Causality-inspired Single-source Domain Generalization for Medical Image Segmentation

Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai and Daniel Rueckert

Abstract—**Deep learning models usually suffer from domain shift issues, 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 is 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. Under 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 tasks: cross-modality (CT-MRI) abdominal image segmentation, cross-sequence (bSSFP-LGE) cardiac MRI segmentation, and cross-center 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]–[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 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 (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*.

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

Shifted-correlation effect: Due to a *confounder* (*i.e.* a "third" variable that spuriously correlates two variable-of-interests) [12], [13], objects in the background might be correlated but **not** *causally* related to the object-of-interests [14]. The network might take these objects in the background as clues for recognizing the object-of-interests [14]. For example, in [15], [16], a model that recognizes `pneumonia` in X-ray images is actually looking at the `hospital mark` 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 object-of-interests. In the above example, the model fails on real-world images where `hospital mark` do not correlate to `pneumonia`.

To mitigate domain shift, previous attempts include *unsupervised domain adaptation* (UDA) [17] and *multi-source domain generalization* (MDG) approaches [18]. 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 towards 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 object-of-interests during training.

Learning causal features and removing confoundings usually require *intervention*: *fixing* the variable-of-interest while incorporating other variables in a fair way [13]. 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 [19] that incorporates possible acquisition shifts via simulation.

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 object-of-interests, in terms of their appearances. We then design a practical method for simulating and independently sampling the appearances of potentially confounded objects durining 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 segmentation networks. 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-center 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 [17], [20], or self-training [21]. Multi-source domain generalization (MDG) usually learns domain-invariant features in an one-off manner from multiple source domains. Recent techniques include meta-learning [22], [23], style transfer [24], [25], transfer learning [26], dynamic networks [27] and so on. In medical imaging, recent methods [28], [29] based on meta-learning [30] have achieved promising results. However, for both UDA and MDG, target domain data or multi-source data is often unavailable due to privacy and/or cost concerns in medical settings.

Single-source domain generalization requires training data from one domain only. Recent works such as [31] propose to remove features that cause the largest loss gradients, which are believed to be domain-dependent. Liu et al. [32] propose to unify statistics of image features, which control image styles [33], 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 [34]–[37]. We review these stream of methods later with more details. Some most recent works such as [38] employ multiple techniques: data augmentation [39], adversarial training [40] and contrastive learning [41].

B. Data augmentation for domain robustness

Theoretical analysis [42] suggests that data augmentation improves generalization by enlarging the span of the data and by regularizing decision boundaries. In practice, Cutout [35] strengthens robustness against feature missing caused by domain shift, by partially occluding training images. Mixup [36], [43] regularizes decision boundaries by interpolating among training samples. RandConv [39] 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 [39]. However, we show in our experiments that the linear filtering in RandConv [39] is oversimplified for accounting for domain gaps that occur in real-world settings. Adversarial data augmentation generates image perturbations that easily flip predictions of classifiers [37], [44]–[46].

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. [47] 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 from segmentation labels to images, which is usually unavailable in most of medical imaging applications. Furthermore, these generative models are often oversimplified. AdvBias [48] 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-center MRI segmentation [49].

C. Leveraging causality for robust deep learning

As discussed in Sec. I, learning causalities and mitigating confoundings usually require *causal intervention* [13]. 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` [50], 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 like grassland and city. By this mean the model would learns that it is the camel rather than the desert that *cause* a `camel` label. Causal relationships are usually modeled using structural causal models (SCM) [12], [13]. The *fixing* operation is usually noted as $do(\cdot)$ [13]. 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* [13].

Causal ideas have been used for discovering image features that are semantically essential and robust [50]–[53]. Invariant risk minimization [50] learns causal image representations by enforcing these representations to be Bayesian optimal in all environments. Mahajan et al. [51] improve domain robustness using contrastive losses. Atzmon et al. [52] 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 [19], [53]. [19] proves that *post-hoc* data augmentation theoretically commutes with "physical" intervention. Mitrovic et al. [53] derive a practical loss function for causality-based domain generalization. Different from [53], we focus on the unanswered practical problem of designing an augmentation model 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. [14], as both works study the adverse effect of confoundings among objects on image segmentation. While Zhang et al. [14] 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 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. 1-B, GIN transfers training images to have diverse appearances while keeping 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 the object-of-interests, in independent and diverse manners. This is implemented as spatially-variable blending 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-of-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 object-of-interests. We then describe how IPA removes confoundings. Finally, we summarize the overall training process.

A. Problem formulation

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 [52], [53], we model the data generation process and the (ideally, domain-invariant) segmentation process using the causal model shown in Fig. 1-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 object-of-interests. 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 an input and predicts Y, by implicitly estimating S.

