbf{Z})$ through computationally inefficient min and max operations, and given the sensitivity to outliers when matching $\mathcal{Z}$ and $\mathcal{Z}^\times$ using distance, this serves as the driving force for our research. Our approach focuses on capturing pairwise similarity via contrastive learning and gives the opportunity to prioritize attribute alignment to maintain sufficient information [55].

This alternative approach aims to design a disentangled and generalizable representation through augmentation (multi-views of a sequence). First, as the supervised contrastive loss we consider the pair dimension (i, j), $\mathbf{z}$ from the cartesian product of each dimension support $\mathbf{Z}_{:,i} \times \mathbf{Z}_{:,j}$ as an anchor, and $\mathbf{z}^+$ from the support $\mathbf{Z}_{:,(i,j)}$ as positive augmentation $\mathcal{T}$ this means for mini-batch we have an augmentation of appliance encoded in dimension i. For $\mathcal{T}$, we use channel dropout and Gaussian noise to create positive pairs, for channel dropout, we randomly mask out a subset of variables while keeping the full-time series intact. The negative will be an element of $\mathcal{Z}$ other than the (i, j) columns i.e $\mathbf{z}^- \in \mathbf{Z}_{:,\neq(i,j)}$, also we considering case where $\mathbf{x}^-$ available the (i, j) of $\mathbf{z}^-$, i.e $\mathbf{z}^- = f_\phi(\mathbf{x}^-)$. Eq 5 is similar to the InfoNCE objective [40], as it reinforces the pre-existing similarity between $\mathbf{Z}_{:,(i,j)}$ and $\mathbf{Z}_{:,i} \times \mathbf{Z}_{:,j}$ (i.e between $\mathcal{Z}$ and $\mathcal{Z}^\times$), and repel $\mathbf{Z}_{:,\neq(i,j)}$ and $\mathbf{Z}_{:,i} \times \mathbf{Z}_{:,j}$, this allowing the support $\mathbf{Z} \approx \mathcal{Z}$ to have a flexible factorization support.

$$\mathcal{L}^{(1)}_{\text{DIS}}(\mathbf{Z}) = \sum_{i=1}^{M+K} \sum_{j=i+1}^{M+K} \sum_{\mathbf{z} \in \mathbf{Z}_{:,i} \times \mathbf{Z}_{:,j}} \sum_{\mathbf{z}^+ \in \mathbf{Z}_{:,(i,j)}} \frac{\exp\left(\mathbf{z} \cdot \mathbf{z}^+\right)}{\exp\left(\mathbf{z} \cdot \mathbf{z}^+\right) + \sum_{\mathbf{z}^- \in \mathbf{Z}_{:,\neq(i,j)}} \exp\left(\mathbf{z} \cdot \mathbf{z}^-\right)} \tag{5}$$

Our results show that this could be more adapted for weak supervised contrastive learning [53], i.e. we need only a positive augmentation in minibatch. The proposed weakly supervised contrastive learning loss combines two terms. The first term enforces axis alignment based on the correlation between $\mathbf{Z}_{:,i}$ and $\mathbf{Z}^+_{:,i}$ (positive augmentation of $\mathbf{Z}_{:,i}$ from $\mathcal{Z}$). This ensures that only one latent variable learns this alignment for fixed attributes (invariant) which forces the approximate posterior of the shared latent $\mathcal{Z}$ to be similar to $\mathcal{Z}^\times$ and its augmentation. The second term minimizes information redundancy by measuring the correlation between $\mathbf{Z}_{:,i}$ and $\mathbf{Z}^+_{:,j\neq i}$ or $\mathbf{Z}_{:,j\neq i}$, which are nearly equivalent in the contrastive sense.

$$\mathcal{L}^{(2)}_{DIS}(\mathbf{Z}) = \sum_{i=1}^{M+K} \sum_{j=i+1}^{M+K} \left(1 - d(\mathbf{Z}_{:,i} \times \mathbf{Z}_{:,j}, \mathbf{Z}^+_{:,i})\right)^2 + \sum_{i=1}^{M+K} \sum_{j=i+1}^{M+K} d(\mathbf{Z}_{:,i} \times \mathbf{Z}_{:,j}, \mathbf{Z}^+_{:,j})^2, \tag{6}$$

where $d(\mathbf{Z}_{:,i}, \mathbf{Z}^+_{:,i})$ (resp. $d(\mathbf{Z}_{:,i}, \mathbf{Z}^+_{:,i})$) is the cosine similarity between $\mathbf{Z}_{:,i}$ and $\mathbf{Z}^+_{:,i}$ (resp. $\mathbf{Z}_{:,i}$ and $\mathbf{Z}^+_{:,i}$) in a mini-batch. During training, for mini-bach augmentation $\mathbf{Z}^+$ affects only one appliance, with others remaining fixed. We assume sufficient augmentation for each factor across the batch.

