A Causal Intervention Scheme for Semantic Segmentation of Quasi-Periodic Cardiovascular Signals

Xingyao Wang, Student Member, IEEE, Yuwen Li, Hongxiang Gao, Xianghong Cheng, Jianqing Li, and Chengyu Liu, Senior Member, IEEE

Abstract—Precise segmentation is a vital first step to analyze semantic information of cardiac cycle and capture anomaly with cardiovascular signals. However, in the field of deep semantic segmentation, inference is often unilaterally confounded by the individual attribute of data. Towards cardiovascular signals, quasi-periodicity is the essential characteristic to be learned, regarded as the synthesis of the attributes of morphology ($A_m$) and rhythm ($A_r$). Our key insight is to suppress the over-dependence on $A_m$ or $A_r$ while the generation process of deep representations. To address this issue, we establish a structural causal model as the foundation to customize the intervention approaches on $A_m$ and $A_r$, respectively. In this article, we propose contrastive causal intervention (CCI) to form a novel training paradigm under a frame-level contrastive framework. The intervention can eliminate the implicit statistical bias brought by the single attribute and lead to more objective representations. We conduct comprehensive experiments with the controlled condition for QRS location and heart sound segmentation. The final results indicate that our approach can evidently improve the performance by up to 0.41% for QRS location and 2.73% for heart sound segmentation. The efficiency of the proposed method is generalized to multiple databases and noisy signals.

Index Terms—Cardiovascular signal, semantic segmentation, QRS-complex, heart sound, representation learning, causal intervention.

Manuscript received 13 February 2022; revised 22 December 2022; accepted 20 April 2023. Date of publication 27 April 2023; date of current version 3 July 2023. This work was supported in part by the National Natural Science Foundation of China under Grants 62001111, 62171213, 62071241, and 81871444, in part by the National Key Research and Development Program of China under Grant 2019YFE0113800, and in part by the Natural Science Foundation of Jiangsu Province under Grants BK20200364, BK20190014, and BK20192004. (Corresponding authors: Xianghong Cheng; Chengyu Liu.)

Xingyao Wang and Chengyu Liu are with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China, and also with the State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China (e-mail: xingyao@seu.edu.cn; chengyu@seu.edu.cn).

Yuwen Li, Hongxiang Gao, Xianghong Cheng, and Jianqing Li are with the School of Instrument Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: liyuwen@seu.edu.cn; hx_gao@seu.edu.cn; li1005578@seu.edu.cn; ljq@seu.edu.cn). Digital Object Identifier 10.1109/JBHI.2023.3270978.

I. INTRODUCTION

The cardiovascular signals implicate rich information about the heart circulation system, including electrocardiograph (ECG), phonocardiogram (PCG) and photoplethysmographic (PPG), etc., commonly used as non-invasive means for monitoring cardiovascular system and diagnosis of organic heart disease and cardiac electrophysiological abnormalities. For paroxysmal arrhythmia and various invisible heart diseases, long-term dynamic monitoring has become an indispensable supplement to conventional test. Automatic analysis with these signals is crucial to alleviate workload for cardiologists, especially for long-term dynamic monitoring, such as Holter or wearable ECG. The first and most critical step for automatic diagnosis is high-precision semantic segmentation of the physiological signal, since the error will be counted up to the subsequent stages.

In clinical applications, peculiarly in dynamic environment, the temporal physiological signals are susceptible to interference from noise and individual variability. Due to the low dominant frequency of the target components, the state identification is always confused by intra-bandpass noise. In the past, the researchers concentrated on the preprocessing and feature extraction of such physiological signal to improve the segmentation performance [1], [2], [3]. The essence of these methods is to amplify the inter-state difference and discrepancy between target signals and noises, such as calculating the slope change and wavelet transforming to locate the QRS-complex and P waves in ECGs [4], [5] and fundamental heart sounds in PCGs [6], [7], modeling PPGs with Gaussian functions [8], [9], etc.. Nonetheless, these classic methods can only deal with static scenes with single-source noise and non-severe variations. The recent researches indicate that the supervised machine learning methods are capable of significantly improving the segmentation performance for pseudo-periodic physiological signals, and are more robust in dynamic databases [10], [11].

Pseudo-periodic is an exclusive characteristics of cardiovascular signals, which are epitomized to two attributes, attribute of rhythm ($A_r$) and attribute of morphology $A_m$, as follows:

- $A_m$: A signal segment of a state needs to own the general morphological characteristics of that state in all cardiovascular signals.
- $A_r$: The dominant frequency of the target components.
A segment of a state needs to obey a repetitive pattern of that state in the same cardiovascular signal. In most cases, the cardiovascular signals naturally contain these two attributes and they are mutual independent. However, due to the abnormalities in electrophysiological activity, such as cardiac arrest, ventricular tachycardia, atrioventricular block, etc., or noise disturbances, such as leads failing, motion artefacts, etc., $A_r$ and $A_m$ would be modified. In Fig. 1, we respectively list two scenarios in QRS-complex location that $A_m$ and $A_r$ hijacks the inference of the segmentation model, respectively.

Assuming $Z$ the deep representation of an ECG episode, $A_m$ and $A_r$ should be joint dependencies of $Z$, yet are highly coupled in the latent space, causing over-dependence of $Z$ on the onefold attribute. In this work, we propose a solution to eliminate the individual effect from $A_m$ and $A_r$ and the intuitive thought is to intervene the attributes in latent space.

In [12], the Independent Causal Mechanisms (ICM) Principle was proposed as follows: The causal generative process of a system’s variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms. Applied to the segmentation of cardiovascular signals, this principle tells us that knowing one of $P(A_r|A_r)$ and $P(Z|A_m)$ does not give any information about the other.

In this article, we propose a novel contrastive learning framework combined with frame-level causal intervention for semantic segmentation of cardiovascular signals, contrastive causal intervention (CCI). There are four main contributions in this article:

1) We establish a structural causal model to depict the implicit dependency relationship between abstracted attributes and the latent representations.

2) A frame-level contrastive training strategy based on the proposed CCI is designed to implement the intervention paradigm on $A_m$ and $A_r$.

3) We evaluate CCI on two classic tasks of cardiovascular signal segmentation, QRS location and heart sound segmentation, and comprehensive experiments for measuring the segmentation performance are implemented on a large number of independent test sets.

4) Additional analytical results including a real-world noise stress test and visualization of latent distributions are presented to illustrate how and why CCI improves robustness and generalization of the segmentation model.

