Causality-Inspired Single-Source Domain Generalization for Medical Image Segmentation

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, *Graduate Student Member, IEEE*, Chen Qin, Wenjia Bai, *Member, IEEE*, and Daniel Rueckert, *Fellow, IEEE*

***Abstract*— Deep learning models usually suffer from the domain shift issue, where models trained on one source domain do not generalize well to other unseen domains. In this work, we investigate the single-source domain generalization problem: training a deep network that is robust to unseen domains, under the condition that training data are only available from one source domain, which is common in medical imaging applications. We tackle this problem in the context of cross-domain medical image segmentation. In this scenario, domain shifts are mainly caused by different acquisition processes. We propose a simple causality-inspired data augmentation approach to expose a segmentation model to synthesized domain-shifted training examples. Specifically, 1) to make the deep model robust to discrepancies in image intensities and textures, we employ a family of randomly-weighted shallow networks. They augment training images using diverse appearance transformations. 2) Further we show that spurious correlations among objects in an image are detrimental to domain robustness. These correlations might be taken by the network as domain-specific clues for making predictions, and they may break on unseen domains. We remove these spurious correlations via causal intervention. This is achieved by resampling the appearances of potentially correlated objects independently. The proposed approach is validated on three cross-domain segmentation scenarios: cross-modality (CT-MRI) abdominal image segmentation, cross-sequence (bSSFP-LGE) cardiac MRI segmentation, and cross-site prostate MRI segmentation. The proposed approach yields consistent performance gains compared with competitive methods when tested on unseen domains.**

***Index Terms*— Domain generalization, image segmentation, causality, data augmentation.**

I. Introduction

Deep learning based medical image segmentation approaches [1], [2], [3], [4], [5], [6] usually achieve state-of-the-art performance when being trained and tested on datasets from a single *domain*, *i.e.* from identically distributed training and testing data. However, in practice, deep learning models perform less well when the testing data is drawn from a different distribution than that of the training data (*i.e.* a different *domain*) [7], [8]. The discrepancy between training and testing domains is termed as the *domain shift* [9]. In medical image segmentation, the most notorious source of domain shift is the differences in image acquisition processes (imaging modalities, scanning protocols, or device manufacturers) [10]. This type of domain shift is therefore termed as the *acquisition shift* [11]. We argue that the performance deterioration under acquisition shift can be attributed to the following two mechanisms: the *shifted domain-dependent features* and the *shifted-correlation effect*.

1) Shifted Domain-Dependent Features: Domain-dependent features include *intensities* and *textures*, which constitute image *appearance*, as shown in Fig. 1-A. Deep networks are susceptible to shifts in intensity and texture [7], [8]. This is in contrast to human annotators: they can easily find the correspondence of the same anatomical structure across different domains [8], usually by focusing on the *shape* information that is domain-invariant and is intuitively *causal* to human-defined segmentation masks, compared with intensity/texture.

2) Shifted-Correlation Effect: Due to a *confounder* (*i.e.* a "third" variable that spuriously correlates two variables of interest) [14], objects in the background might be correlated but **not** *causally* related to the objects of interest [15]. The network might take these objects in the background as clues for recognizing the objects of interest [15]. For example, in [12] and [13], as illustrated in Fig. 1-B, a model that

Manuscript received 27 July 2022; revised 7 October 2022; accepted 13 November 2022. Date of publication 23 November 2022; date of current version 3 April 2023. This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) Programme under Grant EP/P001009/1, in part by EPSRC under Grant EP/W01842X/1, in part by the Innovate U.K. under Grant 104691, and in part by the European Research Council (ERC) Grant 884622. Code is available at https://github.com/cheng-01037/Causality-Medical-Image-Domain-Generalization. *(Corresponding author: Cheng Ouyang.)*

Cheng Ouyang, Chen Chen, Surui Li, and Zeju Li are with the Department of Computing, Imperial College London, SW7 2AZ London, U.K. (e-mail: c.ouyang@imperial.ac.uk; chen.chen15@imperial.ac.uk; surui.li@imperial.ac.uk; zeju.li18@imperial.ac.uk).

Chen Qin is with the Department of Electrical and Electronic Engineering and Imperial-X, Imperial College London, SW7 2AZ London, U.K. (e-mail: c.qin15@imperial.ac.uk).

Wenjia Bai is with the Department of Computing, and the Department of Brain Sciences, Imperial College London, SW7 2AZ London, U.K. (e-mail: w.bai@imperial.ac.uk).

Daniel Rueckert is with the Department of Computing, Imperial College London, SW7 2AZ London, U.K., and also with the Institute for AI and Informatics in Medicine, Klinikum rechts der Isar, Technical University of Munich, 81675 München, Germany (e-mail: daniel.rueckert@tum.de).

Digital Object Identifier 10.1109/TMI.2022.3224067

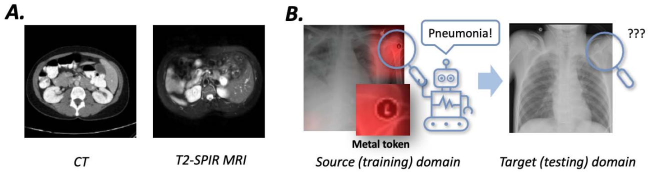


Fig. 1. **A. An example of domain-dependent features.** Appearance varies with imaging modalities (domains). Appearance-based decision rules that are built on one domain (*e.g.* CT) may fail on other domains (*e.g.* MRI). **B. An example of spurious correlations:** In the source domain, the model learns to detect pneumonia by looking at the metal token that is correlated with but not causally leads to the pneumonia label [12], [13][3]. However, these metal tokens may not exist in the target domain, potentially leading to model failure.

recognizes `pneumonia` in X-ray images is actually looking at the `metal token` in the background, which correlates with `pneumonia` due to the confounder: data selection bias. These correlations are often detrimental under domain shift. This is because decision rules based on these correlations may break in the shifted domain: the correlated objects in the background may disappear, or it may not co-shift in the same way as the objects of interest. In the above example, the model may fail or underperform on real-world images, where `metal token` does not exist, or does not correlate with `pneumonia` anymore.

To mitigate domain shift, previous attempts include *unsupervised domain adaptation* (UDA) [16] and *multi-source domain generalization* (MDG) approaches [17]. Unfortunately, UDA or MDG may not always be practical: they rely on training data from the target domain or from multiple source domains, which are often unavailable due to cost or privacy concerns. UDA also requires expertise for fine-tuning on target data, incurring difficulties to its deployment in the real world.

A more practical setting is *single-source domain generalization*: to train a deep learning model to be robust against domain shifts, using training data from only *one* source domain. Since no examples of the target domain are available, we resort to bottom-up approaches that are built on the above causal analysis of acquisition shift. We aim to 1) steer the network toward shape information which is domain-invariant and is intuitively causal to segmentation results; 2) to immunize the segmentation model against the shifted-correlation effect, by removing the confounder that spuriously correlates objects in the background and the objects of interest during training.

Learning causal features and removing confoundings usually require *intervention*: *fixing* the variable of interest while incorporating other variables in a fair way [14]. This is similar to randomized controlled trials. To mitigate acquisition shift, a straightforward intervention is to train the model with images of a fixed cohort of patients that are taken under all possible acquisition processes. Since this is unrealistic, we resort to data-augmentation-based intervention [18] that incorporates possible acquisition shifts via simulations.

In this work, we propose a causality-inspired data augmentation approach for single-source domain generalization. It exposes the network to synthetic acquisition-shifted training samples that incorporate shifts in intensity/texture and shifted correlations. Specifically, to efficiently synthesize diverse appearances (intensities and textures) without losing generality, we employ shallow convolutional networks with random weights that are sampled at each training iteration to augment images. As stable decision rules can hardly be formed on constantly-varying intensities/textures, the network would resort to domain-invariant features such as shapes. To remove the confounder that leads to the shifted-correlation effect, we first reveal that the image acquisition process naturally confounds objects in the background and the objects of interest, in terms of their appearances. We then design a practical method for simulating and independently resampling the appearances of potentially confounded objects during training. This is achieved by applying different appearance transformations in a spatially-variable manner, with the help of pseudo-correlation maps computed using unsupervised algorithms. The overall approach is used as additional stages following standard data augmentations. It is therefore generic to architectures of the segmentation network. In summary, we make the following contributions:

- We investigate single-source domain generalization problem for cross-domain medical image segmentation from a causal view. We propose a simple and effective causality-inspired data augmentation approach.
- We propose 1) global intensity non-linear augmentation (GIN) technique that efficiently transforms images to have diverse appearances via randomly-weighted shallow convolutional networks; 2) interventional pseudo-correlation augmentation (IPA) technique that removes the confounder that leads to the shifted-correlation effect. This is realized by independently resampling appearances of potentially confounded objects. These two components function as cores of the proposed approach.
- We build a comprehensive testing environment for single-source domain generalization for cross-domain medical image segmentation. It covers cross-modality, cross-sequence (MRI) and cross-site settings with various anatomical structures. We hope this testing environment to facilitate future works on domain robustness for medical image segmentation.