$$\mathcal{L}(\mathcal{D}, \phi, \theta) = \mathbb{E}_{\mathbf{X} \overset{b}{\sim} \mathcal{D}}\Big[\gamma \mathcal{L}_{DIS}(f_\phi(\mathbf{X})) + \frac{1}{b} \sum_{\mathbf{x} \in \mathbf{X}} (\mathcal{L}_{\text{rec}}(\mathbf{x}; \phi, \theta) + \beta_1 \mathcal{L}_{\mathbf{KL}}(\mathbf{x}; \phi, \theta) + \beta_2 \mathcal{L}^{\Delta}_{\mathbf{KL}}(\mathbf{x}; \phi, \theta))\Big]. \tag{7}$$

3.2 Attention Variational Auto-encoders

To avoid time locality during dimension reduction, and keep long-range capability we refer to an in-depth Temporal Attention with l-Variational layers. NVAE [43] proposed an in-depth autoencoder for which the latent space $\mathbf{z}$ is level-structured and attended locally [41, 4], this shows an effective results for image reconstruction. In this work, we enable the model to establish strong couplings, as depicted. Our core idea aims to address construct Time context $\hat{T}^l$ that effectively captures the most informative features from a given sequence $T^{<l} = \{T^i\}^l_{i=1}$ across bottom-up and top-down, where $T^{<l}$ is the output of the residual network. Both $\hat{T}^l$ and T^l are features with the same dimensionality: $\hat{T}^l \in \mathbb{R}^{T \times C}$ and $T^i \in \mathbb{R}^{T \times C}$. In our model, we employ Temporal Self-attention [45] to construct either the prior or posterior beliefs of variational layers, which enables us to handle long context sequences with large dimensions τ effectively. The construction of $\hat{T}^l$ relies on a query feature $\mathbf{Q}^l \in \mathbb{R}^{T \times Q}$ of dimensionality Q with $Q \ll C$, and the corresponding context T^l is represented by a key feature $\mathbf{K}^l \in \mathbb{R}^{T \times Q}$. Importantly, $\hat{T}^l(t)$ of time step i in sequence τ depends solely on the time instances in $T^{<l}$. For more consistency, using Multihead-attention [45] allows the model to focus on different aspects of the input sequence simultaneously, which can be useful for capturing various relationships and patterns. which allows the model to jointly attend to information from different representation subspaces at different scales. Instead of computing a single attention function, this method first projects $\mathbf{Q}^l$, $\mathbf{K}^{<l}$), $\mathbf{T}^{<l}$) into h different vectors, respectively. Attention is applied individually to these h projections. The output is a linear transformation of the concatenation of all attention outputs. For the remainder of this paper, we presume that `DISCOV` employs self-attention. Prior works [43] have sought to mitigate against exploding Kullback-Leibler divergence (KL) in Eq 2

by using parametric coordination between the prior and posterior distributions. Motivated by this insight, we seek to establish further communication between them. We accomplish this by allowing the generative model to choose the most explanatory features in $h^{\geq l}$ by generating the query feature $\mathbf{Q}_q^l$. Finally, the holistic conditioning factor for the posterior is:

$$\hat{T}_q^l \leftarrow \textbf{Self-Attention}(h^{\geq l}, \mathbf{Q}_q^l, \mathbf{K}_q^{\geq l}) \text{ for } l = L, L-1, ..., 1. \tag{8}$$

We adopt the Gaussian residual parametrization between the prior and the posterior. The prior is given by $p(\mathbf{z}_l|\mathbf{z}_{<l}) = \mathcal{N}(\mu(T_p^l, \theta), \sigma(T_p^l, \theta)$. The posterior is then given by $q(\mathbf{z}_l|\mathbf{x}, \mathbf{z}_{<l}) = \mathcal{N}(\mu(T_p^l, \theta) + \mathbf{\Delta}\mu(\hat{T}_q^l, \phi), \sigma(T_p^l, \theta) \cdot \mathbf{\Delta}\sigma(\hat{T}_q^l, \phi))$ where the sum $(+)$ and product $(\cdot)$ are pointwise, and T_q^l is defined in Eq 8. $\mu(\cdot)$, $\sigma(\cdot)$, $\mathbf{\Delta}\mu(\cdot)$, and $\mathbf{\Delta}\sigma(\cdot)$ are transformations implemented as temporal convolutions layers.