II. RELATED WORK

Time Series Semantic Segmentation: The common Encoder-Decoder architecture for segmentation task ensures the inherent tension between semantics and location in the training process, which allows researchers to develop different variants of the Encoder structures [13], [14], [15], [16] for more efficient feature fusion. According to existing researches, fully convolutional network (FCN) has been proved a superior performance in semantic segmentation task [17] with controllable computational cost. Subject to the receptive fields, the performance bottleneck has been raised due to the lack of capability for learning long-range dependency information in unconstrained scene images [18] and particularly in time series [19]. To address the limited learning ability of contextual information, DeepLab [16], [20] introduce the dilated convolution to enlarge the receptive field. Alternatively, context modeling is the focus of PSPNet [21] and DeepLabV2 [16]. Decomposed large kernels [22] are also utilized for context capturing. In temporal segmentation, to expand multi-scale receptive fields and leverage the inherent temporal relation, a reasonable approach is to disassemble the network into multi-branches to expand multi-scale receptive fields [11] or distribute sub-networks at each time step [23], [24], [25]. For multi-state segmentation in pseudo-periodic signal, variants of recurrent neural network (RNN) [26], [27] and dynamic inference [28] are utilized for learning state transition probability.

Causal Representation Learning: Although methods for learning causal structure from observations exist [29], [30], [31], variables in a causal graph may be unobserved or unquantifiable (i.e. NN representations), which can make causal inference particularly challenging. It is inevitable to arise statistical dependence caused by internal causal relations so that destruct performance of current machine learning methods, since the i.i.d. assumption is violated. There has been a growing amount of efforts in performing appropriate interventions in several tasks, including image classification [32], visual dialog [33] and scene segmentation [34]. Another dilemma is the entangled factorizations in the latent space, inducing the indecomposable causal mechanisms. Disentanglement of causal effects is crucial for introduction of structural causal models. Recent works are concentrated on disentangled factorization in the latent space while changing background conditions [35], [36], on the basis of the invariance criterion of causal structure.

Mutual Information Estimation: To obtain differentiable and scalable MI estimation, recent approaches utilize deep neural networks to construct variational MI estimators. Barber-Agakov (BA) bound for MI [37] firstly propose to approach the difficulty of computing MI by using a variational distribution. Most of these estimators focus on MI maximization problems through providing MI lower bound. A mainstream method is to treat MI as the Kullback-Leibler (KL) divergence between the joint and marginal distribution and convert it into the dual representation.
Based on this kernel, great efforts have been paid to explore more appropriate transformations and critics using neural networks [38], [39], [40]. Instead of MI maximization, in this article we explicitly use MI upper bound for MI minimization. Most existing MI upper bounds for $I(x; y)$ require the conditional distribution $p(y|x)$ or $P(x|y)$ to be known. Since it is unpractical in most machine learning tasks, multi variational upper bounds were explored [41], also with a Monte Carlo approximation [42].

III. METHOD

A. Notations

Let $X = \{x^0, x^1, \ldots, x^T\}$ be a cardiovascular signal instance with $T$ frames and $Z = \{z^0, z^1, \ldots, z^T\}$ be the corresponding latent feature space, where $z^T = \{z^0, z^1, \ldots, z_T\}$ is a feature vector with $d$ dimensions. In this article, we focus on solving the bias problem induced by the two attributes $A_r$ and $A_m$. A prior hypothesis is proposed as that $A_m$ is distilled from the short-term frame $x_r$ and $A_r$ from the global distribution $P(X)$.

To better understand the causal mechanism and the confounding source, we choose to insert a mediation $A_{mr}$ which defined as follows:

$A_{mr}$: The morphology pattern of each type of state recurs within the same episode and obeys the data distribution of the state-specific waveform family.

As all the attributions are constructed in the cognitive space, we assume the ultimate $Z^r$ for each frame is yielded by $f(A^r_{mr}, U_i)$, where $U_i$ is noise, not providing any information in the latent space. As shown in the proposed structural causal model (SCM) (Fig. 2), $A^r_{mr}$ and $A_r$ are the parent nodes of $A^r_{mr}$, the only dependence of the representation of the $r$th frame, $Z^r$.

In Section 3B, we intervene on $A_m$ and $A_r$ frame by frame respectively to estimate and constrain their direct effect on $z^r$. Thus for each parent attribute, the do-operation will generate a signal set with $T$ variants. Since the perturbation is adopted on each frame, we define the do-variant as $X^\tau \{x^0, x^1, \ldots, x^T\}$, representing the intervention is on the $\tau$th frame. According to the previous definition, $A_m$ represents the local semantic information of a specific state and $A_r$ indicates the global morphology information of a recurring pattern among an episode.

B. Causal Formulation

1) Structural Causal Model: In Fig. 2, we show the generation process of the frame-level representation $Z^r$ generated by the segmentation model for cardiovascular signals from the perspective of causal inference. For most cases, $A_r$ and $A_m$ refined from biased distribution would have the probability degrade the performance of segmentation in signals that are inconsistent with the training distribution. In the proposed SCM, we can clearly see how $A^r_{mr}$ confounds $Z^r$ and $A^r_{mr}$ via the backdoor paths, $z^r \leftarrow A^r_{mr} \rightarrow A^r_{mr}$. Similar causal mechanism exists for $A_r$ via another backdoor path $z^r \leftarrow A_r \rightarrow A^r_{mr}$. We expect to cut off the direct causal links, $A^r_{mr} \rightarrow z^r$ and $A_r \rightarrow z^r$. However, $A_r$ and $A^r_{mr}$ are highly coupled in the latent space while the fitting process, and decoupling out $A_r$ and $A^r_{mr}$ from representations is expensive. Therefore, we choose to perform a constraint while fitting process to reduce the straight influence from $A_r$ and $A^r_{mr}$ to $Z^r$, and the first step is to measure or estimate what degree the model discriminates by $A_r$ and $A^r_{mr}$ in the generation process of $Z$.

2) Causal Intervention Formulation: It is unsteady to do condition on $A^r_{mr}$, as the $A^r_{mr} \rightarrow Z^r$ is confounded by $A_r$ and $A^r_{mr}$, thus a more reasonable manner is to intervene on it. We show a case study for estimating the direct effect of $A^r_{mr} \rightarrow Z^r$ in the following content (the same goes for $A_r \rightarrow Z^r$ relation). According to the ICM principle, there are two sides of conceptions when observe whether the causal link $A^r_{mr} \rightarrow Z^r$ exist, which are shown as follows:

a) The statistical distribution of $z^r$ should not be varying with the change of $A^r_{mr}$ while holding $A^r_{mr}$ steady.

b) The statistical distributions of $z^r$ with mutual independent $A^r_{mr}$ should be irrelevant even with a steady $A^r_{mr}$.