II. Related Works

A. Unsupervised Domain Adaptation and Domain Generalization for Medical Image Segmentation

Considerable efforts have been made to alleviate domain shift for deep networks. Unsupervised domain adaptation (UDA) transfers a model trained on a source domain to a target domain using unlabeled target-domain data [9]. Existing techniques are mainly based on distribution alignment [16], [19], or self-training [20]. The former enforces features from two domains to have the same distribution, by minimizing the distance between two distributions [16], [19]. By this mean the decision rules in the source domain can be applied to the target

[1] Image adapted from [12] under the terms of Creative Commons Attribution License.

domain [9], [16], [19]. The latter first predicts pseudolabels in the target images, and then fine-tunes the source model on pseudo-labeled target images [20]. Applications of UDA in medical image segmentation include [7], [8], [21], [22], [23].

Multi-source domain generalization (MDG) usually learns domain-invariant features in a one-off manner from multiple source domains. Recent techniques include meta-learning [24], [25], style transfer [26], [27], transfer learning [28], dynamic networks [29] and so on. Among them meta-learning optimizes model parameters on synthetic domain generalization sub-problems [24]. Style transfer augments the training data by swapping/mixing appearances among training images from different domains [27]. Transfer learning techniques exploit the mutual-boosting effect between the main task and the carefully-designed auxiliary task [28]. Dynamic networks make the model weights/architectures adaptive to the specific domain of each input [29]. In medical imaging, recent methods [30], [31], [32] have achieved promising results. However, for both UDA and MDG, target data or multi-source data are often unavailable due to privacy or cost concerns.

Single-source domain generalization requires training data from one domain only. Representation self-challenging (RSC) [33] removes features that cause the largest loss gradients, which are believed to be domain-dependent. Liu et al. [34] propose to unify statistics of image features, which control image styles [35], among images from different domains. A major stream of works exploits data augmentation: training the network on deliberately perturbed samples to improve network robustness to real-world perturbations [36], [37], [38]. We review this stream of methods later with more details. Some most recent works such as [39] employ multiple techniques: data augmentation [40], adversarial training [41] and contrastive learning [42].

B. Data Augmentation for Domain Robustness

Theoretical analysis [43] suggests that data augmentation improves domain generalization by enlarging the span of the data and by regularizing decision boundaries. In practice, Cutout [37] strengthens robustness against the feature missing caused by domain shift, by partially occluding training images. Mixup [44], [45] regularizes decision boundaries by interpolating among training samples. RandConv [40] drives the network to learn shape information, which is domain-invariant, by randomly altering image textures using linear filtering. Our method is closely related to RandConv [40]. However, we show in our experiments that the linear filtering mechanism in RandConv [40] is oversimplified for accounting for domain gaps in real-world settings. Adversarial data augmentation generates image perturbations that easily flip predictions of classifiers [38], [46], [47], [48].

In medical imaging, Zhang et al. [4] employ a stack of photometric and geometric transformations to training images to improve domain robustness. Billot et al. [49] propose a contrast-agnostic brain MRI segmentation strategy, which synthesizes training examples by sampling from pre-built generative models of brain images. However, this method necessities well-defined generative models that map segmentation labels to images, which is usually unavailable in most of medical imaging applications. Furthermore, these generative models are often oversimplified. AdvBias [50] is specially designed for medical image segmentation. It employs an adversarial augmentation technique based on a multiplicative bias field model. It outperforms a series of competitive methods on cross-site MRI segmentation [51].

C. Leveraging Causality for Robust Deep Learning

As discussed in Sec. I, learning causalities and mitigating confoundings usually require *causal intervention* [14]. In causal intervention, the variable of interest in a causal relationship is *fixed*, while other variables are fairly incorporated. A model then learns causalities from these intervened samples. For example, a model that recognizes a `camel` might be mistakenly focusing on the background: `desert` [52], since most of pictures of camels are taken in deserts. In this case, intervention can be done by incorporating different backgrounds: training with pictures of camels that are taken in diverse backgrounds such as grassland and city. By this mean the model would learn that it is the camel rather than the desert that *cause* a `camel` label. Causal relationships are usually modeled using structural causal models (SCM) [14]. The *fixing* operation is usually noted as $do(\cdot)$ [14]. The distribution $p(Y|do(X = \texttt{camel}))$ is called *interventional distribution*. Compared with conditional distribution $p(Y|X = \texttt{camel})$ that reflects *correlation* in the observed data, $p(Y|do(X = \texttt{camel}))$ reflects *causation* [14].

Causal ideas have been used for discovering image features that are semantically essential and robust [52], [53], [54], [55]. Invariant risk minimization [52] learns causal image representations by enforcing these representations to be Bayesian optimal in all the environments. Mahajan et al. [53] improve domain robustness using contrastive losses. Atzmon et al. [54] propose a causal mechanism to generalize a model to novel samples with unseen combinations of attributions.

Our idea of using data-augmentation-based intervention is inspired by [18] and [55]. Reference [18] proves that *post-hoc* data augmentation theoretically commutes with “physical” intervention. Mitrovic et al. [55] derive a practical loss function for causality-based domain generalization. Different from [55], we focus on the unanswered practical problem of designing an augmentation technique that is tailored to the real-world problem: cross-domain medical image segmentation. Our work is also related to causal weakly-supervised segmentation by Zhang et al. [15], as both works study the adverse effect of confoundings among objects in image segmentation. While Zhang et al. [15] focus on the intra-domain scenario, we focus on the effect of confoundings under domain shift.

III. Method

Our causality-inspired data augmentation approach aims to improve network robustness against domain shift, in particular, shifts caused by the differences in acquisition processes [10]. Based our analysis in Sec. I, we propose to expose the

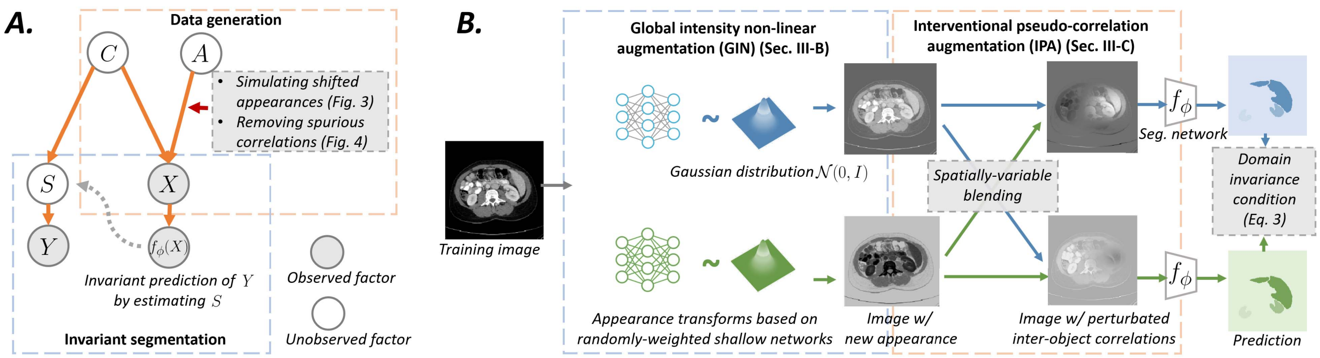

A.
Data generation
C
A
Simulating shifted appearances (Fig. 3)
Removing spurious correlations (Fig. 4)
S
X
Y
$f_\phi(X)$
Invariant prediction of Y by estimating S
Invariant segmentation
Observed factor
Unobserved factor
B.
Global intensity non-linear augmentation (GIN) (Sec. III-B)
Interventional pseudo-correlation augmentation (IPA) (Sec. III-C)
Training image
Gaussian distribution $\mathcal{N}(0, I)$
Spatially-variable blending
Appearance transforms based on randomly-weighted shallow networks
Image w/ new appearance
Image w/ perturbated inter-object correlations
f_ϕ
Seg. network
Domain invariance condition (Eq. 3)
Prediction

Fig. 2. **A. The data generation process of medical images:** An image X is modeled as the effect of two independent factors: 1) a domain-independent content factor C which represents the geometric shapes of the patient anatomy; 2) and a domain-dependent acquisition factor A which controls image appearances. **Mechanism of the domain-invariant segmentation network:** We assume that 1) there exists a domain-invariant representation S that encodes shape information (S is caused by C); 2) the ground-truth Y can be derived from S. A domain-invariant segmentation network $f_\phi(\cdot)$ is trained to implicitly estimate S in spite of different simulated acquisition processes A's. **B. Workflow of the proposed data augmentation approach:** It simulates different possible acquisition processes, hence domains. GIN transforms images to have new appearances, using randomly-weighted shallow networks. IPA blends two GIN-augmented versions of the same image in a spatially-variable manner. This blending operation perturbs/randomizes spurious correlations among patches, in terms of their appearances. The segmentation network $f_\phi(\cdot)$ is trained to make consistent predictions, unaffected by simulated appearances and perturbed correlations.

network to training examples that incorporate simulated intensity/texture shifts and shifted correlations among objects.

