Towards Ultrasound-based Reliable Disease Diagnosis Using Causal Inference

Bolei Chen[1], Jiaxu Kang[1], Haonan Yang[1], Ping Zhong[1*], Yixiong Liang[1], Rui Fan[2], Jianxin Wang[1*]

[1]School of Computer Science and Engineering, Central South University
[2]College of Electronics & Information Engineering, Tongji University
{boleichen, jxkang, haonanyang, ping.zhong, yxliang}@csu.edu.cn, rfan@tongji.edu.cn, jxwang@mail.csu.edu.cn

Abstract

Aligning the decision-making process of deep learning models with that of experienced sonographers is essential for ultrasound-based reliable disease diagnosis. Although existing methods have made significant progress in this aspect, their alignments are primarily associational rather than causal, leading to pseudo-correlations between features and diagnostic results. Such a biased diagnosis blindly models the sonographer's diagnostic skills and attention to specific patterns, which we argue hardly produces an AI diagnoser that is comparable to human experts. To address this issue, we propose a causality-based diagnostic framework to align the model's diagnostic behaviors with those of experts. Specifically, by delving into both conspicuous and inconspicuous confounders within the ultrasound images, the back-door and front-door adjustment causal learning modules are proposed to promote unbiased learning by mitigating potential pseudo-correlations. In addition, we integrate causal inference into a well-designed dual-branch model with feature interaction bridges for compatibility with multimodal ultrasound inputs. To fully evaluate our method, we conduct comparative studies on different diseases and ultrasound modalities. In particular, we publish a carefully constructed multimodal ultrasound dataset for breast lesion diagnosis and segmentation. Sufficient comparative and ablation studies on this dataset emphasize that our method outperforms state-of-the-art methods.

Code — https://github.com/BoLeiChen/Ceusformer

Introduction

Aligning the decision-making process of deep learning models with that of human experts is essential for developing reliable medical diagnosis systems (Zhuang and Hadfield-Menell 2020). For instance, in ultrasound-based breast cancer diagnosis, using models that are misaligned with clinical protocols can lead to shortcuts and spurious correlations between features and diagnostic results, resulting in misdiagnosis and loss of timely treatment. Existing methods (Chen et al. 2024; Gong et al. 2023; Guo et al. 2024; Wan et al. 2023) usually purposefully extract ultrasound features to mimic human experts' diagnostic skills and attention to specific patterns. Despite the promising diagnostic performance, their alignments with expert behaviors is only associational, rather than causal, making their models still biased towards spurious correlated features. As shown in Fig. 1 (a), in **C**ontrast-**E**nhanced **U**ltra**S**ound (CEUS)-based breast cancer diagnosis, two decision chains present similar correlation patterns but have different causal structures. According to clinical experience, microvascular patterns are considered as important attributes to distinguish benign and malignant lesions as they reflect the invasion of the lesion into the surrounding tissues. The causally aligned decision chain can notice such inconspicuous key features, however the unaligned one (DasT (Chen et al. 2024)) focuses on irrelevant features due to the effects of confounding bias.

Sonographers can deal with diverse lesions because they can learn the intrinsic causality of events beyond biased observations and acquire good analogical associations. In this paper, we propose a causality-based diagnostic framework to align the underlying causal logic of the model's decision-making process with that of human experts. Specifically, we first construct a structured causal model (Pearl 2009) by categorizing the bias variables affecting causality into conspicuous and inconspicuous confounders based on clinical experience. As shown in Fig. 1 (b), conspicuous confounders are content-related and easily identifiable. In contrast, inconspicuous confounders imply complex stylistic nuances that are difficult to discern but affect diagnostic decision-making. We then propose **B**ack-door and **F**ront-door **A**djustment **C**ausal **L**earning modules, i.e., BACL and FACL modules, to promote unbiased diagnosis by dealing with these confounders and mitigating pseudo-correlations, as shown in Fig. 1 (c).

Considering sonographers usually contrastively utilize multimodal ultrasound images in clinical practice (Folkman 2002), such as **B**-mode **U**ltra**S**ound (BUS) and CEUS, we design a dual-branch diagnostic model integrating causal inference to compatible with multimodal inputs. In particular, we design **T**emporal **A**ttention (TA) and **C**ontextual **A**ttention (CA) modules to fully extract inconspicuous but valuable features so that they have a fair chance to participate in causality-related predictions. Unlike previous causality-based methods (Yang, Zhang, and Cai 2021; Liu, Li, and Lin 2023) that restrict the causal intervention (Pearl and Mackenzie 2018) to the model's output layer and ignore possible biased features in feature mining and interaction,

*Corresponding authors.