For a, we can fabric the conditional distribution of $Z^r$ through changing $A^r_{mr}$ from $a^0_{mr}$ to $a^T_{mr}$, which is defined as:

$$P_{\theta}^{do(A^r_{mr})} = P_{\theta}(Z_r \mid do(A^r_{mr} = a^r_{mr}), do(A^r_{mr} = a^r_{mr})).$$

Since there is no backdoor path from $A^r_{mr}$ to $z^r$ and $A_r$ is another confounder for $A^r_{mr} \rightarrow z^r$, we can block the other backdoor path through adjusting $A_r$, which gives:

$$P_{\theta}^{do(A_r)} = \int A_r \int P_{\theta}(Z_r \mid A^r_{mr} = a^r_{mr}, A^r_{mr} = a^r_{mr}, A_r = a_r) P(A_r = a_r) .$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
According to \( \mathbf{a} \), \( P^\text{do}(A_m) \) and \( P^\text{do}(A'_m) \) should be consistent, inducing the objective function:

\[
L_m = \min_\theta D_{KL} \left( P^\text{do}(A_m), P^\text{do}(A'_m) \right). \tag{5}
\]

Unfortunately, \( A_r \) is an abstract attribute with an infinite distribution and there is no feasible way to traverse the whole \( A_r \) space. Thus adjusting \( A_r \) to block \( z^T \leftarrow A_r \rightarrow A'_m \text{ } \text{do} \) is unprocurable. Conception \( \mathbf{b} \) provides an inverse logic to hold \( A'_m \) steady instead of \( A'_m \), that is the statistical characteristics of \( Z^T \) depend only on \( A'_m \) regardless of whether \( A_m \) changes. Intuitively speaking, \( A_m \) provides no direct information for \( Z^T \).

If we choose to intervene on \( A'_m \), since there is no backdoor path from \( \bar{A}_m \) to \( Z^T \) in the model, hence we can replace \( \text{do}(a_m) \) with simply conditioning on \( a_m \). The conditional distributions of \( Z_r \) are given as follows:

\[
P^\text{do}(A_m) = P \left( Z_r \mid \text{do}(A_m) = a_m, A'_m = a'_m \right) = P \left( Z_r \mid A'_m = a_m, A'_m = a'_m, A_r = a_r \right), \tag{6}
\]

\[
P^\text{do}(A'_m) = P \left( Z_r \mid \text{do}(A'_m) = a'_m, A'_m = a'_m, A_r = a'_r \right) = P \left( Z_r \mid A'_m = a'_m, A'_m = a'_m, A_r = a'_r \right). \tag{7}
\]

We expect the representation of the target frame with different \( A'_m \) should not derive correlation induced by the invariant \( A'_m \). Here MI is adopted to measure the degree of correlation of the two representations and minimized as a constraint on training. The object function is defined as:

\[
L_m = \min_\theta I \left( P^\text{do}(A_m), P^\text{do}(A'_m) \right). \tag{8}
\]

Symmetrically, we can draw the paradigms for intervention on \( A_r \) and the corresponding object function as:

\[
P^\text{do}(A_r) = P \left( Z_r \mid A_r = a_r, A'_m = a'_m, A'_m = a_m \right), \tag{9}
\]

\[
P^\text{do}(A'_r) = P \left( Z_r \mid A_r = a'_r, A'_m = a'_m, A'_m = a'_m \right), \tag{10}
\]

\[
L_r = \min_\theta I \left( P^\text{do}(A_r), P^\text{do}(A'_r) \right). \tag{11}
\]

3) Intervention Scheme: In this section, we take scenarios of QRS location in an ECG episode to illustrate how to intervene on the two attributes, \( A_m \) and \( A_r \). The first step is to define the associated physical transformation with the controlled \( \text{do} \)-operations. As previously mentioned, \( A'_m \) indicates the state distribution of the local waveform morphology and \( A_r \) the global recurring pattern distribution.

For \( \text{do}(A_m) \), according to (7), we need to solve out how to maintain the subordinate state properties of the local morphology while changing \( A_m \) and \( A'_m \). Here we conduct a straightforward manner of reversing phase (amplitude inversion) on the target frame (as shown in Fig. 3(a)). \( A_r \) is a global attribute, representing how the contextual morphology pattern influence the target frame. Reversing the QRS-complex morphology on the target frame will definitely affect \( A_r \) accordingly.

For \( \text{do}(A_r) \), according to (10), we wish to alternate the state of the target frame while not changing the global recurring pattern. Here we perform a handy intervention, that is zero setting on the chosen target frame. As shown in Fig. 3(b), we simply erase the morphology information of the target frame, not introducing extraneous signals. Since no additional morphological information is introduced, \( A_r \) can be approximately regarded as invariant as \( A_m \) altered.

In practical operation, for the same cardiovascular signal, we performed the above intervention in units of a frame with fixed length. Assuming the signal owns \( T \) frames, given binary masks \( x_{mask} \) with \( T \) dimensions, \( x \otimes x_{mask}[\tau] \) indicates that the \( \tau \)-th frame is set to zeros. Then we have \( \text{do}(A_m)(x) = \{x \otimes x_{mask}[\tau]\}_{\tau=1}^{T} \) and \( \text{do}(A_r)(x) = \{x \otimes (1 - x_{mask}[\tau])\}_{\tau=1}^{T} \).

C. Contrastive Framework for Causal Intervention

In the previous sections, we have confirmed to utilize MI modeling the causal interventions and the specific operations. In this section we will establish the framework so that the intervention of the target frame can form effective constraint while training. Here we adopt the contrastive architecture with a shared-weights Encoder (E) and a Decoder (D), where we should
learn representations from E to separate (contrast) original samples and intervened samples. The designed temporal contrastive learning module is shown in Fig. 3(c).

The general contrastive loss is designed to learn feature representation for positive pairs to be similar, while pushing features from the randomly sampled negative pairs apart. Unlike the conventional contrastive paradigm, the proposed contrastive method should weaken the relevance between representations before and after the intervention, namely the negative pairs in the classic contrastive conception. According to the assumed attributes and causal inference, the frames beside the intervened target frame share the same attribute $A_{int}$ and should own the consistent distributions in the latent space. Thus we deemed these pairs of untreated frames as the positive pairs. The ultimate contrastive paradigm should be:

$$\min_{E,D} \mathcal{L}_{Seg} + \lambda_1 I(z, z^d) - \lambda_2 \frac{1}{T-1} \sum_{i=0,i \neq \tau}^T I(z_i, z^d_{\tau})$$

(12)

Suppose the optimal representations of the frames with the same $A_{int}$ should be completely consistent, i.e., $P(Z_i = Z^d_i) = 1$, then maximizing the MI between these frames can be substituted by cosine similarity distance:

$$\max_{z, z^d} I(z, z^d) \iff \max_{E} \frac{z_i}{\|z_i\|_2} \cdot \frac{z^d_{\tau}}{\|z^d_{\tau}\|_2}$$

(13)

D. Mutual Information Upper Bound Estimation

Denote $P_E(Z|X)$ the distribution of the encoded representation for the original signal, and $P_E(Z|X^d)$ the representation for the intervened signal. For convenience, we apply $P(Z)$ and $P(Z^d)$ representing $P_E(Z|X)$ and $P_E(Z|X^d)$, respectively. The proposed approach to estimate MI upper bound follows Contrastive Log-ratio Upper Bound (CLUB) [41], which estimates MI through narrowing the gap of conditional probabilities between positive and negative sample pairs.