Specifically, our approach is a synergy of a *global intensity non-linear augmentation* technique (GIN) and an *interventional pseudo-correlation augmentation* technique (IPA). As shown in Fig. 2-B, GIN transfers training images to have diverse appearances while it keeps the shapes of anatomical structures unchanged, discouraging the network from biasing towards appearances. IPA resamples possible appearances of potentially spuriously correlated objects (due to confounding) in the background and those of the objects of interest, in an independent and diverse manner. This is implemented as spatially-variable blendings between two GIN-augmented images. The entire approach functions as additional steps in a standard data augmentation pipeline.

In the following sections, we first introduce the general problem formulation of the proposed data augmentation approach. We introduce GIN in Sec. III-B, with a detailed reasoning behind its design choices. IPA is introduced in Sec. III-C, where we firstly reveal that it is the *image acquisition process* that naturally confounds objects in the background and the objects of interest. We then describe how IPA removes confoundings. Finally, we summarize the overall training process.

A. Problem Formulation

1) A Causal View of Image Generation and Segmentation: We first introduce the problem formulation of our data-augmentation-based single-source domain generalization approach. Inspired by recent works [54], [55], we model the data generation process and the (ideally, domain-invariant) segmentation process using the causal model shown in Fig. 2-A. Specifically, we make the following assumptions:

1) $A \rightarrow X \leftarrow C$: An image X is generated from two independent variables (factors): *acquisition* A and *content* C. C represents the shapes of underlying anatomical structures of the patient, while A represents the acquisition process. The factor A maps different types of tissues of the patient in the scanner into different pixel values in the image.

2) $C \rightarrow S \rightarrow Y$: There exists an ideal domain-invariant representation S, determined by C. S is in the form of feature maps of the deep layers of the network and it contains the shape information of the objects of interest. The ground-truth segmentation mask Y can be derived from S.

3) $X \rightarrow f_\phi(X)$: The segmentation network $f_\phi(\cdot)$ takes X as the input and predicts Y, by implicitly estimating S.

A and C are independent: changing A does not affect C, S or Y. Of note, our discussion is constrained to acquisition shift. C is assumed to be unchanged across the source domain and the target domain(s) in our experiments.

2) Causal Intervention for Domain Robustness: According to [55], our argument that S *is invariant to shifts of* A can be formally written as follows:

$$p(Y|S, do(A = a_i)) = p(Y|S, do(A = a_j)), \forall a_i, a_j. \quad (1)$$

Here $p(Y|S, do(A = a_i))$ denotes the distribution that comes from letting images to be generated from a specific acquisition process $A = a_i$ [14], [55], for example MRI. Using a symmetric notation, we let a_j to be another acquisition process, for example CT. Eq. 1 suggests that ideally, this distribution should remain the same regardless of the acquisition processes.

In practice, as shown in Fig.2-A, we use a segmentation network $f_\phi(\cdot)$ parameterized by ϕ to predict Y. To make $f_\phi(\cdot)$ domain invariant, ideally, $f_\phi(\cdot)$ should implicitly estimate S in the last layers of $f_\phi(\cdot)$. However, the condition in Eq. 1 cannot be directly used to train $f_\phi(\cdot)$, as "physical" interventions on A (scanning patients under all possible acquisition processes) is impractical. Fortunately, Ilse et al. [18] have demonstrated that data augmentations can be used as "virtual" causal interventions. Therefore, we assume that for each a_i, there exists a photometric transformation function $\mathcal{T}_i(\cdot)$ that transforms the image X to be like from a_i.[2] We therefore have:

$$p(Y|S, do(A = a_i)) \approx p(Y|f_\phi(\mathcal{T}_i(X)\ )). \tag{2}$$

Combining Eq. 1 and 2 leads to a practical domain invariance condition: minimizing the difference between distributions that come from different photometric transformations $\mathcal{T}_i(\cdot)$ and $\mathcal{T}_j(\cdot)$ [55]. By combining this domain invariance condition and the image segmentation loss, we can now derive our loss function (inspired by [55]). For each iteration, we have

$$\begin{aligned}\mathcal{L}(\phi) = &\ \mathcal{L}_{seg}(f_\phi(\mathcal{T}_i(\mathbf{x})), \mathbf{y}) + \mathcal{L}_{seg}(f_\phi(\mathcal{T}_j(\mathbf{x})), \mathbf{y}) \\ &+ \lambda_{div}\mathcal{D}(p(\mathbf{y}|f_\phi(\mathcal{T}_i(\mathbf{x})))\|p(\mathbf{y}|f_\phi(\mathcal{T}_j(\mathbf{x})))).\end{aligned} \tag{3}$$

Here $(\mathbf{x}, \mathbf{y}) \sim p(X, Y)$ denotes an image-label pair in the training dataset, $\mathcal{L}_{seg}(\cdot, \cdot)$ is a segmentation loss, *e.g.* cross-entropy; $\mathcal{T}_i(\cdot)$ and $\mathcal{T}_j(\cdot)$ are two different photometric transformations that simulate the effect of a_i or a_j respectively, randomly sampled from a family of photometric transformations at each iteration; $\mathcal{D}(\cdot\|\cdot)$ is the Kullback–Leibler divergence: measuring the gaps between two distributions; λ_{div} is a weighting coefficient. Similar divergence terms have also been used for semi-supervised learning [46], [56], although they are not derived from a causal perspective.

As implied by Eq. 3, the photometric transformations $\{\mathcal{T}(\cdot)\}$ serve as the core of the domain invariance condition. Since the target-domain data are unavailable, we resort to build $\{\mathcal{T}(\cdot)\}$ in a bottom-up manner, based on our analysis on ingredients of acquisition shift, as discussed in Sec. I. We simulate $\{\mathcal{T}(\cdot)\}$ as a combination of GIN and IPA.

B. Global Intensity Non-Linear Augmentation

1) Design Choices: GIN is designed to efficiently transform image intensities and textures. We configure GIN as a family of piece-wise linear functions $g(\cdot) \in \mathcal{G}$, operating in pixel level or small local patch level in a spatially-equivariant manner. These functions take a training image $\mathbf{x}$ from the source domain as input, and outputs an image with the same shape information but different intensities/textures, namely, $g(\cdot) : \mathbb{R}^{C_h \times H \times W} \rightarrow \mathbb{R}^{C_h \times H \times W}$, where (H, W) to be the spatial size of a 2-D image $\mathbf{x}$, and C_h to be the number of channels. As shown in Fig. 3, transformations sampled from GIN are instantiated as shallow multi-layer convolutional networks. These networks are composed with 1) random convolutional kernels θ sampled from Gaussian distributions $\mathcal{N}(0, I)$ with small receptive fields (to avoid over-blurring). 2) Leaky ReLU non-linearities between two neighboring convolutional layers (to make transformations non-linear). At each iteration, new kernels are sampled, yielding a variety of transformation functions.[3] Inspired by [40], we perform a linear interpolation between the original and the output of the random network. In the end, as depicted in Fig. 3, to constrain the energy of the augmented image, the output image is re-normalized to have the same Frobenius norm as the original input $\mathbf{x}$. We note the overall transformation with parameters θ as $g_\theta(\cdot)$ and its pure network part as $g_\theta^{Net}(\cdot)$ (see Fig. 3), and note a random interpolation coefficient sampled from uniform distribution $\mathcal{U}(0, 1)$ as α. We can write the transformed image $g_\theta(\mathbf{x})$ as follows:

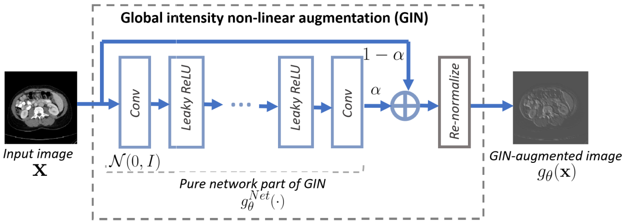


Fig. 3. **Illustration of the proposed global intensity non-linear augmentation (GIN) module.** It transforms image appearances using shallow convolutional networks whose weights are randomly sampled at each iteration. It also contains Leaky ReLU's interleaved between convolutional layers. To maintain spatial resolutions of the input images, these random networks do not contain any downsampling operations.

$$g_\theta(\mathbf{x}) = \frac{\alpha g_\theta^{Net}(\mathbf{x}) + (1-\alpha)\mathbf{x}}{\|\alpha g_\theta^{Net}(\mathbf{x}) + (1-\alpha)\mathbf{x}\|_F} \cdot \|\mathbf{x}\|_F, \tag{4}$$

where $\|\cdot\|_F$ is the Frobenius norm.