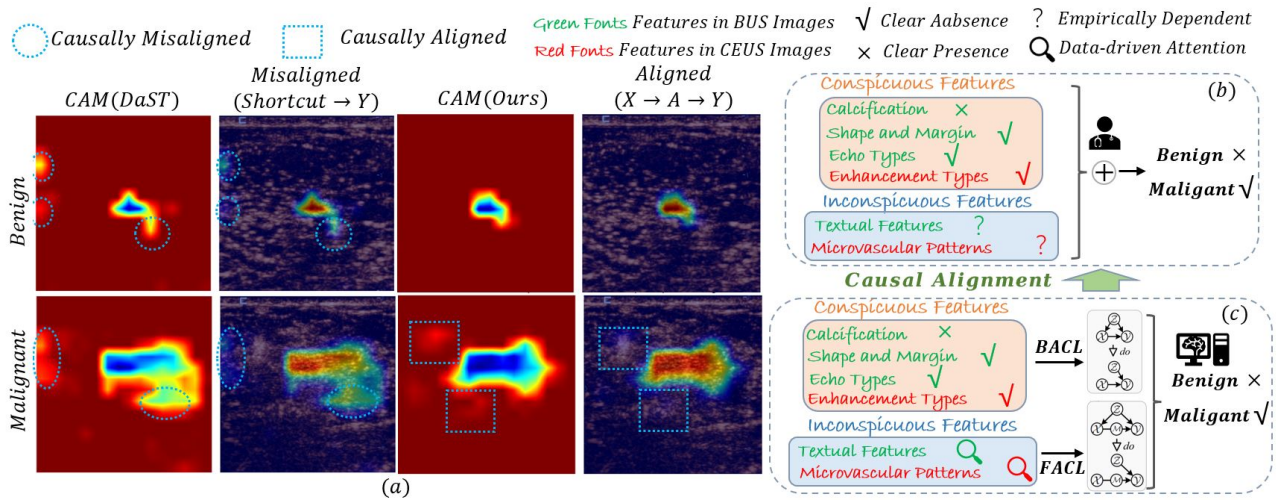

Figure 1: (a) Two decision chains present similar correlation patterns in **C**lass **A**ctivation **M**apping (CAM) visualizations but with different causal structures. The right chain"$Input(X) \rightarrow Attributes(A) \rightarrow Label(Y)$" aligns with sonographers, while the left chain "$Shortcut \rightarrow Label(Y)$" is misaligned due to the confounding bias between the $Shortcut$ and Y. (b) Experienced sonographers make diagnoses by comprehensively considering both conspicuous and inconspicuous features of the ultrasound images based on causal logics. (c) Our causally aligned model can notice inconspicuous but valuable features to eliminate pseudo-correlations and facilitate unbiased diagnosis.

we release the target effect of the causal hypothesis to the feature fusion. Thus, our causal inference modules are integrated into the feature interaction bridges of the dual-branch model to mitigate feature bias caused by confounders.

To fully evaluate our method, we conduct comparative studies on different diseases and ultrasound modalities. In particular, we publish a carefully constructed multimodal ultrasound (BUS-CEUS) dataset for breast lesion diagnosis, which is 1.36 $\sim$ 2.98 times larger than existing non-public BUS-CEUS datasets (Xie et al. 2023; Chen et al. 2021, 2022, 2023, 2024; Xu et al. 2022; Guo et al. 2024). We reveal the advantages of integrating causal inference to deconfound biases through comprehensive comparative and ablation studies. Overall, our contributions are as follows: **(1)** A causality-based diagnostic framework with BACL and FACL modules is proposed to deal with pseudo-correlations caused by confounders and align the model's decision-making process with that of human experts. **(2)** We design a causally aligned dual-branch diagnostic model with feature interaction bridges, which gives inconspicuous features a fair chance to engage in causality-related prediction. **(3)** We conduct comparative studies on different diseases and ultrasound modalities to evaluate our method. Our code and multimodal dataset will be publicly available.

Related Work

Multimodal Ultrasound-based Disease Diagnosis

While promising advances (Wang et al. 2024; Mo et al. 2023) have been made in unimodal ultrasound-based disease diagnosis, researchers are increasingly realizing the advantages of using multimodal ultrasound data to facilitate the diagnosis of tumors. To address the scarcity of multimodal data and improve diagnostic performance, existing methods try to integrate sonographers' domain knowledge into deep learning models through feature mining. For example, some methods (Chen et al. 2021, 2024; Gong et al. 2023) mimic the sonographers's focus on malignant tumor expansion in the spatial dimension and on brightness changes in CEUS along the temporal dimension. In clinical practice, sonographers usually make a diagnosis decision by comprehensively contrasting the morphologic features in BUS with the distribution of microvessels in CEUS. Therefore, most of the recent approaches (Chi et al. 2024; Chen et al. 2024; Gong et al. 2023; Guo et al. 2024) extract features from both modalities separately and fuses them interactively. Several methods (Chi et al. 2024; Chen et al. 2023, 2024) further adopt a multi-task learning strategy based on dual-branch modals to achieve joint tumor diagnosis and segmentation. The regional characteristics provided by the segmentation task and the spatiotemporal patterns represented by the classification task are used as both constraints and references to each other in the learning process (Chi et al. 2024).

Despite the promising progress, existing methods blindly mimic the expert diagnostic experience, which we argue hardly produces an AI diagnoser that is comparable to human experts. Their alignments with expert behaviors is only associational, rather than causal, making their models still biased towards spurious correlated features. In addition, existing methods (Chen et al. 2022; Gong et al. 2023; Guo et al. 2024) fall short in extracting key features and fusing multimodal features. In particular, the neglect of inconspicuous key features leads to their unfair participation in diagnostic decisions, which further causes biased learning. To alleviate these issues, we propose a causality-based diagnostic framework to align the model's decision-making process with that of experts. Such alignment in medical imaging systems is largely understudied. In addition, we design a causally aligned dual-branch diagnostic model with fea-

ture interaction bridges, which gives inconspicuous features a fair chance to engage in causality-related prediction.

Causal Inference

Causal inference is an emerging technique for exploring task causality (Pearl and Mackenzie 2018), with a surge in efforts to combine it with deep learning in tasks such as image recognition (Wang et al. 2021, 2022; Zhang et al. 2022). One popular approach is using adjustment techniques to mitigate the negative effects caused by confounders, and several other studies have explored the use of counterfactuals (Abbasnejad et al. 2020; Niu et al. 2021). Considering its practicality, this paper is the first to use the adjustment method to solve the problem of multimodal ultrasound-based disease diagnosis. In addition, current methods employ back-door (Liu et al. 2022; Zhang et al. 2020) or front-door (Liu, Li, and Lin 2023; Yang, Zhang, and Cai 2021; Yang et al. 2021; Zhang et al. 2024) adjustments separately and lack comprehensive confounder assumptions and complete bias corrections. In this paper, we propose to address both conspicuous and inconspicuous confounders in BUS and CEUS modalities. Our method can effectively mitigate the negative impact of pseudo-correlations caused by confounders on disease diagnosis. Our causality-based disease diagnosis performs well across different diseases and data modalities.

Preliminary

Structural Causal Model and Confounders

For compatibility with multimodal inputs, a structural causal model is constructed to capture the relationships among the key variables in multimodal ultrasound-based disease diagnosis, as shown in Fig. 2. In this directed acyclic graph, the starting and ending points indicate the cause and effect, respectively. Traditional methods focus on learning the observational association $P(Y|X)$, overlooking the ambiguity and pseudo-correlations introduced by confounders Z in the back-door path $X \leftarrow Z \rightarrow Y$. Here, confounders are biased variables that affect causes and effects, e.g., content or specific attributes that are given undue attention. $Z \rightarrow X$ arises because the combined probability of samples is inevitably affected by the limited resources available in the real world when collecting data. In addition, $Z \rightarrow Y$ exists since the data source (patient population) and data labeling also affect the probability of the diagnostic distributions. These confounding connections may cause spurious shortcuts (as shown in Fig. 1 (a)) during training but can be detrimental in new situations.