Difference exists that the intervention is operated frame by frame. For the whole do-operations of the same attribute of a signal episode, the conditional distributions $P(Z | Z^d)$ should be uniformed since they are homogeneous. According to CLUB, a certified unbiased MI upper bound estimation is proposed with $N$ sample pairs $(z_i, z_i^d)_{i=1}^N$ and $T$ frames for each $z_i$ as follows:

$$I_{CCCI} = \frac{1}{N} T \sum_{i=1}^N \sum_{\tau=1}^T \left[ \log p(z_i \mid z_i^d) - \log p(z^d_{\tau} \mid z_i^d) \right].$$

(14)

Unfortunately, $p(z_i \mid z_i^d)$ is unknown so that a variational approximation of the distribution is given as $q_{\theta}(z_i \mid z_i^d)$. Thus, we have the variational upper bound estimation for $I_{CCCI}$:

$$I_{vCCCI} = \frac{1}{N} T \sum_{i=1}^N \sum_{\tau=1}^T \left[ \log q_{\theta}(z_i \mid z_i^d) - \log q_{\theta}(z^d_{\tau} \mid z_i^d) \right].$$

(15)

The prerequisite for the establishment of $I(Z; Z^d) \leq I_{vCCCI}$ is proved to be:

$$KL(p(z_i \mid z_i^d) || q_{\theta}(z_i \mid z_i^d)) \leq KL(p(z_i \mid z_i^d) || p(z_i \mid z_i^d))$$

(16)

where $q_{\theta}(z_i \mid z_i^d) = q_{\theta}(z_i \mid z_i^d)$ is the variational joint distribution induced by $q_{\theta}(z_i \mid z_i^d)$. And $KL(p(z_i \mid z_i^d) || q_{\theta}(z_i \mid z_i^d))$ can be minimized by maximizing the log-likelihood of $q_{\theta}(z_i \mid z_i^d)$. For $I_{vCCCI}$, it is a cross-frame function $L_{\theta}(\theta_q) = \frac{1}{T} \sum_{i=1}^N \sum_{\tau=1}^T \log q_{\theta}(z_i \mid z_i^d)$. A prior Gaussian distribution is provided to solve $q_{\theta}(z_i \mid z_i^d)$. Here we assume that $q_{\theta}(z \mid z_i^d) = N(z \mid \mu(z_i^d), \sigma^2(z_i^d))$. For given samples $(z_i, z_i^d)_{i=1}^N$, we denote $\mu_{\tau} = \mu(z_i^d)$ and $\sigma_{\tau} = \sigma(z_i^d)$. Then we have

$$q_{\theta}(z \mid z_i^d) = \frac{1}{2\pi \sigma_{\tau}^2} \exp \left\{ -\frac{(z - \mu_{\tau})^2}{2\sigma_{\tau}^2} \right\}.$$ 

(17)

Thus the upper bound of the MI between the origin and intervened representation can be solved while training, which is shown in Algorithm 1 in detail.

### IV. EXPERIMENTS AND ANALYSIS

In this section, we conduct comprehensive experiments with the aim of answering the following three key questions.

Q1: What is the role of the proposed intervention approach on each attribution of cardiovascular signals (i.e., the ablation studies of our CCCI)?

---

Algorithm 1: Training Procedure for CCCI

**Require:** $D$: training set

**Require:** $\alpha, \beta, \lambda_1, \lambda_2, N$: batch size, $T$: number of frames

1: **Initialization:** $\theta_1, \theta_2, \theta_3$
2: **while** not converge **do**
3: Sample $\{x^i, y^i\}_{i=1}^N$ from $D$
4: $z_i \leftarrow f_{\theta_1}(x_i)$
5: $y_i \leftarrow g_{\theta_3}(z_i)$
6: for $\tau \leftarrow 1$ to $T$ do
7: $x^{d\tau}_i \leftarrow do(x_{\tau})$
8: $z^{d\tau}_i \leftarrow f(x^{d\tau}_i)_{\theta_e}$
9: Log-likelihood
10: $L_q(\theta_q) = \frac{1}{T} \sum_{i=1}^N \sum_{\tau=1}^T \log q_{\theta}(z_i \mid z_i^d)$
11: **end for**
12: Sampling $k'$ uniformly from $\{1, 2, \ldots, N\}$
13: $U_{\tau} = \log q_{\theta}(z_i \mid z_i^d) - \log q_{\theta}(z_{i\tau} \mid z_i^d)$
14: **end for**
15: $\sum_{i=1}^{T} \sum_{\tau=1}^T U_{\tau}$
16: $\mathcal{L}_{Seg} = \frac{1}{T} \sum_{i=1}^N \sum_{\tau=1}^T D_{CE}(y_i, \hat{y}_i)$
17: $\mathcal{L}_{Sim} = \frac{1}{T} \sum_{i=1}^N \sum_{\tau=1}^T T(z_{ij}, z_{ij}^d)$
18: $\mathcal{L} = \mathcal{L}_{Seg} - \mathcal{L}_{Sim}$
19: **Update** $\theta_e' \leftarrow \theta_e - \beta \nabla \theta_e \mathcal{L}(\theta_e)$
20: **Update** $\theta'_d \leftarrow \theta_d - \beta \nabla \theta_d \mathcal{L}_{Seg}(\theta_d)$
21: **end for**
22: **end while**
Q2: In what domain does the proposed method improve the segmentation performance (i.e., data with physiological variation and noise contamination)?

Q3: How does the proposed contrastive causal intervention influences the generation process of latent representations?