3.3 Evaluating Disentanglement in Time Series

Evaluating disentanglement in series representation is more challenging than established computer vision metrics. Existing time series methods rely on qualitative observations and predictive performance, while metrics like Mutual Information Gap (MIG) [32] have limitations with continuous labels. To address this, we adapted RMIG [10] for continuous labels and used DCI metrics from [15]. Additionally, we employed SAP [27] to measure prediction error differences in the most informative latent dimensions for ground truth attributes. Our evaluation, including β-`VAE`/+HDF and `FactorVAE` scores, . These metrics face challenges with sequential data and do not provide measures of attribute alignment. To overcome this limitation, we introduce the Time Disentanglement Score (TDS) from an information-gain perspective. TDS assesses how well the latent representation $\mathbf{z} := f_\phi(\mathbf{x})$ maintains the invariance of an attribute m in $\mathbf{x}$ when this attribute changes. TDS relies on the correlation matrix between $\mathbf{z}$ and $\mathbf{z}^+$, where $\mathbf{z} := f_\phi(\mathbf{x})$ and $\mathbf{z}^+ := f_\phi(\mathcal{T}(\mathbf{x}))$, with $\mathcal{T}$ denoting an augmentation function. This correlation matrix quantifies the consistency of attribute components. Additionally, TDS evaluates how well $\mathbf{z}$ contributes to the reconstruction of $\mathbf{y}$ and how $\mathbf{z}^+$ contributes to the reconstruction of $\mathcal{T}(\mathbf{y})$. Specifically, it assesses whether each z_m (or z_m^+) can effectively reconstruct the corresponding y_m (or y_m^+). Time series data often exhibit variations that may not always align with conventional metrics, especially when considering the presence or absence of underlying attributes. To address this challenge, we introduce the Time Disentanglement Score (TDS), a metric designed to assess the disentanglement of attributes in time series data. The foundation of TDS lies in an Information Gain perspective, which measures the reduction in entropy when an attribute is present compared to when it's not.

$$TDS = \frac{1}{dim(\mathbf{z})} \sum_{n \neq m} \sum_{k} \frac{||z_m - z_{n,k}^+||^2}{\text{Var}[z_m]}, \tag{9}$$

In the context of TDS, we augment factor m in a time series window $\mathbf{x}$ with a specific objective: to maintain stable entropy when the factor is present and reduce entropy when it's absent. This augmentation aims to capture the essence of attribute-related information within the data. We note that high TDS informativeness signifies strong disentanglement, while a significant distance implies reduced disentanglement and higher attribute correlation, aligning with [16].

4 Experiments

4.1 Experimental Design

Datasets. We conducted experiments on two publicly available datasets, namely UK-DALE [20] and REDD [26]. The dataset UK-DALE [20] consists of 5 dwellings with a varying number of sub-metered devices and includes aggregate and individual aggregate and individual equipment-level power measurements, sampled equipment, sampled at 1/6 Hz.

Evaluation Metrics. To assess the accuracy of all the methods compared, we rely on `RMSE` as a metric for the reconstruction of downstream tasks, in turn for the reconstruction of the load consumption of each individual appliance. Our investigation encompasses both quantitative and qualitative scaling, with a focus on disentangling performance assessed with metrics such as `DCI`/`RMIG`/`TDS`. This aims to measure its effectiveness, notably in scenarios with correlated consumption signatures, which we consider to be out-of-distribution (ODD).

Baseline. We contrast DISCOV against downstream task models in the energy domain, namely Bert4NILM [52] and S2P [48], as well as S2P [11], maintaining the original implementation configurations for these models. Additionally, we present a variety of β-TC/Factor/RNN/-VAE implementations tailored for time series analysis.

Experimental Platform. We conduct 5 rounds of experiments, reporting the averaged results and standard deviation. The experiments are performed on 8× NVIDIA A100 GPUs and 40 Intel(R) 281 Xeon(R) Silver 4210 CPU @ 2.20GHz. The models are implemented in PyTorch.

4.2 Architecture Settings

Our model uses a bi-directional encoder, which processes the input data in a hierarchical manner to produce a low-resolution latent code that is refined latent code that is refined by a series of oversampling layers. This code is then refined by a series of oversampling layers in *Residual Decoders* blocks, which progressively increases the resolution.

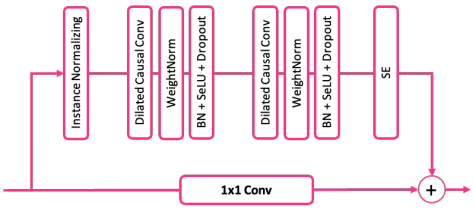

Figure 2: Residual Cell for Encodeur (Inference Model q_ϕ)

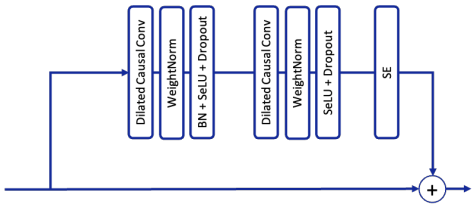

Figure 3: Residual Cell for DISCOV Decodeur (Generative Model p_θ).

Residual Blocs. Activation functions are pivotal for enabling models to learn nonlinear representations, but vanishing and exploding gradients can hinder learning. The Temporal Convolutional Network (TCN) [31] tackles these issues using Rectified Linear Unit (ReLU), weight normalization, and dropout layers. In our Residual model, we simplify the residual block by replacing these components with the Sigmoid Linear Units, which offers advantages and immunity to gradient problems. It reduces training time, efficiently learns robust features, and outperforms weight normalization. SiLU [17] is defined as $\text{SiLU}(x) = x \times \sigma(x)$ where $\sigma(x)$ is the logistic sigmoid.