By this mean, at each iteration, different intensities and textures are given to training images. As stable decision rules can hardly be built on randomly changing intensities/textures, the network would resort to invariant information like shapes.

2) Advantages: We highlight major advantages of the above designs: Firstly, GIN is based on generic assumptions on intensity/texture transformations. It therefore avoids being over-specific to a certain target domain(s). In addition, GIN is computationally efficient: it is in the form of shallow networks and therefore easily exploits GPUs for acceleration. Also, it is differentiable, and therefore can be integrated into adversarial augmentation frameworks [38], [46], [47], [50] to improved data efficiency.

C. Interventional Pseudo-Correlation Augmentation

1) Confounded Objects in the Background Affect Segmentation: Recall in Sec. I, spurious correlations (due to confoundings) between objects in the background and the objects of interest in the source domain might be taken by the network as domain-specific clues for making predictions.[4] These decision rules may break in the target domain, leading to performance downgrades. This is because confounded objects in the background that benefit segmentation in source domain, might not exist in the target domain. Alternatively, they might not co-shift in the same way as those of the objects of interest.

[2] Arguably this transformation can be interpreted as generating *counterfactual* examples [11], [14]: given an observed image from a certain acquisition process, asking how this image could look like should it had been generated from another imaging process. We do not interpret our data augmentation as counterfactual generation since the aim of our approach is not to synthesize the appearance of a *specific* domain.

[3] For more theoretical analysis on Gaussian randomly weighted networks, we refer readers to [57].

[4] In our preliminary experiment, we found that the model does learn background-based decision rules: We distorted the background of images by randomly swapping patches of the background (therefore the global statistics of image intensities would remain unchanged). We then tested a segmentation network on these background-distorted images, and have observed substantial performance downgrades compared with results on the original undistorted images. This phenomenon has also been found in general computer vision and these spurious correlations are sometimes termed as *context bias* [15], [58].

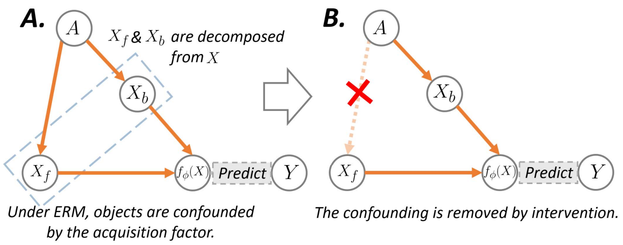

A.
A
X_f & X_b are decomposed from X
X_b
X_f
$f_\phi(X)$
Predict
Y
Under ERM, objects are confounded by the acquisition factor.
B.
A
X_b
X_f
$f_\phi(X)$
Predict
Y
The confounding is removed by intervention.

Fig. 4. **A. Causal graph illustrating spurious correlations between the object of interest X_f and a potentially correlated background patch X_b**, both are from the image X. Spurious correlations between X_f and X_b are due to the acquisition factor A. The network $f_\phi(\cdot)$ may learn source-domain-specific decision rules from these spurious correlations. However, these correlations may break on out-of-domain data. This causal graph is a close-up of the data generation process in Fig. 2-A. The content factor C in Fig. 2-A is assumed to be domain-independent and therefore omitted here. **B. Removing the confounding caused by A**: This is achieved by the causal intervention $do(X_f = \mathrm{x}_f)$, which removes $A \rightarrow X_f$ in the causal graph.

From the perspective of network architectures, objects in the background often affect predictions of the objects of interest by the following ways: 1) Background features can affect global feature statistics at normalization layers [59], since feature statistics are usually calculated across all spatial locations. 2) The large receptive fields often make the pixels of the objects of interest and those of the neighboring background to be inevitably perceived and processed together [60].

2) The Acquisition Process Naturally Confounds Objects: To mitigate the shifted-correlation effect, it is worthwhile to point out that it is the *acquisition factor* A that leads to confounding. It naturally creates spurious correlations between certain objects in the background and the object of interest.[5] To demonstrate this, in an image X, we consider the patch of object of interest X_f and the patch of a potentially correlated unlabeled object X_b in the background. We zoom-in the causal relations in Fig. 2-A using X_f and X_b, and redraw that in Fig. 4-A. We can see:

1) $X_f \rightarrow f_\phi(X) \leftarrow X_b$: Although X_f already contains sufficient information for delineating Y, in practice X_b also affects the network features and the output, as both X_f and X_b are processed by $f_\phi(\cdot)$.

2) $f_\phi(X) \leftarrow X_b \leftarrow A \rightarrow X_f$: More importantly, the confounding effect of A that correlates X_b and X_f is revealed in the path $X_b \leftarrow A \rightarrow X_f$. This corresponds to the fact that given a certain acquisition process, the *same* imaging physical mechanism that maps different tissues to different pixel values, applies to *both* the object of interest and the objects in the background. Without such a path, appearances of X_f and X_b would vary independently in the training dataset. Stable correlations regarding their appearances could unlikely be established and learned.

Of note, as we assume the content factor C to remain unchanged across domains, we ignore the confoundings caused by C, and omit C in Fig. 4. Under this assumption, the content information (rather than the appearance information) in the background could still be utilized for locating the object of interest.

3) Removing Confoundings by Intervention: We propose to mitigate the shifted correlation effect during the training stage, by removing the confounding $X_b \leftarrow A \rightarrow X_f$ using the intervention $do(X_f = \mathbf{x}_f)$ [14]. This operation resamples the appearances of correlated objects X_b's, independent of X_f's. This intervention in effect removes $A \rightarrow X_f$. Formally, we are learning the interventional distribution $p(Y|do(X_f = \mathbf{x}_f))$ based on the intervened causal diagram Fig. 4-B:

$$\begin{aligned} p(Y|do(X_f = \mathbf{x}_f)) &= \sum_{\mathbf{x}_b} p(Y|\mathbf{x}_f, \mathbf{x}_b)p(\mathbf{x}_b) \\ &= \sum_{\mathbf{x}_b}\sum_{a} p(Y|\mathbf{x}_f, \mathbf{x}_b)p(\mathbf{x}_b|a)p(a). \end{aligned} \tag{5}$$

Here $\mathbf{x}_f, \mathbf{x}_b \in \mathbf{x}$; $(\mathbf{x}, \mathbf{y}) \sim p(X, Y)$; $a \sim p(A)$ and $p(A)$ is a prior of possible acquisition processes. Eq. 5 translates to independently sampling possible appearances of X_b.

Unfortunately, to compute Eq. 5 we are faced with three practical issues: 1) we do not know which object is correlated with the object of interest and there is no ground-truth map of it; 2) there might be more than one object in the background that correlates with the object of interest, and their effects might be entangled; 3) directly fixing $\mathbf{x}_f$ using ground-truth masks $\mathbf{y}$ would make $\mathbf{x}_f$ unnaturally stand out from the background, providing shortcuts for the network to recognize $\mathbf{x}_f$. Crucially, this intervention *cannot* be realized by GIN alone: If two patches share the same appearance, their appearances would remain the same after being transformed by GIN.

4) Spatially-Variable Blending: As a practical solution, we employ *interventional pseudo-correlation augmentation* (IPA), which approximates the intervention $do(X_f = \mathbf{x}_f)$. IPA is built on appearance transformations of GIN. We use *pseudo-correlation maps* as surrogates of label maps of $\mathbf{x}_b$'s, for allocating transformation functions to different pixels of the image: Pixels that correspond to different values in the pseudo-correlation map would be given different transformation functions. Pseudo-correlation maps are generated using the unsupervised algorithm [61]. To account for different potential spurious correlations, we use different randomly-sampled maps at each iteration. To avoid the shortcuts caused by fixing $\mathbf{x}_f$ using $\mathbf{y}$, we apply pseudo-correlation maps to both $\mathbf{x}_b$'s and $\mathbf{x}_f$ (*i.e.* to the entire image).

To improve computation efficiency, as shown in Fig. 5-A, we use pseudo-correlation maps as coefficients for blending pixels from two GIN-augmented versions of a same image. Considering a pseudo-correlation map $\mathbf{b} \in \mathbb{R}^{C_h \times H \times W}$ where all entries $b \in \mathbf{b}$ satisfy $b \in [0, 1]$, we have the output image of IPA $\mathcal{T}_1(\mathbf{x})$ as:

$$\mathcal{T}_1(\mathbf{x}; \theta_1, \theta_2, \mathbf{b}) = g_{\theta_1}(\mathbf{x}) \odot \mathbf{b} + g_{\theta_2}(\mathbf{x}) \odot (1 - \mathbf{b}). \tag{6}$$

Here $\mathcal{T}_1(\cdot)$ denotes the overall combined effect of GIN and IPA; $\odot$ denotes the Hadamard product; $(g_{\theta_1}(\cdot), g_{\theta_2}(\cdot))$ are two random appearance transformations sampled from GIN.

[5] For the ease of illustration, in the following analysis, we focus on the scenario where only one object of interest is available. Our conclusion naturally holds for multi-class segmentation as well, and has been validated by our experiments on abdominal and cardiac segmentations.