A. Experiment Setups

1) Database:

a) QRS location: We use CPSC2019-Train database [43] for training and five other open-access ECG databases with beat annotation in the experiments, including CPSC2019-Test, MIT-BIH 2-lead Arrhythmia Database (MITDB) [44], QT 2-lead Database (QTDB) [45] and INCART 12-lead Arrhythmia Database (INCART) [46]. The beat annotations in MITDB, QTDB and INCART are on the basis of multi-lead ECGs. Since the proposed CCI scheme is aimed to address the segmentation task for the single-channel cardiovascular signals, we evaluated the effectiveness of CCI for QRS-complex location on each lead. The summary of the ECG databases for training and testing is shown in Table I.

b) Heart sound segmentation: We selected 100 recordings randomly from training-a in PhysioNet/CinC Challenge 2016 [47] and slice them into 5-second samples for training. The remaining recordings in training-a and other data sets including training-b ‘f and hidden test sets (Test-b ‘e, Test-g and Test-i) from PhysioNet/CinC Challenge 2016 are utilized for testing. They are restructured into one database, which is named as PhysioNet-Test. Meanwhile, in order to comprehensively evaluate the impact for the segmentation model trained with CCI on pathological heart sounds, we introduced the abnormal heart sound databases constructed in our previous work [10] as the supplementary test sets. In [10], we designed several indicators to reflect the characteristics of heart sounds and constructed the three-level data sets from normal to pathological, including LEVEL-I, LEVEL-II and LEVEL-III. In LEVEL-II 16.5% of the records are with severe noise and murmur and 13.5% with arrhythmia while the proportion is 70% and 30% in LEVEL-III. More detailed information can be found in [10].

c) Noise Stress Test: Two main noise databases are utilized in our experiments. For QRS location, the three noise records from MIT-BIH Noise Stress Database [48] are chosen to test the method’s robustness when facing the real-world ECG noise, including baseline wander (bw), muscle (EMG) artifact (ma), and electrode motion artifact (em). Except for the three types of noises, power line interference noise is also utilized in noise stress test for QRS location. For heart sound segmentation, we focus on the influence brought by lung sounds recorded from the electronic stethoscope simultaneously. The lung sound samples are extracted randomly from the database constructed in [49]. The noisy records were synthesized from CPSC2019-Test for QRS location and Training-A from PhysioNet/CinC Challenge 2016 (other than the 100 records for training) by adding different types of noises. The noise stress test was conducted under signal-noise-ratio (SNR) controlled condition and the specific SNRs were 0, 5, 10, 15, 20 dB.

2) Pre-Processing and Post-Processing: The pre- and post-processing in the experiments is designed to be plain and unified for the backbone model training with and without CCI proposed in this article.

a) QRS location: Considering the energy of QRS-complex is mainly concentrated at 8–50 Hz [50], we perform band-pass filtering from 0.5–50 Hz as well as mean filtering on each 10-second episode. Since the magnitudes are not uniform across databases, standardization is conducted on ECG records after filtering and the input episodes are re-sampled at 250 Hz. The ultimate outputs of the segmentation model are activated by Sigmoid function, approximated to the probability of the corresponding time step belonging to the QRS-complex. Thus the decision of QRS-complex is to find candidate intervals with consecutive probabilities over a fixed threshold of 0.5. Referring to the effective refractory period in ECGs, some of the intervals will be excluded if they are less than 200 ms.

b) Heart sound segmentation: The majority of the frequency content in S1 and S2 sounds is below 150 Hz, usually with a peak around 50 Hz [51], and murmur is around 400 Hz. Thus, all the heart sound recordings were downsampled into 800 Hz. Moreover, different digital stethoscopes vary widely in response of heart sound and noise. Therefore, we adopted an adaptive local Wiener filter proposed in [10] to suppress the in-band noise from system and increase the amplitude resolution of alternate segments between heart sound states. The outputs of the models are functioned by the Softmax activation and the the time step is assigned the state with the maximal probability. We only determine the onsets of S1, systole, S2 and diastole by positioning the alternating time steps.

3) Evaluation Metrics: For evaluation, Sensitivity (Se), positive predictive rate (P+), error rate (Er) and F1 are calculate in all databases. These metrics are defined as follows:

\[
Se = \frac{TP}{TP + FN} \times 100\% ,
\]

\[
P_+ = \frac{TP}{TP + FP} \times 100\% ,
\]

\[
Er = \frac{FP + FN}{TP + FP + FN} \times 100\% ,
\]

\[
F_1 = \frac{2 \times SE \times P_+}{SE + P_+} \times 100\% ,
\]

where \( TP \) is true positive, \( FP \) is false positive and \( FN \) is false negative. The standard grace period of 150 ms is used for beat-by-beat comparison in QRS location [52] and 100 ms for state-by-state comparison in heart sound segmentation [6].

| Database            | Total Records | Total Beats | Database            | Total Records | Total Beats |
|---------------------|---------------|------------|---------------------|---------------|------------|
| CPSC2019-Train      | 2,000         | 19,655     | PhysioNet-Train     | 1000          | 3,688      |
| CPSC2019-Test       | 3,132         | 47,231     | PhysioNet-Test      | 1,923         | 51,059     |
| MITDB               | 48            | 109,494    | LEVEL-I             | 200           | 2,296      |
| QTDB                | 105           | 3,622      | LEVEL-II            | 200           | 2,439      |
| INCART              | 75            | 175,918    | LEVEL-III           | 150           | 1,906      |
### Implementation Details

In this work, the Decoders for the two tasks are fixed with two-layer dense block. A multi-branch 1D convolutional neural network (MBCNN) architecture is adopted as the backbone Encoder to comprehensively analyze CCI’s performance. MBCNN can distribute varying receptive fields into different branches as necessary to merge the full contextual information and avoid bloating due to long sequence inputs. Meanwhile, we compare various baseline networks with 1D convolution and dense layers for Encoder, including DenseNet [53] and SENet [54]. Except for the two common neural networks, we also evaluate the efficiency of CCI while training with the state-of-the-art models, CRNN [11] for QRS complex location and temporal-framing adaptive network (TFAN) [10] for heart sound segmentation. Since CRNN and TFAN are both seq2seq models, they can be trained with frame-by-frame constraint of deep representations under CCI scheme. More specifically, we trained the final dense layer of CRNN and TFAN as Decoder and the previous layers as Encoder. CCI was implemented to constrain the representations generated from these Encoders.

The training set was sliced into 5 folds for training and evaluation. We unified the training settings and hyperparameters for each task. The batch size was set to 100 for QRS-complex location and 50 for heart sound segmentation. As illustrated in Alg. 1, we adopted two optimizers, one with learning rate of \( \alpha \) to update \( \theta_q \) and the other with \( \beta \) to update \( \theta_e \) and \( \theta_d \). The both optimizers were Adam optimizer and the two learning rate, \( \alpha \) and \( \beta \), were both set to 0.001. For \( L_{Seg} \), we utilized binary cross entropy for QRS location and softmax cross entropy for heart sound segmentation. An early-stopping training strategy was adopted as follows: when the model failed to achieve the best validation accuracy in 10 consecutive epochs, the training is terminated.