Squeeze-and-Excitation on Spatial and Temporal. SE block enhances our neural networks by selectively emphasizing important features and suppressing less relevant ones. It does this through global information gathering (squeezing) and feature recalibration (excitation). We find that extending SE for time series data improves the capture of significant temporal patterns in sequence. Our Residual encoders (Inference Model q_ϕ) in Fig 2 and Decodeur (Generative Model p_θ) in Fig 3.

4.3 Performance and Informativity of Contrastive

Finding: *DISCOV retains its robustness in correlated scenarios and achieves comparable performance to baseline models.*

In evaluating the robustness of DISCOV regarding correlations in appliance signatures or consumption, we consider several pairs of appliances. Firstly, there's the `No.Corr` scenario, where we examine the correlation between the refrigerator's signature and the dishwasher's signature. These appliances are typically active at different times, resulting in less correlated signatures. Moving on to specific pairs, `Washing Machine/dryer` involves analyzing the correlation between the washing machine's signature and the dryer's signature. Given that these appliances are often used sequentially, their signatures might exhibit some level of correlation. In `Microwave/Oven`, the focus is on evaluating the correlation between the microwave's signature and the oven's signature. These appliances have distinct power profiles and usage patterns, potentially leading to lower correlation. Since these appliances are often used independently, their signatures may exhibit a lower level of correlation. Lastly, the `Random Device` approach involves selecting two random appliances from a dataset.

Sc.	Methods	No.Corr $\sigma = \infty$			Washing Machine/dryer $\sigma = 0.3$			Microwave/Oven $\sigma = 0.4$			Random Device $\sigma = 0.8$		
	Metrics –>	DCI↓	TDS↓	RMSE ↓	DCI↓	TDS↓	RMSE ↓	DCI↓	TDS↓	RMSE ↓	DCI↓	TDS↓	RMSE ↓
REDD [26]	Bert4NILM	-	-	56.4 ±2.58	-	-	70.2±1.45	-	-	72.08±0.96	-	-	70.92±1.15
	S2S	-	-	54.3 ±3.12	-	-	69.5±3.56	-	-	72.31±2.45	-	-	69.95±3.26
	β-VAE	72.4±3.10	0.96±.15	48.6 ±2.32	72.4±3.10	0.96±.15	52.6 ±2.31	72.4±3.10	0.96±.15	54.73±1.54	74.29±2.04	1.08±.09	52.99±1.91
	β-TCVAE	78.0±1.09	0.94±.13	43.2 ±2.23	78.0±1.09	0.94±.13	49.2 ±1.13	77.23±0.76	0.94±.13	50.87±1.17	79.74±0.84	1.07±.11	49.65±1.43
	FactorVAE	68.4±2.41	0.97±.03	47.7 ±1.35	68.4±2.41	0.97±.03	53.2 ±1.02	69.78±1.43	0.97±.03	54.32±0.64	69.95±1.63	1.00±.02	53.45±0.82
	HDF	79.8 ± .10	0.64 ± .05	57.2 ± 2.15	79.8 ± .10	0.64 ± .05	61.3 ± 1.82	79.56±0.28	0.64±.05	62.33±1.23	80.37±.05	0.72±.03	61.64±1.52
	β-VAE + HDF	73.1±1.01	0.69±.02	34.4±1.89	73.1±1.01	0.69±.02	38.1±1.34	73.59±0.86	0.69±.04	39.65±0.87	74.25±0.59	0.73±.05	38.48±1.04
	β-TCVAE + HDF	**67.2±2.01**	**0.52±.02**	**24.3 ±1.81**	**67.2±2.01**	**0.52±.02**	**27.4 ±1.13**	**67.51±1.84**	**0.52±.07**	**28.94±0.66**	**68.79±1.27**	**0.58±.04**	**27.77±0.83**
	DISCOV	**63.5±1.35**	**0.49±.02**	**19.6±1.95**	**69.3±1.2**	**0.4±.02**	**22.3±1.79**	**70.3±0.82**	**0.49±.02**	**23.97±1.19**	**67.12±0.91**	**0.51±.01**	**22.63±1.49**
UK-DALE [20]	Bert4NILM	-	-	57.85 ±1.88	-	-	68.8±1.12	-	-	73.41±1.35	-	-	72.78±0.88
	S2S	-	-	56.38 ±2.22	-	-	67.8±2.76	-	-	73.95±1.91	-	-	70.92±2.25
	β-VAE	73.78±2.68	1.08±.09	50.14 ±1.87	75.47±1.98	0.82±.10	51.7±1.79	70.8±2.62	0.85±.11	55.98±1.27	76.18±1.54	1.16±.08	54.83±1.58
	β-TCVAE	79.57±0.84	1.07±.11	45.72 ±1.68	80.23±0.54	0.81±.09	48.3±0.94	76.2±0.54	0.83±.10	51.74±0.94	80.88±0.53	1.15±.10	51.15±1.10
	FactorVAE	70.14±1.89	1.00±.02	49.02 ±1.05	71.89±1.24	0.94±.02	52.4±0.85	68.7±1.13	0.92±.02	55.24±0.42	71.57±1.27	1.06±.01	54.68±0.64
	HDF	80.12±.05	0.72±.03	58.49 ±1.45	80.26±.03	0.56±.03	6.0±1.42	78.8±0.15	0.58±.03	63.79±0.97	80.61±.02	0.80±.02	63.22±1.17
	β-VAE + HDF	74.47±0.61	0.73±.05	36.09±1.25	75.12±0.41	0.67±.02	37.4±1.04	72.8±0.52	0.64±.03	40.92±0.66	75.07±0.43	0.75±.03	39.68±0.80
	β-TCVAE + HDF	**68.54±1.36**	**0.58±.04**	**25.88 ±1.20**	**69.28±1.01**	**0.46±.01**	**26.7±0.88**	**66.7±1.51**	**0.45±.02**	**29.82±0.51**	**7.04±0.93**	**0.72±.02**	**40.49±0.64**
	DISCOV	**64.42±0.96**	**0.51±.01**	**21.35 ±1.80**	**65.11±0.66**	**0.39±.01**	**21.5±1.44**	**69.5±0.43**	**0.48±.01**	**24.94±0.87**	**65.05±0.71**	**0.55±.01**	**24.05±1.30**