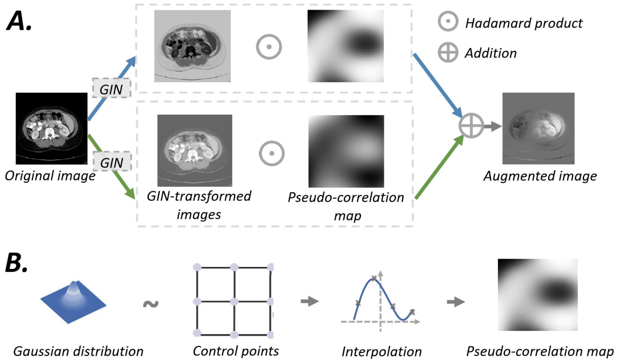


Fig. 5. **A. Implementation of IPA:** In each iteration, a new pseudo-correlation map is generated, and it is used as pixel-wise coefficients for blending two GIN-transformed images. This is equivalent to assigning different appearance transformations to different pixels/patches, according to their corresponding values in the pseudo-correlation map. As the pseudo-correlation maps used in each iteration are different, during the training process, IPA can approximate the operation of resampling appearances of potentially confounded objects. **B. Computing pseudo-correlation maps:** These maps are generated by interpolating along a lattice of randomly-valued control points.

We simultaneously obtain one additional augmented image $\mathcal{T}_2(\mathbf{x})$ by swapping the positions of $\mathbf{b}$ and $1-\mathbf{b}$ in Eq. 6. Here the subscript 1 or 2 of $\mathcal{T}(\cdot)$'s denotes whether it is $g_{\theta_1}(\mathbf{x})$ or $g_{\theta_2}(\mathbf{x})$ to be multiplied with $\mathbf{b}$. Of note, this operation can also be interpreted as an extension to AugMix [62] which is designed for image classification. Different from AugMix, IPA necessities strict spatial correspondence between pixels and labels to ensure the accuracy of this pixel-wise prediction. Also, the blending coefficients of IPA are intentionally made to be spatially variable to simulate our causal intervention.

5) Pseudo-Correlation Maps: As shown in Fig. 5-B, we configure pseudo-correlation maps as a field of continuous randomly-valued scalars with a low spatial frequency. They are interpolated from a lattice of randomly-valued control points, using cubic B-spline kernels [61], based on the efficient implementation from [63]. The spacing between two neighboring control points is empirically set to be 1/4 of image length to avoid introducing unnaturally large image gradients. This configuration features the following advantages: 1) It allows spatially-variable intensity transformation while does not severely distort shape information due to its low spatial frequency. 2) It further increases diversities of appearances by interpolating between two appearances.

D. Training Objective

The overall training is end-to-end using the loss function described in Eq. 3. For the ease of implementation we let the output of $f_\phi(\cdot)$ to be in the form of raw logits, and let $p(\mathbf{y}|f_\phi(\cdot))$ to be the probabilities obtained by passing the output of $f_\phi(\cdot)$ to a softmax function. For $\mathcal{L}_{seg}$, we employ a sum of multi-class cross-entropy loss and soft Dice loss. We set the weighting coefficient λ_{div} in Eq. 3 to be 10.0, same as in [40]. After training, the segmentation network $f_\phi(\cdot)$ is ready to be directly applied to unseen testing domains. The overall algorithm flow is summarized in Algorithm 1.

Algorithm 1 End-to-End Training With Causality-Inspired Data Augmentation

```
Require: Training dataset $\{(\mathbf{x},\mathbf{y})\}$ from the source domain, segmentation network $f_\phi(\cdot)$, global intensity non-linear augmentaton (GIN), interventional pseudo-correlation augmentation (IPA), number of iterations N, learning rate $l_r^{(t)}$ at iteration $t$.
for $t = 1 \ldots$N do
    Sample an image-label pair $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ from $\{(\mathbf{x},\mathbf{y})\}$.
    Sample intensity/texture transformations $g_{\theta_1}^{(t)}(\cdot), g_{\theta_2}^{(t)}(\cdot)$ using GIN.
    Apply $g_{\theta_1}^{(t)}(\cdot), g_{\theta_2}^{(t)}(\cdot)$ as described in Eq. 4 to $\mathbf{x}^{(t)}$ and obtain $g_{\theta_1}^{(t)}(\mathbf{x}^{(t)}), g_{\theta_2}^{(t)}(\mathbf{x}^{(t)})$.
    Compute a pseudo-correlation map $\mathbf{b}^{(t)}$ by interpolating along randomly valued control points as in Fig. 5-B.
    Compute augmented images $\mathcal{T}_1^{(t)}(\mathbf{x}^{(t)}), \mathcal{T}_2^{(t)}(\mathbf{x}^{(t)})$ using IPA, as described in Eq. 6.
    Compute training loss $\mathcal{L}^{(t)}(\phi)$ using Eq. 3.
    Update parameters of the segmentation network $f_\phi(\cdot)$: $\phi^{(t+1)} := \phi^{(t)} - l_r^{(t)} \frac{\partial \mathcal{L}^{(t)}(\phi)}{\partial \phi}$.
end for
```

IV. Experiments

A. Datasets and Evaluation Protocols

The proposed approach is evaluated in three cross-domain settings: 1) cross-modality abdominal segmentation between CT and T2-SPIR MRI (Abdominal Cross-modality), 2) cross-sequence cardiac segmentation from bSSFP MRI to LGE MRI (Cardiac Cross-sequence), and 3) prostate segmentation on MRI across six sites (Prostate Cross-site). Details of the datasets and the source-target splits are summarized in Table I. All these datasets are originally in 3-D and have been reformatted to 2-D, then resized to 192 × 192, and padded along the channel dimension to fit into the network. For the abdominal CT dataset, we applied a window of [-125, 275] [70] in Housefield values. For all MRI images, we clipped the top 0.5% of the histograms. We normalized all the 3-D scans to have zero mean and unit variance. For fairness of comparisons, for all the methods evaluated (including ERM), conventional geometric augmentations: affine transformations and elastic transformations; and intensity augmentations: brightness, contrast, gamma transformations and additive Gaussian noises were applied by default.

We employed the commonly-used Dice score (0-100) as the evaluation metric for measuring the overlap between the prediction and the ground truth. For abdominal and prostate segmentations, for the source domain, we used a 70%-10%-20% split for training, validation and testing sets; for the target domain(s), we used all the images for testing, same as in [30]. For cross-site prostate segmentation, each time we took one domain as the source and the rest five domains as targets, and we computed Dice scores averaged by target domains. This 1-versus-5 experiment is repeated for each of all six domains. For cardiac segmentation, we employed the same data split as in the cross-sequence segmentation challenge [66].

TABLE I
DETAILS OF CROSS-DOMAIN SEGMENTATION DATASETS USED IN THIS STUDY

Name	Label(s)	View	Split	Domain(s)	No. of 3-D scans	Origin
Abdominal Cross-modality	Liver, L-kidney, R-kidney, Spleen	Axial	Source/Target Target/Source	Computed tomography (CT) T2 spectral presaturation with inversion recovery (SPIR) MRI	30 20	[64] [65]
Cardiac Cross-sequence	L-ventricle, Myocardium, R-ventricle	Short-axis	Source Target	Balanced steady-state free precession (bSSFP) MRI Late gadolinium enhanced (LGE) MRI	45 45	[66]
Prostate Cross-site	Prostate	Axial	1 Source 5 Targets	Prostate MRI from 6 sites	30, 30, 19 13, 12, 12	[30] [67]–[69]

TABLE II
SEGMENTATION RESULTS ON THREE CROSS-DOMAIN SCENARIOS, WHERE A MODEL IS TRAINED ON THE SOURCE DOMAIN AND TESTED ON THE TARGET DOMAIN(S). DICE SCORE IS USED AS THE EVALUATION METRIC. THE HIGHEST SCORES ARE IN RED, THE SECOND-HIGHEST SCORES ARE IN BLUE

Method	Abdominal Cross-modality		Cardiac Cross-sequence	Prostate Cross-site						
	CT to MRI	MRI to CT	bSSFP to LGE	A to Rest	B to Rest	C to Rest	D to Rest	E to Rest	F to Rest	Overall
Upper bound	90.85	91.76	88.15	83.75	84.78	84.92	84.98	86.68	84.92	85.01
ERM	77.31	67.59	75.99	71.81	65.56	43.98	71.97	48.39	37.82	56.59
Cutout [37]	80.12	74.84	78.87	78.36	69.08	63.45	66.39	61.88	60.19	66.56
RSC [33]	74.09	71.50	77.51	72.81	70.18	49.18	74.11	54.73	43.69	60.78
MixStyle [26]	77.80	69.62	75.21	73.24	58.06	44.75	66.78	49.81	49.73	57.06
AdvBias [50]	80.17	75.00	79.62	78.15	62.24	54.73	72.65	53.14	51.00	61.98
RandConv [40]	80.66	74.69	83.73	77.28	60.77	53.54	66.21	52.12	36.52	57.74
Proposed	86.31	79.62	85.01*	82.14	67.21	59.11	73.16	67.38	73.23	70.37
Improve. over baseline	+9.01	+12.04	+9.02	+10.33	+1.66	+15.13	+1.20	+18.99	+35.40	+13.79

Single-sided Wilcoxon singed-rank tests are performed when the performance gains of the proposed method are below 2% in Dice score, compared with the second-highest method(s). *:p-value $\ll 1 \times 10^{-4}$.