For the value setting of the two loss weights in (12), we traced the impact of different \( \lambda_1 \) and \( \lambda_2 \) on the model performance based on the 5-fold cross validation. As shown in Fig. 4, we found out that the performance was insensitive to the two loss weights. The possible reason for this phenomenon is that \( z \) and \( z^{do} \) are generated by shared Encoder. The essence of CCI is an adversary learning to constrain deep representations instead of multi-task learning. Thus no matter how \( \lambda_1 \) and \( \lambda_2 \) be tuned, the second and the third loss term in (12) are going to balance out eventually. To avoid ambiguity, \( \lambda_1 \) and \( \lambda_2 \) were both set to the same value of 0.5.

### Ablation Studies for Causal Intervention

To understand the assumed causal mechanisms and the respective effects of interventions on \( A_r \) and \( A_m \) in (5) and (11) while fitting process, we conduct ablation studies on the two pseudo-periodic segmentation tasks, QRS-complex location and heart sound segmentation. For the multi-lead ECG records in the test sets, each lead is deemed as a single-lead ECG, sharing the ground truth of QRS locations while testing. We firstly conduct ablation studies on the attributes for training with CCI with the backbone Encoder, MBCNN. The ablation results for QRS-complex location and heart sound segmentation are shown in Tables II and IV, respectively. Then we implement DenseNet, SENet and the two SOTA models for QRS-complex location and heart sound segmentation as different Encoders to evaluate the adaption of CCI. The corresponding results are shown in Tables III and V.

#### Table II: Ablation Results (%) for QRS Location with MBCNN as the Backbone Encoder

| Database       | Method            | \( SE \) | \( F_1 \) | \( P_1 \) |
|----------------|-------------------|---------|---------|---------|
| CPSC2019-Test  | MBCNN             | 98.79   | 99.09   | 98.94   |
|                | MBCNN+CCI (\( A_m \)) | 99.32   | 99.31   | 99.31   |
|                | MBCNN+CCI (\( A_r \)) | 99.28   | 99.37   | 99.32   |
|                | MBCNN+CCI (\( A_m \) & \( A_r \)) | 99.26   | 99.45   | 99.35   |
| MITDB          | MBCNN             | 99.20   | 99.44   | 99.32   |
|                | MBCNN+CCI (\( A_m \)) | 99.37   | 99.49   | 99.43   |
|                | MBCNN+CCI (\( A_r \)) | 99.41   | 99.50   | 99.45   |
|                | MBCNN+CCI (\( A_m \) & \( A_r \)) | 99.38   | 99.56   | 99.47   |
| INCART         | MBCNN             | 99.35   | 99.13   | 99.24   |
|                | MBCNN+CCI (\( A_m \)) | 99.45   | 99.28   | 99.37   |
|                | MBCNN+CCI (\( A_r \)) | 99.48   | 99.23   | 99.35   |
|                | MBCNN+CCI (\( A_m \) & \( A_r \)) | 99.45   | 99.32   | 99.39   |
| QT             | MBCNN             | 99.93   | 99.90   | 99.92   |
|                | MBCNN+CCI (\( A_m \)) | 99.97   | 99.93   | 99.95   |
|                | MBCNN+CCI (\( A_r \)) | 99.98   | 99.92   | 99.95   |
|                | MBCNN+CCI (\( A_m \) & \( A_r \)) | 99.95   | 99.95   | 99.95   |

The results are average of five sub models under 5-fold cross validation and the better results are bold-faced.

#### Table III: The \( F_1 \) Results of Common Networks and the SOTA Solution Trained as Encoders With and Without CCI for QRS Location

| Database | CPSC-Te. | MITDB | INCART | QT |
|----------|----------|-------|--------|----|
| DenseNet | 98.86    | 99.22 | 99.10  | 99.80 |
| DenseNet+CCI | 99.11 | 99.36 | 99.28 | 99.93 |
| SENet [54] | 99.09 | 99.47 | 99.36 | 99.88 |
| SENet+CCI | 99.31 | 99.51 | 99.44 | 99.93 |
| CRNN [11] | 99.31 | 99.45 | 99.39 | 99.95 |
| CRNN+CCI | 99.41 | 99.69 | 99.50 | 99.97 |

The better results are bold-faced.
(A_r or A_m) and both attributes. Here we evaluated the proposed assumption on the four independent and classic databases, CPSC2019-Test, MITDB, INCART and QT. According to the results of the ablation study, intervention on the morphology attribute (A_m) and the rhythm attribute (A_r) in the latent space is effective and superimposed, which confirms our assumption on SCM with abstract attributes. We see steady gains when training model with CCI. The backbone model has reached a bottleneck in performance on most databases, yet for long-term ECGs, improvement of 0.1% on F1 may be equivalent to reducing hundreds or thousands of FPss and FNss. This can immensely alleviate the workload of cardiologists and reduce the cumulative burden of errors in subsequent diagnostics. The improvement of performance brought by CCI is mainly reflected in complex ECGs.

We show four typical examples with severe pathological variation and noise contamination in Fig. 5. It is apparent to see that the model trained with CCI can significantly reduce errors when recognizing variant QRS-complex or QRSized noise. Meanwhile, CCI can also weaken the response to the repeated P wave pattern for ECGs with severe auriculo-ventricular block, in which the relative position of the P wave and QRS-complex is unfixed.

Table III summarizes the performance gain brought by training with CCI for different Encoders for QRS-complex location. From these results, it can be seen that CCI is effective for the common network architectures and CRNN [11], which is consistent with the tendency when using MBCNN as the Encoder.

b) Results of heart sound segmentation: We report the evaluation metrics of the model training with and without CCI on databases from PhysioNet/CinC Challenge 2016 and LEVEL-I, II, III from [10] in Table IV. Similar observations to QRS location can be obtained in heart sound segmentation. For heart sound segmentation, the improvement of F1 induced by CCI is more significant. The model training with CCI outperforms the backbone method by at least 1.0% on most databases, even 2.37% and 2.25% on the databases constructed with pathological heart sounds, LEVEL-II and LEVEL-III.
Moreover, CCI causes a consistent $F_1$ performance promotion when segmenting different states. Also we can conclude that on the whole sub databases, the performance is further improved when we implement CCI with both attributes. In addition, the improvement in performance for the model trained with CCI is more evident when facing complex heart sounds. Such as for LEVEL-III, intervention on both attributes improves the $F_1$ by 1.72% compared to intervention on $A_m$ solely.

Similar as QRS-complex location, we also evaluated the efficiency of CCI while training different models. As shown in Table V, training with CCI can promote all the models including the SOTA solution. With CCI as the training paradigm, the enhancement in performance is especially clear when the baseline model is potent. For example, training TFAN with CCI brought almost 3% performance improvements on LEVEL-III. We also noticed that the performance gain is not sufficient when the Encoder is not ideally fitted.