Table 1: Average scores DCI, TDS, and RMSE vary from No Correlation (left) to every appliance correlated with one confounder (right) on uncorrelated test data. Red to blue, with bold indicating the best performance per correlation. (↓ lower is better, ↑ higher is better [**Top-1**,**Top-2**]).

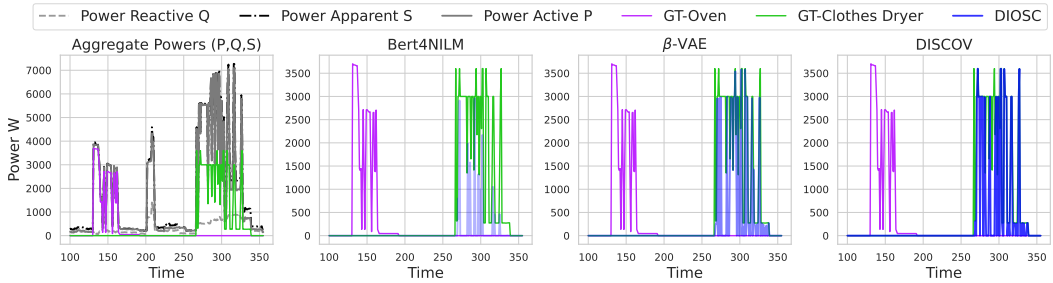

Figure 4: Prediction Clothes dryer under in correlated case (left) and uncorrelated case (right) over a time window of 256min. Moving from left to right, the graph illustrates the aggregated power (**P**,**Q**,**S**) alongside the ground-truth (**Green**) to be identified under correlation with Oven (**Magenta**).

5 Ablation Studies

In this section, we conduct ablation experiments to assess DISCOV effectiveness and robustness in comparison to traditional variant VAEs. Our experiments utilize the Uk-Dale, REDD, and REFIT datasets with a fixed random seed.

(1) A Self-Attention Variational Autoencoder Learn an Effective Representation.

Finding: *DISCOV with increasing depth, the representation becomes over 20% more separable (40% in terms of TDS), downtasking improves performance by 50%, and attention mechanisms contribute to a 10% enhancement in results.*

Table 2, we observe notable differences in performance as the depth (L) of the model architecture varies including Root Mean Square Error (RMSE), Relative Mutual Information Gain (RMIG), and Task Discriminative Score (TDS) for various methods, with a particular emphasis on DISCOV variants with and without attention as the depth (L) increases. Regarding RMSE, which measures the accuracy of the models, we find that the baseline methods VAE, β-TCVAE, and DIP-VAE exhibit consistently higher RMSE values compared to the DISCOV variants. Furthermore, introducing the DISCOV significantly improves RMSE values across all methods, indicating the effectiveness of the DISCOV loss in enhancing model performance. Additionally, as depth (L) increases from 4 to 16, we observe that the DisCoV variants consistently outperform the baseline methods in terms of RMSE.

Notably, when L reaches 16, both DISCOV and DISCOV attention achieve the lowest RMSE value of 0.48, showcasing the superior performance of DISCOV based models with higher depth. It is also worth mentioning that RMIG and TDS metrics follow a similar trend, with DISCOV variants demonstrating superior performance, especially as L increases. These findings suggest that increasing the depth of the model architecture and incorporating DISCOV loss play pivotal roles in improving model accuracy and task discriminative capabilities, highlighting the significance of attention mechanisms in enhancing performance.