To reduce the effect of confounders, our model categorizes them into conspicuous and inconspicuous categories to align with the diagnostic process of human experts. Concretely, conspicuous confounders are content-related and easily identifiable, e.g., the lesion shapes, margins, and calcifications in BUS Z^c_{BUS} and the contrastive enhancement types in CEUS Z^c_{CEUS}. In contrast, inconspicuous confounders imply complex stylistic nuances that are difficult to discern but affect diagnostic decision-making, e.g., the rich textural features in BUS Z^i_{BUS} and the microvascular patterns in CEUS Z^i_{CEUS}, as shown in Fig. 1. Since we cannot explicitly model inconspicuous confounders Z^i_*, the additional mediators M are inserted between X and Y to establish front-door paths $X \rightarrow M \rightarrow Y$. Therefore, the diagnostic reasoning can be divided into a feature selector $X \rightarrow M$ and a diagnostic predictor $M \rightarrow Y$ which are utilized to select the appropriate attributes M from X and to predict Y using M, respectively.

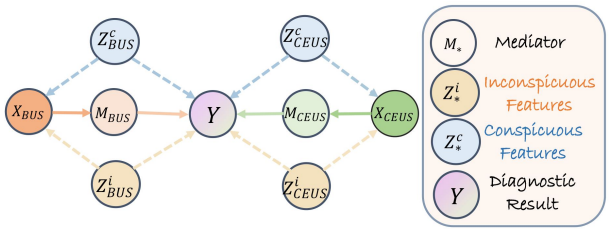

Figure 2: An illustration of the structural causal model for multimodal ultrasound-based disease diagnosis.

Methodology

Back-door Adjustment Causal Learning (BACL)

According to Bayes' theorem, the typical observation likelihood is $P(Y|X) = \sum_z P(Y|X,z)\underline{P(z|X)}$, where $P(z|X)$ may introduce biased weights during disease diagnosis. In our work, *do*-operator (Pearl and Mackenzie 2018) is employed to break the backdoor link between Z and X, providing a scientifically sound method for determining causal effects. Based on the invariance and independence rules (Pearl 2016), we have:

$$\begin{aligned} P(Y|do(X)) &= \sum_z P(Y|do(X),z)P(z|do(X)) \\ &= \sum_z P(Y|X,z)\underline{P(z)} \end{aligned} \tag{1}$$

In this case, the causal intervention is achieved by blocking the backdoor path $Z \rightarrow X$ so that X has a fair chance to incorporate causality-related factors for prediction. Eq. (1) is implemented as:

$$\mathcal{B}(x,z) = \mathbb{E}_z[f(x,z)], \tag{2}$$

where $f(x,z) = f_x(x) + f_z(z)$ is the specific network module, which is a linear approximation in our work. Notably, if $f(x,z)$ is considered nonlinear, it needs to be considered in the joint space and $\mathbb{E}_z[f(x,z)]$ needs to be calculated using integration or sampling methods. Such models may have stronger modeling capabilities, but are computationally complex and unfriendly to small sample learning. Therefore, Eq. (2) becomes $f_x(x) + \mathbb{E}_z[f_z(z)]$ in our case, as shown in Fig. 2 (a). We obtain $\mathbb{E}_z[f_z(z)]$ by utilizing attention-based methods:

$$\mathbb{E}_z[f_z(z)] = \sum_i \frac{exp(hz_i^\top)}{\sum_i exp(hz_j^\top)} f_z(z_i), \tag{3}$$

where h denotes hidden features of the ultrasound images. $\frac{exp(hz_i^\top)}{\sum_i exp(hz_j^\top)}$ indicate the correlation weight between the image features and confounder z_i.

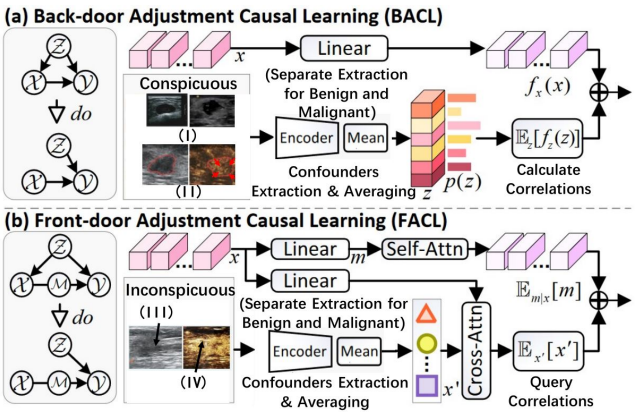

Figure 3: Details of how our BACL and FACL modules work. These modules are integrated into the feature interaction bridges, where x is the feature that transmits from one branch to the other. (I) Morphological Features and Calcification. (II) Centripetal Enhancement. (III) Texture Features. (IV) Microvascular Patterns.

In practice, we need to explicitly extract conspicuous confounders Z_*^c from ultrasound images for BACL. Based on clinical experience, morphologic features such as shape, margin, calcification, and echo types in the BUS are considered conspicuous confounders. The lesion expansion, internal filling defect, and contrastive enhancement types are regarded as inconspicuous confounders in CEUS. We give examples of extracting confounders using medical pre-training models in the supplementary material. The confounder dictionaries are denoted as $\hat{D}_*^b$ and $\hat{D}_*^m$ according to the benign and malignant nature, respectively. On this basis, for example, the BACL process in BUS is formulated as follows:

$$\hat{x}_B = LN(\psi(\mathcal{B}(x_B, Z_B))), Z_B = LN(\phi([\hat{D}_B^b, \hat{D}_B^m], \quad (4)$$

where ϕ and ψ are full-connection layers. LN is LayerNorm. x_B and $\hat{x}_B$ are initial and causally aligned BUS features, respectively. The BACL process in CEUS is similar.