2) Q2: Test With SNR Controllered Samples: In this section, we mainly analyze the changes brought by training with CCI when segmenting noisy cardiovascular signals. We conducted a noise stress test using the aforementioned typical noise records, including the bw, ma and em noises from [48] and the artificial power line interference noises for QRS-complex location, the lung sounds from [49] and the artificial gaussian noises for heart sound segmentation. We tested all the sub-models from 5-fold evaluation under different noise types and signal-to-noise ratios (SNR), and calculated the mean and standard deviation of the corresponding error rates. The noisy samples were generated from CPSC2019-Test and Training-a from PhysioNet/CinC Challenge 2016 by adding different types of noise according to the controlled SNRs (0, 5, 10, 15, 20 dB).

a) Main Results: Figs. 6 and 7 show the $Er$s at different SNRs for the segmentation model trained with and without CCI. It is evident that the model trained with CCI has the highest performance among all noise levels. CCI also results in the slowest performance decay compared to the backbone method. As shown in Fig. 6, the model trained with CCI can keep the $Er$ of QRS-complex location below 5% when SNR is over 5 dB, whereas the $Er$ is at least 1% higher for the model trained without CCI. Among all the noise categories, muscle artifact (ma) and power line interference noises have the greatest influence on QRS-complex location performance. It is worth noting that when the SNR is 0 dB, the performance of QRS-complex location models is drastically diminished. A similar phenomenon also occurs during noise stress tests on heart sound segmentation. This indicates that the original signals have been severely contaminated by noises when SNR is 0 dB, and are beyond the segmentation capacity of the models. However, the model trained with CCI still maintains higher performance and is more stable than one trained without CCI.

C. Q3: Visualization of Feature Density

According to the previous assumption, CCI should eliminate the implicit statistical bias brought by the single attribute and lead to more objective representations. For the conventional training, $A_m$ and $A_r$ may confounding the Encoder in distilling the intrinsic factors for state discrimination. Therefore, the feature representation learned with CCI ought to be more concentrated. To empirically verify this, we compress the deep features encoded by MBCNN to the unit hypersphere to visualize the latent distribution of different states. The representations of each state are grouped by frames with interval of 16 ms for QRS location and 100 ms for heart sound segmentation. A Gaussian kernel with bandwidth estimated by Scott’s Rule [55]...
is applied to estimate the probability density function of the generated representations after dimensionality reduction with principle component analysis (PCA) and normalization. Darker areas have more concentrated features, and if the feature space (the 2-dim sphere) is covered by dark areas, it has more diversely placed features.

The visualization results are shown in Figs. 8 and 9. It can be observed that, with CCI, the deep features of different states form more tight and concentrated clusters. Intuitively, they are potentially more separable from each other. In contrast, features learned without CCI are distributed in clusters that have more overlapped parts. The evident discrepancy occurs in non-QRS representations for QRS location and S2 representations for heart sound segmentation. This demonstrate why CCI improves the segmentation performance to a certain extent.

V. DISCUSSION AND CONCLUSION

In this work, we introduce a contrastive causal intervention scheme (CCI) for learning semantic representations of cardiovascular signals. CCI is a frame-level constraint for the training process to eliminate the implicit confounding factors induced...
by $A_m$ and $A_f$. We show that training with CCI can effectively improve the segmentation performance and adapt to other independent databases and various networks. Furthermore, the proposed method is considerably efficient to train a segmentation model generalizing to noisy cardiovascular signals. According to the visualization results of the latent distribution encoded with and without CCI, it is sensible to attribute the performance gain to a more separable and state-concentrated deep feature space brought by CCI. Since the proposed CCI does not introduce additional parameters for inference, it is suitable for quasi-periodic segmentation methods based on deep learning.

The proposed CCI currently has several disadvantages as a training paradigm for cardiovascular signal segmentation models. First of all, training with CCI will cause increase of computational complexity of training process since the CCI scheme should do intervention on representations frame-by-frame. Meanwhile, extra modules need to be introduced for variational approximation of the latent distribution while training with CCI. To quantify and compare the computational complexity for training with CCI, we provide the train parameters and train time in Table VI. As shown in Table VI, training with CCI will take about 10 times longer than training without it. The hardware platform for the training process is an AMD Ryzen5-2600x processor with 32 GB RAM and an NVIDIA Geforce RTX 2080Ti graphic card. Except for the extra computational cost for training, another disadvantage of the CCI scheme is that the intervened representations are generated by altering the contributions on the input data. Therefore, the proposed CCI in this work is not friendly to some signal processing methods. In the future, we will focus on direct causal intervention on attributions in latent space without manipulating the input data.

Contrastive learning has been out-standingly successful for CV and NLP, especially in self-supervised tasks. However, the existing work of contrastive learning in CV and NLP seems inappropriate to be applied for cardiovascular signals. For example, in CV and NLP, masking a part of the data and aligning the representation of the masked area with the original representation is a common framework. Yet for event-based analysis of cardiovascular signals, the disappearance of heartbeats may correspond to cardiac arrest, not a noise masking. In this work, we propose a contrastive learning framework based on the causal attributes of cardiovascular signals, and summarize several suggestions for further exploring.

1) If it is effective to construct a causal graph of the inner attributes of the data, how can we explore the intrinsic causality of more complex task with cardiovascular signals?

Our assumption of causal intervention on rhythm and morphology attribution was based on a prior intuition, corresponding to the cognition when we identify each state in a cardiac cycle. However, except for semantic segmentation, the classification of cardiovascular signals requires the abstracted causal mechanism more detailed. One possible direction is to establish the preliminary research on the binary classification task, like diagnosis of atrial fibrillation. Other attribute such as Markov chain of state transition is also a critical causal dependency when we doing deep representation learning for cardiovascular signals.

2) Excluding the interference of confounders in a specific task, should a better concentrated representation be obtained? In [56], the researchers have presented a connection between contrastive loss and the alignment and uniformity properties. The analysis is set in image classification and unsupervised learning. In the visualization results of this work, we have found out that when utilizing CCI in training process, the generated representations of different states are more aligned, less uniformed. As the organizer of CPSC2019 and the trimmer of PhysioNet/CinC Challenge 2016, we understand that the annotations of the training data from CPSC2019 and Training-a can be 100% confident through contextual information. Therefore, it is reasonable to have more concentrated representations when we reduce the confounding impact. Nonetheless, the cardiovascular signals own low frequency band and are always highly uncertain while testing due to variation and noise contamination. The research of whether the introduction of representation uniformity can measure and distinguish the uncertain state is a worthy investment.

References

[1] A. Santini, E. Diez, and M. Llamedo, “Versatile detector of pseudo-periodic patterns,” in Proc. Workshop Inf. Process. Control, Bahia Blanca, Argentina, 2019, pp. 37–41.