B. Network Architecture and Training Configurations

We configured the segmentation network $f_\phi(\cdot)$ as a U-Net [1], the most commonly used network architecture for medical image segmentation, with an EfficientNet-b2 backbone [71]. For our proposed method, we trained the segmentation network using an Adam optimizer [72] with an initial learning rate of 3×10^{-4} with learning rate decay. We evaluated our method at the 2k-th epoch where the learning rate decays to zero.

C. Quantitative and Qualitative Results

We compared our method with the empirical risk minimization (ERM) baseline and several recent single-source domain generalization methods. Among them Cutout [37] enforces the model to be robust to corruptions by deliberately removing patches from training images. RSC [33] defines features that lead to the largest gradients as non-robust features and removes them in training. MixStyle [26] synthesizes novel domains by mixing instance-level feature statistics [35]. AdvBias [50] is designed for medical images. It augments images using adversarial perturbations. Closely related to our work is RandConv [40], which employs a random linear intensity transformation model to synthesize novel domains.

Table II summarizes performances on three cross-domain segmentation scenarios, where a network is trained on the source domain and evaluated on the target domain(s). The proposed approach in general outperforms peer methods. In particular, the performance gains of our method compared with the closely-related RandConv [40] suggest that our approach simulates domain shifts in a more effective manner, leading to stronger robustness upon unseen domains. We also provide the upper bounds: *i.e.* training and testing in the target domain, in Table II. Qualitative examples are shown in Fig. 6.

TABLE III
EFFECTS OF NUMBER OF CONVOLUTIONAL LAYERS (LEFT) AND NUMBER OF CHANNELS IN HIDDEN LAYERS (RIGHT) ON THE PERFORMANCE OF ABDOMINAL CT-MRI SEGMENTATION

No. of layers	Average Dice Score	No. of channels in hidden layers	Average Dice Score
2	84.26	2 (proposed)	86.04
4 (proposed)	86.04	4	85.49
8	85.66	8	83.21
16	79.17	16	81.90

To visualize the feature spaces, in Fig. 7 we show t-SNE of the target domain features collected at the last hidden layer of abdominal segmentation networks. As can be seen, for our proposed method, for the same class, target domain features stay close to those of the source domain; while features of different classes are separated.

D. Ablation Studies

1) Configurations of GIN: To investigate the effect of configurations of GIN on generalization performance, we first conducted ablation studies on two key design choices: the number of convolutional layers and the number of channels in hidden layers. Intuitively, only one or two layers may be insufficient for simulating non-linear transformations across domains in real world, while a too-large number of layers may lead to unrealistically aggressive augmentations that deviate from reality. The effect of numbers of channels in hidden layers is difficult to postulate, due to the non-linearity of GIN.

We show quantitative results in Table III by varying the number of layers, and the number of channels in hidden layers from the default setting (4 layers and 2 channels). These experiments were conducted under the abdominal segmentation scenario, with IPA turned off for the ease of analysis.

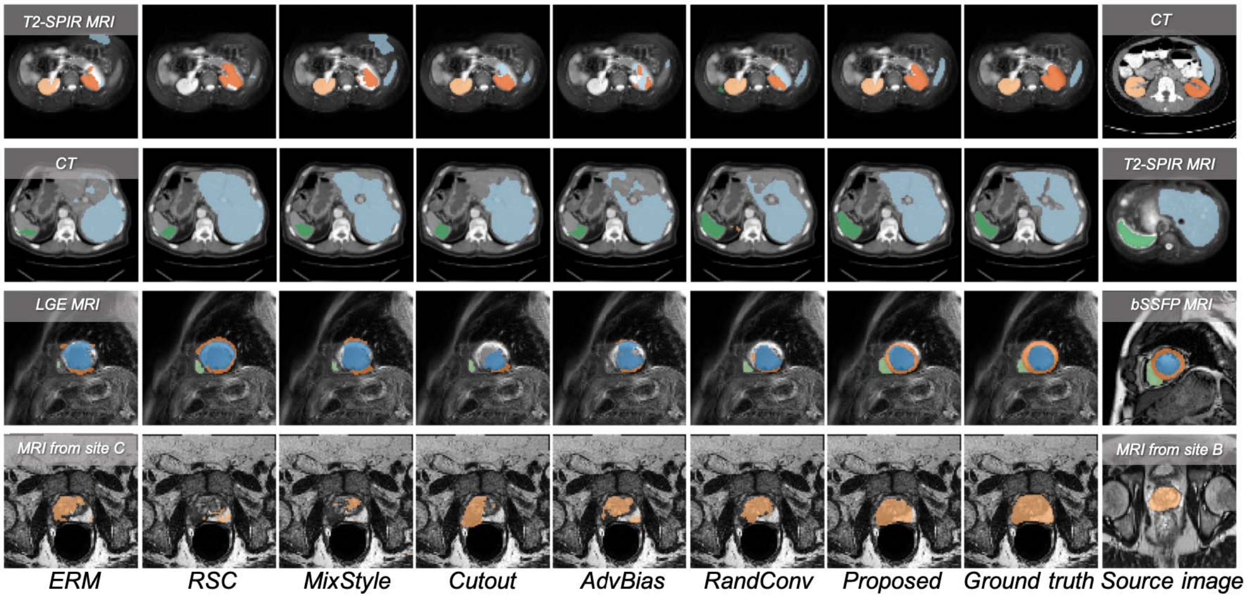
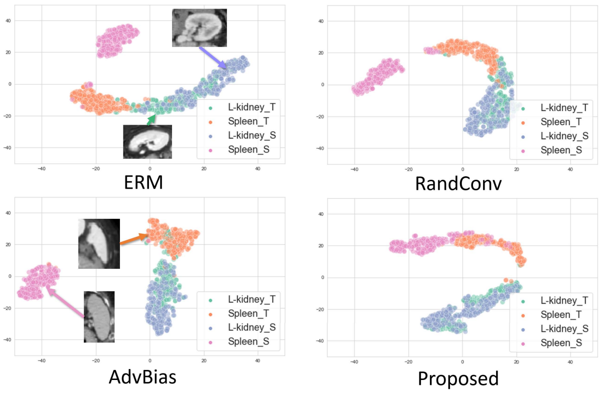


Fig. 6. Qualitative results on cross-domain segmentation under three scenarios: cross-modality segmentation for abdominal CT and MRI (the first and the second row), cross-sequence segmentation for cardiac MRI (the third row) and cross-site segmentation for prostate MRI (the fourth row). Examples of source domain (training dataset) images are shown in the rightmost column.



ERM

RandConv

AdvBias

Proposed

Fig. 7. t-SNE visualizations of the last hidden layer's features of segmentation networks, trained using different domain generalization techniques, for the Abdominal CT-MR segmentation scenario. Suffixes _S or _T denote the source (CT) or the target domain (MRI).

TABLE IV

Effects of Different Mechanisms for Incorporating Original Images in Training Batches, Evaluated on Abdominal CT to MRI Segmentation. Upper: The Effect of Different Interpolation Mechanisms Between Original and Network-Augmented Images. Lower: The Effect of Adding Original Images to the Training Batches

GIN configuration	Average Dice Score
No interp. ($\alpha = 1$)	5.41
Interp. w $\alpha \sim \mathcal{U}(0, 1)$ (default, proposed)	86.04
Interp. w $\alpha \sim \mathcal{B}eta(0.2, 0.2)$	85.84
Interp. w $\alpha \sim \mathcal{B}eta(0.4, 0.4)$	85.69
Interp. w $\alpha \sim \mathcal{B}eta(2.0, 2.0)$	85.61
Default (proposed)	86.04
Default w. 25% orig. image	85.87
Default w. 50% orig. image	84.20
Default w. 75% orig. image	80.05

The left column of Table III agrees with our intuition regarding the number of convolutional layers.

We also investigated the importance of the interpolation mechanism between the original image and the network-augmented image in GIN, as described in Sec. III-B and Eq. 4. As discussed in Sec. III-B, interpolating with original images allows to preserve essential semantic information in the augmented training samples, avoiding to generate too many over-aggressively augmented examples that may break the training process. Specifically, we removed the interpolation part, or varied the distribution of the interpolation coefficient α in Eq. 4 from a uniform distribution $\mathcal{U}(0, 1)$ to Beta distributions, as suggested in [44]. As shown in the upper part of Table IV, without interpolating with original images, the method would fail, while replacing uniform distribution to Beta distributions does not lead to significant improvement. We also replace $\{25\%, 50\%, 75\%\}$ of our default interpolated training images with original images. As shown in the lower part in Table IV, once the default interpolation mechanism has been applied, adding extra original images would not yield significant extra benefit.