Front-door Adjustment Causal Learning (FACL)

Although BACL can deal with the shortcuts and pseudo-correlations caused by conspicuous confounders, there are additional inconspicuous confounders that cannot be explicitly captured and pre-modeled. In this section, we introduce the FACL technique (Pearl 2016) to address this issue.

As shown in Fig. 2, an additional mediator M is inserted between inputs and outputs to construct the front door path $X \rightarrow M \rightarrow Y$. In the feature selector $X \rightarrow M$, an attention-based model $P(Y|X) = \sum_m P(Y|m)P(m|X)$ will select key features M from inputs X for the diagnostic predictor $M \rightarrow Y$. In addition, the *do*-operation is simultaneously applied to X and M to eliminate spurious correlations introduced by the inconspicuous confounder Z_*^i:

$$\begin{aligned} P(Y|do(X)) &= \sum_m P(Y|do(m))P(m|do(X)) \\ &= \sum_{x'} P(x') \sum_m P(Y|m, x')P(m|X) \\ &= \mathbb{E}_{x'}\mathbb{E}_{m|x}[P(Y|x', m)], \end{aligned} \quad (5)$$

where x' denotes potential input samples of the whole representation space, different from current inputs $X = x$. Please see the supplementary material for detailed derivations of the formulas. Similar to BACL, based on the linear mapping model, Eq. (5) becomes $\mathbb{E}_{m|x}[m] + \mathbb{E}_{x'}[x']$, as shown in Fig. 3 (b). Since the expectation is difficult to obtain a closed-form solution in a complex representation space, the estimation of the expectation is achieved through a query mechanism. In particular, we employ two embedding functions (Yang, Zhang, and Cai 2021) to transmit input x into two query sets $g_1 = q_1(x)$ and $g_2 = q_2(x)$, respectively. Then, the FACL is approximated as:

$$\begin{aligned} \mathbb{E}_{m|x}[m] &\approx \sum_m P(m|g_2)m = \sum_i \frac{exp(g_2 m_i^\top)}{\sum_j exp(g_2 m_j^\top)} m_i, \\ \mathbb{E}_{x'}[x'] &\approx \sum_{x'} P(x'|g_1)x' = \sum_i \frac{exp(g_1 x_i'^\top)}{\sum_j exp(g_1 x_j'^\top)} x_i', \\ F(x, x') &= \mathbb{E}_{x'}[x'] + \mathbb{E}_{m|x}[m]. \end{aligned} \quad (6)$$

The above process can be efficiently implemented using multi-head attention (Vaswani 2017), which is seamlessly integrated into a dual-branch diagnostic model introduced in the following section.

In practice, inconspicuous confounders are not as easy to distinguish as conspicuous ones. Therefore, we employ medical pre-training models to frame-by-frame extract texture features depicting the style of BUS images and video-by-video model the brightness changes representing the microvascular infiltration in CEUS, respectively. Please see the supplementary material for more details. Similar to BACL, the confounder dictionaries are denoted as $\tilde{D}_*^b$ and $\tilde{D}_*^m$ according to the benign and malignant nature, respectively. On this basis, the causally aligned features $\tilde{x}_B$, and $\tilde{x}_C$ are calculated as follows:

$$\tilde{x}_B = F(x_B, [\tilde{D}_B^b, \tilde{D}_B^m]), \tilde{x}_C = F(x_C, [\tilde{D}_C^b, \tilde{D}_C^m]), \quad (7)$$

where x_B and x_C denote the initial BUS and CEUS features, respectively.

A Dual-branch Model for Multimodal Ultrasound

As shown in Fig. 4, inspired by clinical experience, our dual-branch model extracts complementary features from multi-modal ultrasound images and exploits them contrastingly.

CNN Branch for BUS. This branch employs a feature pyramid structure with N convolutional blocks, in which the resolution of the feature maps decreases with the network depth as the number of channels increases. As shown in Fig. 4 (a), according to the definition in ResNet (He et al. 2016), the bottleneck contains a 1×1 down-projected convolution, a 3×3 spatial convolution, a 1×1 up-projected convolution, and a residual connection between the bottleneck inputs and outputs. In a CNN block, the convolutional kernels slide over overlapping feature maps, possibly preserving fine lesion local texture features.

Transformer Branch for CEUS. Following ViT (Alexey 2020; Touvron et al. 2021), this branch contains N repeated transformer blocks. As shown in Fig. 4 (a), each transformer block mainly consists of a TA module, a CA module, and

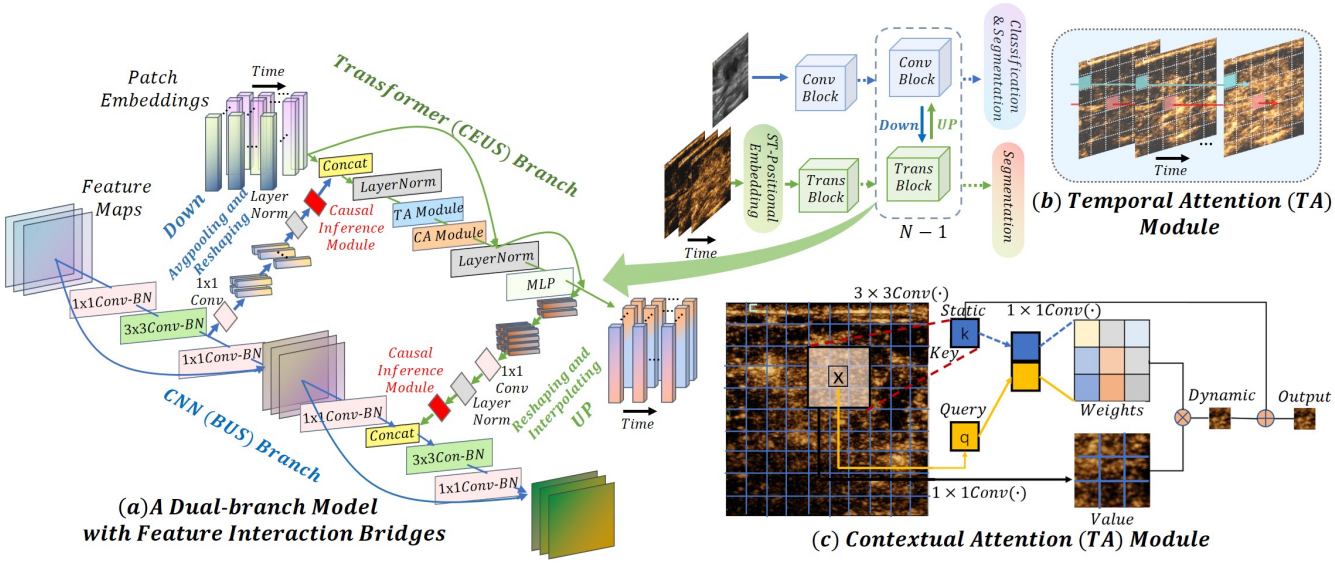

Figure 4: Architecture and details of the dual-branch diagnostic model. (a) The CEUS branch takes as input a clip of a CEUS video and the BUS branch takes as input a single BUS image. (b) and (c) illustrate the TA and CA modules in the Transformer (CEUS) branch, respectively.