Method	Depth (L)	NRMSE ↓	RMIG ↓	TDS ↓
VAE (baseline)	-	0.928	0.921	0.935
VAE (baseline)+DISCOV	-	0.929	0.924	0.931
β-TCVAE	-	0.931	0.918	0.937
β-TCVAE+DISCOV	-	0.930	0.922	0.933
DIP-VAE	-	0.932	0.915	0.939
DIP-VAE+DISCOV	-	0.928	0.926	0.930
DISCOV	8	0.50	0.73	0.71
DISCOVw/o Attention	8	0.54	**0.71**	0.72
DISCOV	16	**0.49**	0.74	0.70
DISCOVw/o Attention	16	0.52	**0.72**	0.73
DISCOV	32	**0.48**	0.75	**0.69**

Table 2: Average Normalized RMSE, RMIG, and TDSScores for Variants DISCOVw/,w/o Attention, as L Increases. (↓ lower values are better [**Top-1**,**Top-2**], the Red row the worst on average, and the Blue the best).

(2) Robustness, Disentanglement, and Strong Generalization

Finding: *DISCOV demonstrates robust disentanglement performance across varying M.*

We report the disentanglement performance of DISCOV and FactorVAE on the Uk-dale dataset as we change λ and β. FactorVAE [18] is the closest TC-based method: it uses a single monolithic discriminator and the density-ratio trick to explicitly approximate $TC(\mathbf{z})$. Computing $TC(\mathbf{z})$ is challenging to compute as M increases. The average disentanglement scores for DISCOV $\lambda = 0.5$ and $\lambda = 0.6$ are very close, indicating that its performance is robust in disentangling. This is not the case for FactorVAE it performs worse on all metrics when m increases. Interestingly, FactorVAE seems to recover its performance on most metrics with higher β than is beneficial for FactorVAE.

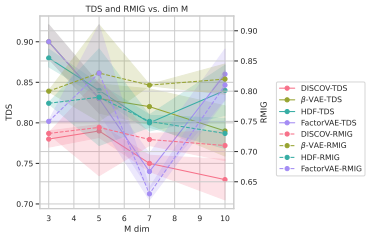

Figure 5: The resilience of various disentanglement metrics, as dimension M varies.

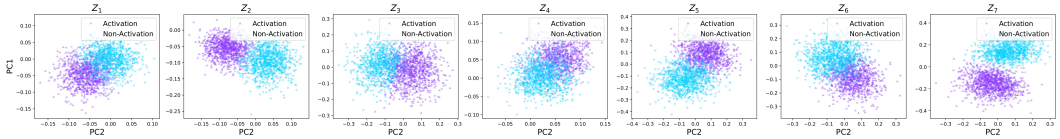

Figure 6: PCA visualization on the latent variable of DISCOV. A distinct separation between samples corresponds to the activation and inactivation of an appliance. The latent space exhibits clear distinguishability between instances of activation and non-activation of each appliance from the lefet to right: Washing Machine, Oven, Dishwasher, Cloth Dryer, and Fridge. (cyan represents non-activated samples, while blue-purple indicates the activation of appliances in those samples).

6 Related Work

Disentanglement Representation. Since the work of [5], many methods have been proposed to learn disentangled representations based on various heuristics [19, 12, 23, 28, 7]. Following the work of [36], which highlighted the lack of identifiability in modern deep generative models, many works have proposed more or less weak forms of supervision motivated by identifiability analyses [37, 25, 46, 1, 3, 54]. A similar line of work have adopted the causal representation learning perspective [30, 29, 34, 33, 2, 51, 8]. Recent work [9, 39] has produced promising results. However, they are confronted with problems of interpretability, generalization, and robustness. Various approaches have been proposed to solve these problems. For instance, [9] introduced Convolutional Neural Networks (CNNs) for feature extraction from power consumption data, showing promise on the UK-DALE dataset [20]. Generalization concerns persist despite leveraging Gated Recurrent Units (GRUs) and attention mechanisms. Other works attempt meaningful representation of time series, but disentangling remains challenging [47, 44, 38]. Recurrent VAE (RVAE) [14] for sequential data, D3VAE [32] improves prediction using a diffusion model after decoding the latent space. In representation learning, [53] employs contrastive learning, but in correlated data scenarios it is not explored. [35] based on specific propriety of time series like frequency and amplitude to disentangling Time series, but disentangling the latent space through data-driven methods poses a challenge. Nevertheless, recent approaches like Support Factorization as described in the works of [55, 42] show promise in addressing this challenge and have yielded encouraging results.