2) Configurations of IPA: To examine the benefit of interventional pseudo-correlation augmentation, we performed ablation studies on IPA. These results are summarized in Table V. By comparing the results of *GIN-only* (the second row) with those of *GIN + IPA* (the last row) we can observe the benefit of mitigating the shifted-correlation effect using IPA, especially for the cardiac and the prostate settings. The benefit of introducing IPA has been further verified by the fact that simply mixing-up two GIN-augmented images in

TABLE V
ABLATION STUDY ON IPA. DICE SCORE IS USED AS THE EVALUATION METRIC. THE HIGHEST SCORES ARE IN RED, THE SECOND-HIGHEST SCORES ARE IN BLUE

Method	Abdominal Cross-modality		Cardiac Cross-sequence	Prostate Cross-site					
	CT to MRI	MRI to CT	bSSFP to LGE	A to Rest	B to Rest	C to Rest	D to Rest	E to Rest	F to Rest
ERM	77.31	67.59	75.99	71.81	65.56	43.98	71.97	48.39	37.82
GIN-only	86.04	79.40	83.83	79.30	62.28	47.38	67.71	64.84	69.11
GIN + Mixup (scalar) [44]	85.87	79.30	84.01	79.40	64.20	49.27	71.08	68.32	64.52
GIN + IPA (superpixel-based)	85.09	78.99	85.97	79.96	63.97	51.70	68.30	63.47	71.81
GIN + IPA (proposed)	$86.31^{\dagger}$	$79.62^{\dagger}$	85.01^{*}	82.14	67.21	59.11	73.16	67.38	73.23

Single-sided Wilcoxon singed-rank tests are performed when the performance gains of GIN + IPA (proposed) are below 2% in Dice score, compared with those of GIN-only.
†:p-value > 0.05 *:p-value $\ll 1 \times 10^{-4}$.

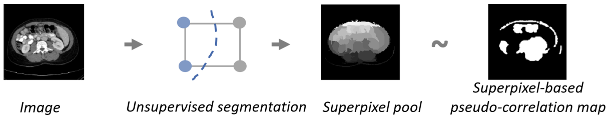


Image
Unsupervised segmentation
Superpixel pool
Superpixel-based pseudo-correlation map

Fig. 8. An alternative configuration of pseudo-correlation maps: randomly sampled superpixels.

TABLE VI
QUANTITATIVE RESULTS OF THE MODEL ENSEMBLING MECHANISM FOR RETAINING SOURCE-DOMAIN PERFORMANCE, EVALUATED ON ABDOMINAL AND CARDIAC IMAGES

Model	Abdominal				Cardiac	
	CT to MRI		MRI to CT		bSSFP to LGE	
	Source	Target	Source	Target	Source	Target
ERM	91.76	73.86	89.79	63.17	91.91	73.04
Proposed	89.81	86.31	88.70	79.62	91.86	85.01
Ensemble	92.07	86.38	90.28	79.64	92.37	84.58
Improve. over proposed	+2.25	+0.07	+1.58	+0.02	+0.51	-0.43

using a random global scalar coefficient [44], does not lead to comparable performance gains, as reflected by the third row of Table V.

As a further exploration, we also examined an alternative design of pseudo-correlation maps: superpixels that are randomly sampled at each iteration [73], as depicted in Fig. 8. We present its results in the last row of Table V. We postulate that the sub-optimal performance of the superpixel-based maps is due to the fact that superpixels often coincide with real ground-truth masks $\mathbf{y}$'s, making $\mathbf{x}_f$'s unnaturally stand out and thus become shortcuts for the network.

3) Domain Invariance Condition: The domain invariance condition proposed by [55], which is instantiated as the Kullback-Leibler divergence term in Eq. 3, provides an additional supervision signal for enforcing domain invariance in together with the segmentation loss on augmented images. To examine its utility we performed ablation study by dropping this divergence term, for the abdominal CT to MRI setting. We observe that without this term, the average Dice score drops from 86.31 to 85.67.

E. Performance in the Source Domain

Although the focus of our work is to improve performance in unseen target domains, we also investigated the performance in source domains. Intuitively, suppressing domain-specific features might hurt the source-domain performance. As shown in the first two rows (*ERM* and *Proposed*) in Table VI and those in Table VII, the proposed method experiences slight performance downgrades in some of the source domains, compared with models trained by the standard ERM.

TABLE VII
QUANTITATIVE RESULTS OF THE MODEL ENSEMBLING MECHANISM FOR RETAINING SOURCE-DOMAIN PERFORMANCE, EVALUATED ON PROSTATE IMAGES

Model	Prostate					
	A to Rest		B to Rest		C to Rest	
	Source	Targets	Source	Targets	Source	Targets
ERM	91.25	70.77	86.51	56.54	84.07	42.56
Proposed	91.28	82.14	86.13	67.21	85.42	59.11
Ensemble	91.63	82.34	87.91	67.57	85.60	59.15
Improve. over proposed	+0.35	+0.20	+1.77	+0.35	+0.18	+0.05

Model	D to Rest		E to Rest		F to Rest	
	Source	Targets	Source	Targets	Source	Targets
ERM	86.22	64.36	84.03	44.16	85.29	36.38
Proposed	85.14	73.16	76.62	67.38	85.44	73.23
Ensemble	84.77	73.65	78.97	67.54	85.82	73.11
Improve. over proposed	-0.37	+0.15	+2.35	+0.15	+0.39	-0.11

In the case that the domain of a test image is known, to retain the source-domain performance, we suggest to simply ensemble the predictions of our domain-generalized model (noted as $f^g(\cdot)$) and those of a standard model trained by ERM[6] (noted as $f^e(\cdot)$). This ensembling mechanism is based on the fact that predictions of both models may reasonably agree with each other when confronted with a source-domain image, while they may not agree when confronted with an out-of-domain image: The standard ERM model may fail while the proposed domain-generalized model would remain less affected. Specifically, the ensembling process is as follows: if two predictions overlap well (*e.g.* has a Dice similarity of $> 95\%$), we take the prediction by $f^e(\cdot)$ as the final prediction; if the overlapping falls between 85%∼95%, we take the average of two predictions; otherwise, we trust the output of $f^g(\cdot)$, as a large disagreement between two predictions implies an out-of-domain image, on which ERM model may fail.

As shown in the last two rows in Table VI-VII, the ensembling mechanism has compensated for the performance drop in the source domain caused by the suppression on domain-specific features.

V. DISCUSSION AND CONCLUSION

Domain robustness has been a challenge for deep learning based medical image computing for a long time. In this work, we propose a causality-inspired data augmentation approach

[6] Unlike in Table II, here model selection for ERM is based on the best performance on *source* domain.

for single-source domain generalization. From a methodological perspective, while previous multi-source domain generalization (MDG) and unsupervised domain adaptation (UDA) methods are *top-down* solutions that learn *a priori* knowledge of out-of-domain data (assumed to be available), our data augmentation is a *bottom-up* approach based on the causal mechanism of acquisition shifts. Although challenging, pursing causalities and designing bottom-up methods encourage further theoretical investigations on domain shift, which in turn facilitates more principled techniques for robust learning. From a practical perspective, unlike UDA or MDG, our method does not require target-domain data or multi-source data to be available during training. Also, compared with UDA, our method is easier to deploy in real world: it does not require fine-tuning on the target domain (which more or less relies on expertise). Compared to peer single-source generalization techniques, our approach demonstrates consistently superior performances in our experiments. Meanwhile, we note that single-source domain generalization and UDA are not mutually exclusive: a well-generalizable source model can facilitate adaption process, by providing a well-reduced domain gap for UDA to start from.

In terms of clinical value, domain generalization is critical for the success of deep models in clinical deployment: training one standalone model for each possible domain is impractical, and cross-institute data sharing is often prohibited. Instead, solving the fundamental problem of model generalizability would significantly expand the applications of deep models, by reducing the reliance on domain-specific datasets. It also boosts the trustworthy of deep models on safety-critical medical imaging applications. This is achieved by reducing the risk of erroneous predictions caused by out-of-domain data.