an **M**ulti-**L**ayer **P**erceptron (MLP) block. The TA module is used to model microvascular patterns reflected by time-varying brightness features. In practice, it is computed as the self-attention of each patch along the temporal dimension, as shown in Fig. 4 (b). Please see the supplementary material for more implementation details. The CA module is designed to model the fine-grained invasion of the lesion into the surrounding tissues. As shown in Fig. 4 (c), the CA module first contextually encodes the input keys by 3×3 convolution to obtain a static contextual representation of the input. Furthermore, the encoded keys are concatenated with the input query to learn the multi-head attention matrix by two consecutive 1×1 convolutions. The learned attention matrix is multiplied by the input values to simulate feature interactions and realize the dynamic contextual representation of the input. Finally, the fusion of the static and dynamic contextual representations is used as the output. Notably, the local attention matrix for each spatial location is learned based on query features and contextual key features, rather than isolated query key pairs, which enhances self-attention learning by tapping into the additional guidance of the static context. In addition, the CA module unifies context mining among keys and self-attentive learning over 2D feature maps in a single architecture, thus avoiding additional branches for context mining.

Feature Interaction Bridge. Previous causality-based approaches (Yang, Zhang, and Cai 2021; Liu, Li, and Lin 2023) restrict the intervention to only the final Softmax layer of the network, ignoring possible biased features in the feature mining and interaction. Since trained neural networks implicitly incorporate conditional probabilities in pattern recognition (Nie, Zheng, and Ji 2018), we release the target effect of the causal hypothesis to the feature fusion rather than outputs. The learned unbiased feature fusion will cause unbiased predictions. Therefore, we integrate causal inference (BACL and FACL) modules into the bridging design to mitigate feature bias caused by confounders, as shown in Fig. 4 (a). In addition, our bridging design utilizes a 1×1 convolution to align the channel size. Average pooling, interpolating, and reshaping operations are used to align the feature resolution. LayerNorm (Lei Ba, Kiros, and Hinton 2016) and BatchNorm (Ioffe 2015) are employed to align the feature values. Causally aligned features $\hat{x}_*$ or $\tilde{x}_*$ from one branch are concatenated with the features of another branch to participate in the forward propagation of that branch.

Experiments

Experimental Setup

BUSI Dataset (Al-Dhabyani et al. 2020) is a publicly available breast BUS dataset, which is gathered from 600 females aged 25 to 75 at Baheya Hospital in Cairo, Egypt, with a total of 780 images (133 normal, 437 benign and 210 malignant). We perform comparative experiments using only 647 abnormal images.

SYSU-FLL-CEUS Dataset (Liang et al. 2015) is gathered from the First Affiliated Hospital, Sun Yat-sen University. The equipment used was Aplio SSA-770A (Toshiba Medical System), and all videos included in the dataset are collected from pre-operative scans. The dataset consists of CEUS data of focal liver lesions in three types: 186 HCC, 109 HEM and, 58 FNH instances (i.e. 186 malignant and 167 benign instances).

BUS-CEUS Multimodal Dataset. Since there is no publicly available multimodal dataset, we utilize our self-constructed BUS-CEUS dataset of breast lesions to perform comparative and ablation studies. Our dataset consists of 486 BUS-CEUS pairs collected from different patients at an internationally leading hospital (anonymous for review), including 237 benign and 249 malignant samples. The 486

Method	BUSI			
	Acc (%) ↑	Sens (%) ↑	Spec (%) ↑	F1 (%) ↑
HoVer-Trans (Mo et al. 2023)	$85.5_{\pm 2.4}$	$87.6_{\pm 2.3}$	$83.8_{\pm 2.2}$	$87.2_{\pm 2.1}$
Li et.al. (Li et al. 2025)	$90.2_{\pm 2.3}$	$\mathbf{90.0}_{\pm 1.9}$	$90.2_{\pm 2.6}$	$80.2_{\pm 1.9}$
Ours	$\mathbf{91.6}_{\pm 1.8}$	$\underline{89.8_{\pm 1.7}}$	$\mathbf{93.8}_{\pm 2.3}$	$\mathbf{88.3}_{\pm 1.6}$

Table 1: Comparitive studies for BUS-based breast cancer diagnosis. Bold and underline indicate optimal and suboptimal results, respectively.

Method	SYSU-FLL-CEUS			
	Acc (%) ↑	Sens (%) ↑	Spec (%) ↑	F1 (%) ↑
Liang et.al. (Liang et al. 2015)	$\underline{92.7_{\pm 1.5}}$	$89.5_{\pm 2.4}$	$91.4_{\pm 1.6}$	$88.1_{\pm 2.1}$
Trans-CEUS (Chen et al. 2023)	$91.7_{\pm 2.5}$	$\underline{90.3_{\pm 2.2}}$	$\underline{92.6_{\pm 2.4}}$	$\underline{89.5_{\pm 1.7}}$
Ours	$\mathbf{94.2}_{\pm 1.6}$	$\mathbf{91.6}_{\pm 2.0}$	$\mathbf{95.8}_{\pm 1.3}$	$\mathbf{91.6}_{\pm 2.4}$

Table 2: Comparative studies for focal liver lesion diagnosis.

cases are collected by different sonographers and each case is labeled with the corresponding biopsy result. Notably, the scale of our breast BUS-CEUS dataset is $1.36 \sim 2.98$ times larger than that of existing work (Xie et al. 2023; Chen et al. 2021, 2022, 2023, 2024; Xu et al. 2022; Guo et al. 2024). Both modalities are acquired simultaneously using a Mindray Resona R9 color Doppler US diagnostic instrument and a linear array probe 9L at a frequency of 3-9 MHz. The image sizes of the BUS image and frames in the CEUS videos are 550×582 pixels. All the BUS-CEUS samples in the dataset are labeled with a dual-mode mask by several experienced sonographers. That is, there is a corresponding segmentation mask for each BUS image or complete CEUS video. Please note that the participating hospital's ethics committees have approved the study protocol, and the data will be made public after anonymous review.