On The Non-Intrusive Load Monitoring and Representation Learning. Recent work [9, 39] has produced promising results for separation source power. Nevertheless, they encounter challenges related to generalization and robustness when confronted with out-of-distribution scenarios. Several approaches have been suggested to address these challenges. Some methods tackle them through either transfer learning or by enhancing the learned representations for each individual appliance. Exploring ways to enhance representation learning in this field has been the focus of recent studies [47, 44, 38]. However, achieving an informative and disentangled representation remains an open and challenging question. Existing models, like RNN-VAE [14] for sequential data and D3VAE [32], assume

statistically independent attributes. This assumption hampers their performance on real-world data and makes them less applicable to out-of-distribution scenarios. Developing models that effectively capture informative and disentangled representations in a realistic and versatile manner continues to be a significant challenge.

7 Conclusion

Enhancing the generalization capabilities and interpretability of current machine learning models are primary objectives when learning disentangled representations of time series data. However, existing methods for disentangling time series have primarily focused on identifying independent factors of variation, disregarding the fact that the causal factors underlying real-world data often exhibit correlations. To address this limitation and capture the correlated nature of real-world data, we propose an approach that enables the model to encode a diverse range of generative attributes in the latent space, thereby facilitating disentanglement. Our method, DisCo, demonstrates that promoting pairwise factorized support is adequate for traditional disentanglement techniques. Moreover, our results suggest that `DISCOV` competes effectively in downstream tasks, such as NILM methods, and achieves significant relative improvements of over +55% on standard benchmarks across datasets with varying correlation shifts.

8 Reproducibility Statement

We have made diligent efforts to ensure the reproducibility of our work. In the main paper, we describe every aspect of our learning objective and all the assumptions made. Concerning our experiments, all training details and dataset information are made available as open source at https://oublalkhalid.github.io/DISCOV/.

9 Acknowledgements

The authors would like to express their gratitude to the organizers of Unirep NeurIPS 2023 for their invaluable discussions, and to the Area Chair and anonymous reviewers for their valuable comments and insightful suggestions. This research was supported by access to the HPC resources of IDRIS under the allocation AD011014921 granted by GENCI (Grand Equipement National de Calcul Intensif). Additionally, this work received funding from the TotalEnergies Individual Fellowship through One Tech. Special thanks are extended to Thierry Luci, Head of the Applied Scientist AI Team at One Tech, for his leadership and support, as well as to the team for their active participation in insightful discussions.

References

[1] K. Ahuja, J. Hartford, and Y. Bengio. Properties from mechanisms: an equivariance perspective on identifiable representation learning. In *International Conference on Learning Representations*, 2022.

[2] K. Ahuja, J. Hartford, and Y. Bengio. Weakly supervised representation learning with sparse perturbations, 2022.

[3] K. Ahuja, D. Mahajan, V. Syrgkanis, and I. Mitliagkas. Towards efficient representation identification in supervised learning. In *First Conference on Causal Learning and Reasoning*, 2022.

[4] I. Apostolopoulou, I. Char, E. Rosenfeld, and A. Dubrawski. DEEP ATTENTIVE VARIATIONAL INFERENCE. 2022.

[5] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

[6] Y. Bengio, A. Courville, and P. Vincent. Representation Learning: A Review and New Perspectives, Apr. 2014. arXiv:1206.5538 [cs].

[7] D. Bouchacourt, R. Tomioka, and S. Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[8] J. Brehmer, P. De Haan, P. Lippe, and T. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 2022.

[9] G. Bucci, E. Fiorucci, S. Mari, and A. Fioravanti. A New Convolutional Neural Network-Based System for NILM Applications. *IEEE Transactions on Instrumentation and Measurement*, 2021.

[10] M.-A. Carbonneau, J. Zaidi, J. Boilard, and G. Gagnon. Measuring Disentanglement: A Review of Metrics, May 2022. arXiv:2012.09276 [cs].

[11] K. Chen, Q. Wang, Z. He, K. Chen, J. Hu, and J. He. Convolutional sequence to sequence non-intrusive load monitoring. *the Journal of Engineering*, 2018(17):1860–1864, 2018. Publisher: Wiley Online Library.

[12] R. T. Q. Chen, X. Li, R. G., and D. Duvenaud. Isolating sources of disentanglement in vaes. In *Advances in Neural Information Processing Systems*, 2018.

[13] R. T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[14] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A Recurrent Latent Variable Model for Sequential Data. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[15] K. Do and T. Tran. Theory and Evaluation Metrics for Learning Disentangled Representations, Mar. 2021. arXiv:1908.09961 [cs, stat].

[16] C. Eastwood and C. K. I. Williams. A FRAMEWORK FOR THE QUANTITATIVE EVALUATION OF DISENTANGLED REPRESENTATIONS. 2018.

[17] S. Elfwing, E. Uchibe, and K. Doya. Sigmoid-Weighted Linear Units for Neural Network Function Approximation in Reinforcement Learning, Nov. 2017. arXiv:1702.03118 [cs].

[18] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. Nov. 2016.

[19] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

[20] J. Kelly and W. Knottenbelt. The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Scientific data*, 2, 2015. Publisher: Nature Publishing Group.

[21] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

[22] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised Contrastive Learning, Mar. 2021. arXiv:2004.11362 [cs, stat].