In the current approach several limitations remain: First, although consistent performance gains are shown in all three scenarios, some crucial hyper-parameters like the number of layers in GIN and the configurations of IPA still require empirical choices. A more elegant augmentation technique that requires less empirical choices is desirable. In addition, as discussed in Sec. IV-E, as domain-specific information is suppressed during training, our method experiences slight performance downgrades in some of source domains, when the ensembling is turned off. A potential solution to this issue may reside in the network architecture side [29]. Also, sampling weights for random networks in GIN for every training image and every iteration inevitably incurs computational overhead during training. For the abdominal CT to MRI experiment, on a single Nvidia RTX 2080Ti GPU, using the setting described in Sec. IV-B, the proposed method takes $\sim$19 seconds per epoch $\times$ 2k epochs ($\sim$10.6 hours in total), with a GPU memory consumption of 10.6 GB. This is in comparison with ERM that takes $\sim$9 seconds per epoch and $\sim$ 4.3 GB of GPU memory. Optimization can be done by sharing the same random network in GIN within a batch or across several batches. As our method is based on data augmentation during training, it does not incur additional test-time computational overhead. Although the ensemble strategy in Sec. IV-E approximately doubles the computation time in testing, it is still in most cases acceptable, as the inference process of segmentation models is relatively inexpensive compared with that of training.

Starting from the current methodology, several potential extensions arise: As our method can efficiently produce huge amount of domain-shifted images, it is natural to combine the proposed method with multi-source domain generalization techniques [24], [30]. In addition, as our work focuses on image appearance, it is interesting to design methods targeting at domain shifts in terms of anatomical shapes.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[2] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[3] F. Isensee et al., "nnU-Net: Self-adapting framework for U-Net-based medical image segmentation," 2018, *arXiv:1809.10486*.

[4] L. Zhang et al., "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE Trans. Med. Imag.*, vol. 39, no. 7, pp. 2531–2540, Jul. 2020.

[5] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[6] M. Antonelli et al., "The medical segmentation decathlon," 2021, *arXiv:2106.05735*.

[7] K. Kamnitsas et al., "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *Proc. Int. Conf. Inf. Process. Med. Imag. (IPMI)*. Cham, Switzerland: Springer, 2017, pp. 597–609.

[8] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng, "Unsupervised cross-modality domain adaptation of ConvNets for biomedical image segmentations with adversarial loss," 2018, *arXiv:1804.10916*.

[9] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Mach. Learn.*, vol. 79, nos. 1–2, pp. 151–175, May 2010.

[10] B. Glocker, R. Robinson, D. C. Castro, Q. Dou, and E. Konukoglu, "Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects," 2019, *arXiv:1910.04597*.

[11] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Commun.*, vol. 11, no. 1, pp. 1–10, Dec. 2020.

[12] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002683.

[13] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[14] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[15] D. Zhang, H. Zhang, J. Tang, X. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," 2020, *arXiv:2009.12547*.

[16] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 1180–1189.

[17] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 10–18.

[18] M. Ilse, J. M. Tomczak, and P. Forré, "Selecting data augmentation for simulating interventions," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2021, pp. 4555–4562.

[19] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," 2014, *arXiv:1412.3474*.

[20] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 289–305.

[21] C. Ouyang, K. Kamnitsas, C. Biffi, J. Duan, and D. Rueckert, "Data efficient unsupervised domain adaptation for cross-modality image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2019, pp. 669–677.

[22] N. Karani, E. Erdil, K. Chaitanya, and E. Konukoglu, "Test-time adaptable neural networks for robust medical image segmentation," *Med. Image Anal.*, vol. 68, Feb. 2021, Art. no. 101907.

[23] M. Bateson, H. Kervadec, J. Dolz, H. Lombaert, and I. Ben Ayed, "Source-free domain adaptation for image segmentation," 2021, *arXiv:2108.03152*.

[24] Q. Dou, D. C. De Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 6450–6461.

[25] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. Hospedales, "Episodic training for domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1446–1455.

[26] K. Zhou, Y. Yang, Y. Qiao, and T. Xiang, "Domain generalization with MixStyle," 2021, *arXiv:2104.02008*.

[27] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multimodal MR synthesis via modality-invariant latent representation," *IEEE Trans. Med. Imag.*, vol. 37, no. 3, pp. 803–814, Mar. 2017.

[28] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2229–2238.

[29] P. Chattopadhyay, Y. Balaji, and J. Hoffman, "Learning to balance specificity and invariance for in and out of domain generalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 301–318.

[30] Q. Liu, Q. Dou, and P.-A. Heng, "Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2020, pp. 475–485.

[31] X. Liu, S. Thermos, A. O'Neil, and S. A. Tsaftaris, "Semi-supervised meta-learning with disentanglement for domain-generalised medical image segmentation," 2021, *arXiv:2106.13292*.

[32] R. Gu, J. Zhang, R. Huang, W. Lei, G. Wang, and S. Zhang, "Domain composition and attention for unseen-domain generalizable medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2021, pp. 241–250.

[33] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 124–140.

[34] Z. Liu et al., "Generalize ultrasound image segmentation via instant and plug & play style transfer," in *Proc. Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 419–423.

[35] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1501–1510.

[36] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.

[37] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, *arXiv:1708.04552*.

[38] R. Volpi, H. Namkoong, O. Sener, J. Duchi, V. Murino, and S. Savarese, "Generalizing to unseen domains via adversarial data augmentation," 2018, *arXiv:1805.12018*.

[39] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 834–843.

[40] Z. Xu, D. Liu, J. Yang, C. Raffel, and M. Niethammer, "Robust and generalizable visual representation learning via random convolutions," 2020, *arXiv:2007.13003*.

[41] R. D. Hjelm et al., "Learning deep representations by mutual information estimation and maximization," 2018, *arXiv:1808.06670*.

[42] P. Khosla et al., "Supervised contrastive learning," 2020, *arXiv:2004.11362*.

[43] S. Wu, H. Zhang, G. Valiant, and C. Ré, "On the generalization effects of linear transformations in data augmentation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 10410–10420.

[44] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2017, *arXiv:1710.09412*.

[45] V. Verma et al., "Manifold mixup: Better representations by interpolating hidden states," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6438–6447.

[46] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.

[47] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12556–12565.

[48] L. Li et al., "Progressive domain expansion network for single domain generalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 224–233.

[49] B. Billot, D. Greve, K. Van Leemput, B. Fischl, J. E. Iglesias, and A. V. Dalca, "A learning strategy for contrast-agnostic MRI segmentation," 2020, *arXiv:2003.01995*.

[50] C. Chen et al., "Realistic adversarial data augmentation for mr image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent. (MICCAI)*. Cham, Switzerland: Springer, 2020, pp. 667–677.

[51] C. Chen, K. Hammernik, C. Ouyang, C. Qin, W. Bai, and D. Rueckert, "Cooperative training and latent space data augmentation for robust medical image segmentation," 2021, *arXiv:2107.01079*.

[52] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019, *arXiv:1907.02893*.

[53] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 7313–7324.

[54] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik, "A causal view of compositional zero-shot recognition," 2020, *arXiv:2006.14610*.

[55] J. Mitrovic, B. McWilliams, J. Walker, L. Buesing, and C. Blundell, "Representation learning via invariant causal mechanisms," 2020, *arXiv:2010.07922*.

[56] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[57] R. Giryes, G. Sapiro, and A. M. Bronstein, "Deep neural networks with random Gaussian weights: A universal classification strategy?" *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3444–3457, Jul. 2016.

[58] L. Hoyer, M. Munoz, P. Katiyar, A. Khoreva, and V. Fischer, "Grid saliency for context explanations of semantic segmentation," 2019, *arXiv:1907.13054*.

[59] C. Burns and J. Steinhardt, "Limitations of post-hoc feature alignment for robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2525–2533.

[60] O. Oktay et al., "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.

[61] C. De Boor and C. De Boor, *A Practical Guide to Splines*, vol. 27. Cham, Switzerland: Springer, 1978.

[62] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," 2019, *arXiv:1912.02781*.

[63] R. Sandkühler, C. Jud, S. Andermatt, and P. C. Cattin, "AirLab: Autograd image registration laboratory," 2018, *arXiv:1806.09907*.

[64] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge," in *Proc. Multi-Atlas Labeling Beyond Cranial Vault Workshop Challenge (MICCAI)*, vol. 5, 2015, p. 12.

[65] A. E. Kavur et al., "CHAOS Challenge–combined (CT-MR) healthy abdominal organ segmentation," *Med. Image Anal.*, vol. 69, Apr. 2021, Art. no. 101950.

[66] X. Zhuang et al., "Cardiac segmentation on late gadolinium enhancement MRI: A benchmark study from multi-sequence cardiac MR segmentation challenge," 2020, *arXiv:2006.12434*.

[67] N. Bloch et al., "NCI-ISBI 2013 challenge: Automated segmentation of prostate structures," *Cancer Imag. Arch.*, vol. 370, no. 6, pp. 1–5, 2015.

[68] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau, "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review," *Comput. Biol. Med.*, vol. 60, pp. 8–31, May 2015.

[69] G. Litjens et al., "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge," *Med. Image Anal.*, vol. 18, no. 2, pp. 359–373, 2014.

[70] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert, "Self-supervision with superpixels: Training few-shot medical image segmentation without annotation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2020, pp. 762–780.

[71] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.

[72] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[73] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vis.*, vol. 59, no. 2, pp. 167–181, Sep. 2004.