Implementation Details. Our method is implemented using PyTorch and is trained on an NVIDIA GeForce RTX 3090 GPU. The BUS branch accepts a single BUS image as input, while the CEUS branch incorporates $k = 7$ frames extracted from each CEUS video. Unless specifically declared, the input BUS and CEUS images are resized to a resolution of 384×384 pixels. In the parametric studies, we will experimentally evaluate the effect of different k and resolutions on the diagnostic performance. Data augmentation techniques, such as horizontal flip and rotation are also applied. During training, the batch size is set to 4, the epoch is set to 50, the learning rate is set to 10^{-4}, and the Adam optimizer (Kingma 2014) is used to update network parameters. The CEUS branch has a feature embedding dimension of 384 and the number of attention heads is 8.

For disease diagnosis, a linear layer is used to project the BUS branch's output into benign or malignant categories. The binary cross-entropy loss is used as the categorization loss. In addtion, a 4-layer decoder consisting of CNNs and upsampling operations is used for lesion segmentation in BUS images. The reshaping operation and 1D convolution are adopted to reshape the shape and dimension of the CEUS branch's output to predict the lesion segmentation of CEUS. The cross-entropy, Dice, and IoU losses are used as the segmentation losses with weights of 1:1:1. In the supplementary material, we discuss the mutual promotion between segmentation and diagnostic tasks.

Method	BUS-CEUS Multimodal Dataset			
	Acc (%) ↑	Sens (%) ↑	Spec (%) ↑	F1 (%) ↑
TSDBN (Yang et al. 2020)	82.22	80.72	68.42	79.77
DKG (Chen et al. 2021)	85.31	79.02	72.27	85.03
Huang et al. (Huang et al. 2023)	82.41	76.76	63.55	81.77
AHAF (Gong et al. 2023)	$84.53_{\pm 2.72}$	$\underline{83.37_{\pm 1.59}}$	$83.90_{\pm 2.07}$	$79.47_{\pm 1.93}$
KAMnet (Guo et al. 2024)	$85.42_{\pm 2.82}$	$74.19_{\pm 2.67}$	$89.04_{\pm 2.83}$	$83.97_{\pm 2.03}$
DaST (Chen et al. 2024)	$86.28_{\pm 2.61}$	$82.79_{\pm 1.97}$	$79.81_{\pm 2.73}$	$82.53_{\pm 1.82}$
UAC-T (Chi et al. 2024)	$\underline{86.33_{\pm 2.81}}$	$81.62_{\pm 2.49}$	$\underline{92.36_{\pm 3.43}}$	$\underline{87.28_{\pm 2.79}}$
Ours	$\mathbf{89.80}_{\pm 2.58}$	$\mathbf{85.08}_{\pm 1.80}$	$\mathbf{95.83}_{\pm 3.04}$	$\mathbf{89.36}_{\pm 2.00}$

Table 3: Comparative studies on multimodal ultrasound-based diagnosis of breast cancer.

Evaluation Metrics. The commonly used classification accuracy (Acc), sensitivity (Sens), specificity (Spec), and F1-score are selected as evaluation metrics for disease diagnosis. In addition, mean intersection over union (mIoU) and mean Dice coefficients (mDice) are employed to evaluate lesion segmentation in BUS and CEUS modalities. All the datasets are randomly divided into training and validation sets, with a split ratio of 4:1. We evaluate our method using 5-fold cross-validation and report the results of statistical significance using paired student's t-tests ($p<0.05$).

Comparisons on Unimodal Dataset

We first compare the proposed method with several strong baselines on two publicly available unimodal ultrasound datasets. Notably, in this case, the inputs to both branches of the model are either BUS images or CEUS images. The experimental results of BUS-based breast cancer diagnosis and CEUS-based focal liver lesion diagnosis are shown in Tab. 1 and Tab. 2, respectively. Li et.al. and HoVer-Trans respectively incorporate clinical experience and anatomical structure knowledge into the diagnostic model to mimic the thinking of a human expert diagnosing breast lesions. Liang et.al. and Trans-CEUS respectively incorporate contrastive enhancement types and dynamic microvascular perfusion patterns into the diagnostic model to establish associations between CEUS characteristics and diagnostic results. Unlike them, our method aligns the model's decision-making process with that of human experts by modeling the causal logic behind disease diagnosis, thus achieving better diagnostic performance. In addition, unlike the vanilla attention mechanisms used by HoVer-Trans and Trans-CEUS, our CA module can reveal inconspicuous but valuable features that facilitate reliable disease diagnosis.

Comparisons on Multimodal Dataset

We compare the proposed method with several baselines and state-of-the-art methods on the BUS-CEUS multimodal dataset and the experimental results are shown in Tab. 3. Notably, TSDBN and Huang et al. proposed different dual-branch diagnostic models similar to our method. DKG leverages keyframes' ROI and the brightness change curve of CEUS as domain knowledge to enhance breast cancer diagnosis, achieving respectable Acc and F1 metrics. Statis-

Model	Para(M) ↓	FLOPs(G) ↓
Trans-CEUS (Chen et al. 2023)	88.63	15.49
UAC-T (Chi et al. 2024)	40.55	67.34
Our method	63.41	48.28

Table 4: Comparisons of number of parameters and computational complexity.