[23] H. Kim and A. Mnih. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[24] H. Kim and A. Mnih. Disentangling by Factorising, July 2019. arXiv:1802.05983 [cs, stat].

[25] D. A. Klindt, L. Schott, Y. Sharma, I. Ustyuzhaninov, W. Brendel, M. Bethge, and D. M. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *9th International Conference on Learning Representations*, 2021.

[26] J. Z. Kolter and M. J. Johnson. REDD: A public data set for energy disaggregation research. In *Workshop on data mining applications in sustainability (SIGKDD), San Diego, CA*, volume 25, 2011. Issue: Citeseer.

[27] A. Kumar, P. Sattigeri, and A. Balakrishnan. VARIATIONAL INFERENCE OF DISENTANGLED LATENT CONCEPTS FROM UNLABELED OBSERVATIONS. 2018.

[28] A. Kumar, P. Sattigeri, and A. Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

[29] S. Lachapelle and S. Lacoste-Julien. Partial disentanglement via mechanism sparsity. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

[30] S. Lachapelle, P. Rodriguez Lopez, Y. Sharma, K. E. Everett, R. Le Priol, A. Lacoste, and S. Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *First Conference on Causal Learning and Reasoning*, 2022.

[31] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. Temporal Convolutional Networks: A Unified Approach to Action Segmentation, Aug. 2016. arXiv:1608.08242 [cs].

[32] Y. Li, X. Lu, Y. Wang, and D. Dou. Generative Time Series Forecasting with Diffusion, Denoise, and Disentanglement, Jan. 2023. arXiv:2301.03028 [cs].

[33] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. iCITRIS: Causal representation learning for instantaneous temporal effects. In *UAI 2022 Workshop on Causal Representation Learning*, 2022.

[34] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. CITRIS: Causal identifiability from temporal intervened sequences, 2022.

[35] S. Liu, X. Li, G. Cong, Y. Chen, and Y. Jiang. MULTIVARIATE TIME-SERIES IMPUTATION WITH DIS- ENTANGLED TEMPORAL REPRESENTATIONS. 2023.

[36] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[37] F. Locatello, B. Poole, G. Raetsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[38] L. Maaløe, M. Fraccaro, V. Liévin, and O. Winther. BIVA: A Very Deep Hierarchy of Latent Variables for Generative Modeling, Nov. 2019. arXiv:1902.02102 [cs, stat].

[39] C. Nalmpantis and D. Vrakas. On time series representations for multi-label NILM. *Neural Computing and Applications*, 32(23), 2020.

[40] A. v. d. Oord, Y. Li, and O. Vinyals. Representation Learning with Contrastive Predictive Coding, Jan. 2019. arXiv:1807.03748 [cs, stat] version: 2.

[41] K. Oublal, S. Ladjal, D. Benhaiem, F. Roueff, et al. Temporal attention bottleneck is informative? interpretability through disentangled generative representations for energy time series disaggregation. 2023.

[42] K. Roth, M. Ibrahim, Z. Akata, P. Vincent, and D. Bouchacourt. Disentanglement of Correlated Factors via Hausdorff Factorized Support, Feb. 2023. arXiv:2210.07347 [cs, stat].

[43] A. Vahdat and J. Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, 2020.

[44] A. Vahdat and J. Kautz. NVAE: A Deep Hierarchical Variational Autoencoder, Jan. 2021. arXiv:2007.03898 [cs, stat].

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[46] J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[47] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. Hoi. COST: CONTRASTIVE LEARNING OF DISENTANGLED SEASONAL-TREND REPRESENTATIONS FOR TIME SERIES FORECASTING. 2022.

[48] M. Yang, X. Li, and Y. Liu. Sequence to Point Learning Based on an Attention Neural Network for Nonintrusive Load Decomposition. *Electronics*, 2021.

[49] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9588–9597, Nashville, TN, USA, June 2021. IEEE.

[50] Y. Yang, J. Zhong, W. Li, T. A. Gulliver, and S. Li. Semisupervised Multilabel Deep Learning Based Nonintrusive Load Monitoring in Smart Grids. *IEEE Transactions on Industrial Informatics*, 16(11):6892–6902, Nov. 2020.

[51] W. Yao, Y. Sun, A. Ho, C. Sun, and K. Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022.

[52] Z. Yue, C. R. Witzig, D. Jorde, and H.-A. Jacobsen. BERT4NILM: A Bidirectional Transformer Model for Non-Intrusive Load Monitoring. In *Proceedings of the 5th International Workshop on Non-Intrusive Load Monitoring*, NILM'20, pages 89–93, New York, NY, USA, Nov. 2020. Association for Computing Machinery.

[53] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow Twins: Self-Supervised Learning via Redundancy Reduction, June 2021. arXiv:2103.03230 [cs, q-bio].

[54] Y. Zheng, I. Ng, and K. Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, 2022.

[55] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel. Contrastive Learning Inverts the Data Generating Process, Apr. 2022. arXiv:2102.08850 [cs].