[2] H. U. Voss, “Hypersampling of pseudo-periodic signals by analytic phase projection,” Comput. Biol. Med., vol. 98, pp. 159–167, 2018.

[3] P. Noman, S. H. Salleh, C. M. Ting, S. B. Samdin, H. Obmao, and H. Hussain, “A Markov-switching model approach to heart sound segmentation and classification,” IEEE J. Biomed. Health Inf., vol. 24, no. 3, pp. 705–716, Mar. 2020.

[4] J. P. Madeiro, P. C. Cortez, F. I. Oliveira, and R. S. Siqueira, “A new approach to QRS segmentation based on wavelet bases and adaptive threshold technique,” Med. Eng. Phys., vol. 29, no. 1, pp. 26–37, 2007.

[5] L. Maršánová, A. Němčová, R. Smíšek, M. Vítek, and L. Smital, “Advanced P wave detection in ECG signals during pathology: Evaluation in different arrhythmia contexts,” Sci. Rep., vol. 9, no. 1, pp. 1–11, 2019.

[6] D. B. Springer, L. Tarassenko, and G. D. Clifford, “Logistic regression-HSMM-based heart sound segmentation,” IEEE Trans. Biomed. Eng., vol. 63, no. 4, pp. 822–832, Apr. 2016.

[7] C. Liu, D. Springer, and G. D. Clifford, “Performance of an open-source heart sound segmentation algorithm on eight independent databases,” Physiol. Meas., vol. 38, no. 8, 2017, Art. no. 1730.

[8] C. Liu, D. Zheng, A. Murray, and C. Liu, “Modeling carotid and radial artery pulse pressure waveforms by curve fitting with Gaussian functions,” Biomed. Signal Process. Control, vol. 8, no. 5, pp. 449–454, 2013.

[9] C. Liu et al., “Modelling arterial pressure waveforms using Gaussian functions and two-stage particle swarm optimizer,” Biomed. Res. Int., vol. 2014, 2014, Art. no. 923260.

[10] X. Wang, C. Liu, Y. Li, X. Cheng, J. Li, and G. D. Clifford, “Temporal-framing adaptive network for heart sound segmentation without prior knowledge of state duration,” IEEE Trans. Biomed. Eng., vol. 68, no. 2, pp. 650–663, Feb. 2021.

[11] W. Cai and D. Hu, “QRS complex detection using novel deep learning neural networks,” IEEE Access, vol. 8, pp. 97082–97089, 2020.

[12] B. Scholkopf et al., “Toward causal representation learning,” Proc. IEEE, vol. 109, no. 5, pp. 612–634, May 2021.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[14] H. Noh, S. Hong, and B. Han, “Learning deconvolution network for semantic segmentation,” in Proc. IEEE Int. Conf. Comput. Vis., Santiago, Chile, 2015, pp. 1520–1528.

[15] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Lecture Notes Comput. Sci., Munich, Germany, 2015, pp. 234–241.
[37] D. B. F. Agakov, “The IM algorithm: A variational approach to information maximization,” in Proc. Adv. Neural Inf. Process. Syst., Vancouver, BC, Canada, 2004, pp. 201–208.

[38] M. I. Belghazi et al., “Mutual information neural estimation,” in Proc. Int. Conf. Mach. Learn., Stockholm, Sweden, 2018, pp. 531–540.

[39] X. Nguyen, M. J. Wainwright, and M. I. Jordan, “Evaluating divergences functional and the likelihood ratio by convex risk minimization,” IEEE Trans. Inf. Theory, vol. 56, no. 11, pp. 5847–5861, Nov. 2010.

[40] R. D. Hjelm et al., “Learning deep representations by mutual information estimation and maximization,” in Proc. Int. Conf. Learn. Representations, New Orleans, LA, USA, 2019.

[41] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, “CLUB: A contrastive log-ratio upper bound of mutual information,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 1779–1788.

[42] B. Poole, S. Ozair, A. V. D. Oord, A. Alemi, and G. Tucker, “On variational bounds of mutual information,” in Proc. Int. Conf. Mach. Learn., Long Beach, CA, USA, 2019, pp. 5171–5180.

[43] H. Gao et al., “An open-access ECG database for algorithm evaluation of QRS detection and heart rate estimation,” J. Med. Imag. Health Inf., vol. 9, no. 9, pp. 1853–1858, 2019.

[44] G. B. Moody and R. G. Mark, “The impact of the MIT-BIH arrhythmia database,” IEEE Eng. Med. Biol. Mag., vol. 20, no. 3, pp. 45–50, May/Jun. 2001.

[45] P. Laguna, R. G. Mark, A. Goldberg, and G. B. Moody, “Database for evaluation of algorithms for measurement of QT and other waveform intervals in the ECG,” in Proc. Comput. Cardiol., Lund, Sweden, 1997, pp. 673–676.

[46] A. Taddei et al., “The European ST-T database: Standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography,” Eur. Heart J., vol. 13, no. 9, pp. 1164–1172, 1992.

[47] C. Y. Liu et al., “An open access database for the evaluation of heart sound algorithms,” Physiol. Meas., vol. 37, no. 12, pp. 2181–2123, 2016.

[48] G. B. Moody, W. Muldrow, and R. G. Mark, “Noise stress test for arrhythmia detectors,” in Proc. Comput. Cardiolog., Salt Lake City, UT, USA, 1984, pp. 381–384.

[49] M. Fraiwan, L. F. Frank, B. Khassawneh, and A. Ibni, “A dataset of lung sounds recorded from the chest wall using an electronic stethoscope,” Data Brief, vol. 35, 2021, Art. no. 106913.

[50] S. A. Israel, J. M. Irvine, A. Cheng, M. D. Wiederhold, and B. K. Wiederhold, “ECG to identify individuals,” in Proc. Int. Conf. Learn. Representations, New Orleans, LA, USA, 2020, pp. 6348–6359.

[51] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and M. J. Wainwright, “Densely connected convolutional networks,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, 2016, pp. 2043–2051.

[52] C. Mao, A. Cha, A. Gupta, H. Wang, J. Yang, and C. Vondrick, “Generative interventions for causal learning,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Virtual, Online, USA, 2021, pp. 3947–3956.

[53] T. Wang, J. Huang, H. Zhang, and Q. Sun, “Visual commonsense representation learning via causal inference,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops, Virtual, Online, USA, 2020, pp. 378–379.

[54] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, “Unbiased scene graph generation from biased training,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., Virtual, Online, USA, 2020, pp. 3716–3725.

[55] D. W. Scott. Multivariate Density Estimation: Theory, Practice, and Visualization. Hoboken, NJ, USA: Wiley, 2015.

[56] T. Wang and P. Isola, “Understanding contrastive representation learning through alignment and uniformity on the hypersphere,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 9929–9939.