Ablations						Metrics			
TA	CA	B-Down	B-Up	BACL	FACL	**Acc** (%) ↑	**Sens** (%) ↑	**Spec** (%) ↑	**F1** (%) ↑
✓	-	-	-	-	-	79.25	74.61	84.17	78.83
✓	✓	-	-	-	-	83.67	78.34	88.83	82.60
✓	✓	✓	-	-	-	84.39	79.67	89.01	83.96
✓	✓	✓	✓	-	-	85.03	80.90	90.83	84.79
✓	✓	✓	✓	-	✓	88.37	83.83	94.16	88.40
✓	✓	✓	✓	✓	-	89.07	84.28	94.77	88.96
✓	✓	✓	✓	✓	✓	**89.80**	**85.08**	**95.83**	**89.36**

Table 5: Ablation studies of each component of our method.

tically, our method absolutely improves $4.49\% \sim 7.58\%$, $3.36\% \sim 4.36$, $23.56\% \sim 32.28\%$, and $4.33\% \sim 9.59\%$ on four metrics relative to these baseline methods.

AHAF proposes to combine non-imaging clinical data with multimodal ultrasound data to boost the breast cancer diagnostic performance. In addition, AHAF also utilizes the clearest BUS frames that are captured separately. For a fair comparison, we compare our method with AHAF using only multimodal ultrasound data on our dataset without using additional data. KAMnet, DaST, and UAC-T all utilize a dual-branch network design. The difference is that KAMnet focuses on knowledge mining, representation, and integration. DaST and UAC-T mimic expert diagnostic experience by emphasizing feature interactions between modalities. Despite the competitive diagnostic performance, their alignment with expert behaviors is only associational, rather than causal, making their models still biased towards spurious correlated features. Thanks to the causality-based diagnostic framework and dual-branch model design, our method significantly improves diagnostic performance relative to existing methods. Statistically, Our method absolutely improves $3.47\% \sim 5.27\%$, $1.71\% \sim 10.89$, $3.47\% \sim 16.02\%$, and $2.08\% \sim 9.89\%$ on four metrics relative to these state-of-the-art methods. Further, Tab. 4 reports the number of parameters and computational complexity of our method compared to state-of-the-art unimodal and multimodal methods.

In the supplementary material, we further compare the proposed method with existing lesion segmentation methods on the multimodal BUS-CEUS dataset.

Ablation Studies

Based on the BUS-CEUS multimodal dataset, Tab. 5 reports the results of the ablation studies for each component of our method. The method that retains the TA module is used as the baseline to capture CEUS's spatiotemporal features. By replacing the vanilla spatial attention module (Vaswani 2017) with a CA module, all four metrics are significantly improved. The integration of both B-Down and B-Up improves the performance of breast cancer diagnosis, reflecting the effectiveness of the bridging designs for multimodal feature interactions. However, we find that feature interaction bridges without causal inference modules contribute little to diagnostic performance. In the feature extraction and interaction phases, the attendance of BACL and FACL can align the model's decision-making process with that of sonographers, thus substantially improving diagnostic performance. Benefiting from explicitly extracted conspicuous confounders, the contribution of BACL is larger than that of FACL. The complete model achieves the best breast cancer diagnostic performance.

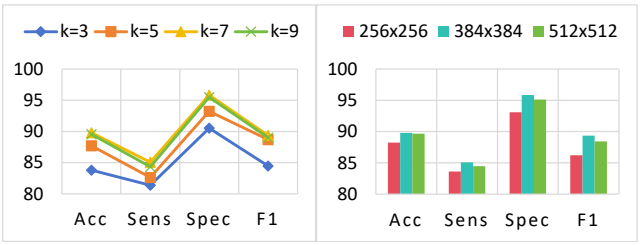

Figure 5: The effects of different CEUS frames k and different image sizes on multimodal ultrasound-base breast cancer diagnosis.

Parametric Studies

We conduct parametric studies on the BUS-CEUS multimodal ultrasound dataset. Fig. 5 illustrates the effects of different CEUS frames k and different image sizes on disease diagnostic performance. Using appropriate k or image size facilitates the trade-off between diagnostic performance and computational overhead. By adjusting the hyper-parameters, we find that $k = 7$ and image size 384×384 are favorable to achieve the best diagnostic performance. Increasing the value of k or the image size consumes more computational resources and degrades the diagnostic performance.

Conclusion

In this paper, we propose a causality-based diagnostic framework to align the diagnostic model's decision-making process with that of human experts. Specifically, we employ BACL and FACL techniques to address spurious correlations caused by conspicuous and inconspicuous confounders, respectively, for unbiased causal learning. For compatibility with multimodal ultrasound inputs, we design a causally aligned dual-branch model with feature interaction bridges and a CA module. The bridging design integrated with causal inference modules helps to achieve causally aligned multimodal feature fusion. The CA module facilities mining inconspicuous key features so that they can participate fairly in causality-based diagnostic predictions. To fully evaluate our method, we publish a carefully constructed multimodal ultrasound (BUS-CEUS) dataset for breast lesion diagnosis, whhich is $1.36 \sim 2.98$ times larger than existing non-public BUS-CEUS dataset. Sufficient comparative studies on different diseases and ultrasound modalities demonstrate the superiority of our method.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under 62272489, 62332020, and 62350004, in part by the Natural Resources Science and Technology Plan Project of Hunan Province under 2021-17, and in part by the Open Competition Project of Xiangjiang Laboratory under 23XJ01011. This work was carried out in part using computing resources at the High-Performance Computing Center of Central South University.

References

Abbasnejad, E.; Teney, D.; Parvaneh, A.; Shi, J.; and Hengel, A. v. d. 2020. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10044–10054.

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in brief*, 28: 104863.

Alexey, D. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv: 2010.11929*.

Chen, C.; Wang, Y.; Niu, J.; Liu, X.; Li, Q.; and Gong, X. 2021. Domain knowledge powered deep learning for breast cancer diagnosis based on contrast-enhanced ultrasound videos. *IEEE Transactions on Medical Imaging*, 40(9): 2439–2451.

Chen, F.; Han, H.; Ning, G.; Wen, B.; Liao, H.; Kong, W.; and Zhang, D. 2022. Immunohistochemical index prediction of breast cancer by using hybrid ultrasound data. *IEEE Transactions on Biomedical Engineering*, 70(4): 1401–1412.

Chen, F.; Han, H.; Wan, P.; Chen, L.; Kong, W.; Liao, H.; Wen, B.; Liu, C.; and Zhang, D. 2024. Do as Sonographers Think: Contrast-enhanced Ultrasound for Thyroid Nodules Diagnosis via Microvascular Infiltrative Awareness. *IEEE Transactions on Medical Imaging*.

Chen, F.; Han, H.; Wan, P.; Liao, H.; Liu, C.; and Zhang, D. 2023. Joint segmentation and differential diagnosis of thyroid nodule in contrast-enhanced ultrasound images. *IEEE Transactions on Biomedical Engineering*, 70(9): 2722–2732.

Chi, J.; Chen, J.-h.; Wu, B.; Zhao, J.; Wang, K.; Yu, X.; Zhang, W.; and Huang, Y. 2024. A Dual-Branch Cross-Modality-Attention Network for Thyroid Nodule Diagnosis Based on Ultrasound Images and Contrast-Enhanced Ultrasound Videos. *IEEE Journal of Biomedical and Health Informatics*.

Folkman, J. 2002. Role of angiogenesis in tumor growth and metastasis. In *Seminars in oncology*, volume 29, 15–18. Elsevier.

Gong, X.; Yuan, S.; Xiang, Y.; Fan, L.; and Zhou, H. 2023. Domain knowledge-guided adversarial adaptive fusion of hybrid breast ultrasound data. *Computers in Biology and Medicine*, 164: 107256.

Guo, D.; Lu, C.; Chen, D.; Yuan, J.; Duan, Q.; Xue, Z.; Liu, S.; and Huang, Y. 2024. A multimodal breast cancer diagnosis method based on Knowledge-Augmented Deep Learning. *Biomedical Signal Processing and Control*, 90: 105843.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, Y.; Hu, H.; Zhu, Y.; and Xu, Y. 2023. Breast Lesion Diagnosis Using Static Images and Dynamic Video. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–5. IEEE.

Ioffe, S. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lei Ba, J.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *ArXiv e-prints*, arXiv–1607.

Li, M.; Gong, W.; Yan, P.; Li, X.; Jiang, Y.; Luo, H.; Zhou, H.; and Yin, S. 2025. Joint Lesion Detection and Classification of Breast Ultrasound Video via a Clinical Knowledge-Aware Framework. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1): 45–61.

Liang, X.; Lin, L.; Cao, Q.; Huang, R.; and Wang, Y. 2015. Recognizing focal liver lesions in CEUS with dynamically trained latent structured models. *IEEE Transactions on Medical Imaging*, 35(3): 713–727.

Liu, B.; Wang, D.; Yang, X.; Zhou, Y.; Yao, R.; Shao, Z.; and Zhao, J. 2022. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18041–18050.

Liu, Y.; Li, G.; and Lin, L. 2023. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 11624–11641.

Mo, Y.; Han, C.; Liu, Y.; Liu, M.; Shi, Z.; Lin, J.; Zhao, B.; Huang, C.; Qiu, B.; Cui, Y.; et al. 2023. Hover-trans: Anatomy-aware hover-transformer for roi-free breast cancer diagnosis in ultrasound images. *IEEE Transactions on Medical Imaging*, 42(6): 1696–1706.

Nie, S.; Zheng, M.; and Ji, Q. 2018. The deep regression bayesian network and its applications: Probabilistic deep learning for computer vision. *IEEE Signal Processing Magazine*, 35(1): 101–111.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12700–12710.

Pearl, J. 2009. *Causality*. Cambridge university press.

Pearl, J. 2016. *Causal inference in statistics: a primer*. John Wiley & Sons.

Pearl, J.; and Mackenzie, D. 2018. The book of why : the new science of cause and effect. *Science*, 361(6405): 855.2–855.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wan, P.; Xue, H.; Liu, C.; Chen, F.; Kong, W.; and Zhang, D. 2023. Dynamic perfusion representation and aggregation network for nodule segmentation using contrast-enhanced us. *IEEE Journal of Biomedical and Health Informatics*, 27(7): 3431–3442.

Wang, J.; Qiao, L.; Zhou, S.; Zhou, J.; Wang, J.; Li, J.; Ying, S.; Chang, C.; and Shi, J. 2024. Weakly supervised lesion detection and diagnosis for breast cancers with partially annotated ultrasound images. *IEEE Transactions on Medical Imaging*.

Wang, T.; Zhou, C.; Sun, Q.; and Zhang, H. 2021. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3091–3100.

Wang, Y.; Li, X.; Qi, Z.; Li, J.; Li, X.; Meng, X.; and Meng, L. 2022. Meta-causal feature learning for out-of-distribution generalization. In *European Conference on Computer Vision*, 530–545. Springer.

Xie, X.; Wang, Y.; Chen, C.; Wang, R.; Liu, X.; and Niu, J. 2023. IMAN: An Iterative Mutual-Aid Network for Breast Lesion Segmentation on Multi-modal Ultrasound Images. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 3954–3961. IEEE.

Xu, Z.; Wang, Y.; Chen, M.; and Zhang, Q. 2022. Multi-region radiomics for artificially intelligent diagnosis of breast cancer using multimodal ultrasound. *Computers in Biology and Medicine*, 149: 105920.

Yang, X.; Zhang, H.; and Cai, J. 2021. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 12996–13010.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9847–9857.

Yang, Z.; Gong, X.; Guo, Y.; and Liu, W. 2020. A temporal sequence dual-branch network for classifying hybrid ultrasound data of breast cancer. *Ieee Access*, 8: 82688–82699.

Zhang, H.; Xiao, L.; Cao, X.; and Foroosh, H. 2022. Multiple adverse weather conditions adaptation for object detection via causal intervention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3): 1742–1756.

Zhang, S.; Jiang, T.; Wang, T.; Kuang, K.; Zhao, Z.; Zhu, J.; Yu, J.; Yang, H.; and Wu, F. 2020. Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4373–4382.

Zhang, Y.; Huang, Z.-A.; Hong, Z.; Wu, S.; Wu, J.; and Tan, K. C. 2024. Mixed prototype correction for causal inference in medical image classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4377–4386.

Zhuang, S.; and Hadfield-Menell, D. 2020. Consequences of misaligned AI. *Advances in Neural Information Processing Systems*, 33: 15763–15773.