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

Causal intervention for domain robustness: According to [53], our argument that S *to be 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 raised by letting images to be generated from a specific acquisition process $A = a_i$ [13], [53], 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.1-A, we use a segmentation network $f_\phi(\cdot)$ parameterized by ϕ to predict Y. To make
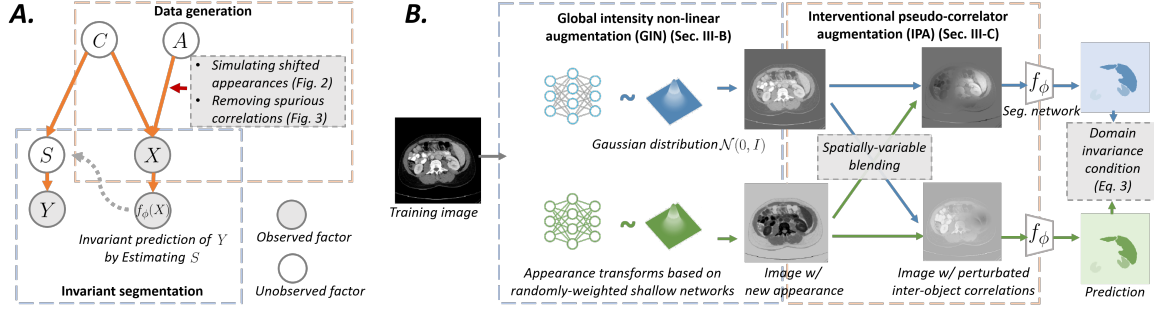
Fig. 1. **A. Data generation process of medical images:** An image X is modeled as the effect of two independent factors: 1) a domain-independent content factor C which represents the geometric shapes of the patient anatomy; 2) and a domain-dependent acquisition factor A which controls image intensities and textures (appearances). **Mechanism of a domain-invariant segmentation network:** We assume that 1) there exists an ideal domain-invariant representation S that contains shape information and is in the form of feature maps. S is caused by C; 2) the pixel-wise ground-truth Y can be derived from S. Our goal is to learn a network $f_\phi(\cdot)$ that predicts Y by implicitly estimating S. To train $f_\phi(\cdot)$, our approach simulates different acquisition processes A's and hence X's. These simulated images encourage the network to distill S from them.
B. Workflow of the proposed data augmentation approach: Our approach simulates different possible acquisition processes. GIN first transforms input images to have new appearances (intensities and textures), using randomly-weighted shallow networks. IPA then blends two GIN-augmented versions of a same image in a spatially-variable manner. This blending operation perturbs/randomizes spurious correlations among objects in an image, in terms of their appearances. Augmenting data with IPA allows the network to be immune to the *shifted-correlation effect* (*i.e.* the breaking of decision rules that are built on spurious correlations in the source domain, as described in Sec. I). During training, the segmentation network $f_\phi(\cdot)$ is encouraged to make accurate and consistent predictions, unaffected by these simulated domain shifts.

$f_\phi(\cdot)$ domain invariant, we 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. [19] 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 $\mathcal{T}_i(\cdot)$ being able to transform the image X to be like from a_i [1]. We therefore have:

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

Combining Eq. 1 and 2 leads to a practical domain invariance condition: minimizing the difference between distributions raised by different photometric transformations $\mathcal{T}_i(\cdot)$ and $\mathcal{T}_j(\cdot)$ [53]. By combining this domain invariance condition and the image segmentation loss, we can now derive our loss function (inspired by [53]). For each iteration, we have

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

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 simulates 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 differences between two distributions; λ_{div} is a weighting coefficient. Similar divergence terms have also been

used for semi-supervised learning [44], [54], 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 is 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.
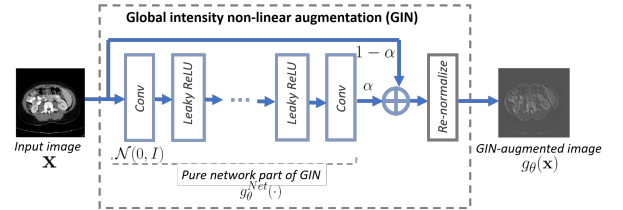
B. Global intensity non-linear augmentation

Fig. 2. **Illustration of the proposed global intensity non-linear augmentation (GIN) module.** It transforms image appearances using shallow convolutional networks with their weights 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.

Design-of-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-invariant 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, $g(\cdot) : \mathbb{R}^{C_h \times H \times W} \to \mathbb{R}^{C_h \times H \times W}$, 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. 2, transformations sampled from GIN are instantiated as shallow multi-layer convolutional networks

[1]Arguably this transformation can be interpreted as generating *counterfactual* examples [11], [13]: given an observed image from a certain acquisition process, asking how this image would look like, had it been generated from another imaging process. We do not label our data augmentation as counterfactual since the aim of our approach is not to synthesize the appearance of a *specific* domain.

$g_\theta(\cdot)$'s. 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 $g_\theta(\cdot)$'s are sampled, yielding a variety of transformation functions. Inspired by [39], we perform linear interpolation between the original and the output of the random network. In the end, as depicted in Fig. 2, 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 pure network part of the transformations $g_\theta(\cdot)$ as $g_\theta^{Net}(\cdot)$ (see Fig. 2), 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:

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

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. **Advantages:** We highlight major advantages of the above configurations: 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 be integrated into adversarial augmentation frameworks [37], [44], [45], [48] to improved data efficiency.

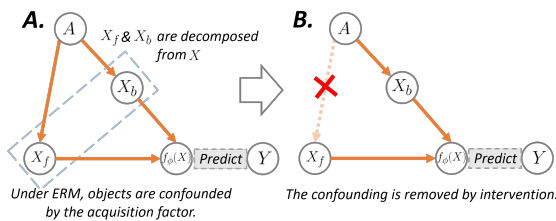### C. Interventional pseudo-correlation augmentation

Fig. 3. **A. Causal graph illustrating spurious correlations between the object-of-interest X_f and a potentially correlated object X_b**, where X_f and X_b are patches decomposed from a same image X. Spurious correlations between X_f and X_b are due to the acquisition factor A which naturally confounds them. In the source domain, the network $f_\phi(\cdot)$ may build domain-specific decision rules on these spurious correlations. In a shifted domain, these spurious correlations may break, and they are therefore detrimental to out-of-domain robustness of the network. This causal graph is a close-up of the data generation process in Fig. 1-A. The content factor C in Fig. 1-A is assumed to be unchanged across domains and they are therefore omitted here. **B. Removing the confounding caused by** A: This is achieved by the causal intervention $do(X_f = \mathbf{x}_f)$, which removes $A \to X_f$ in the causal graph. This intervention is conducted by independently sampling possible appearances of potentially correlated objects X_b's.

Confounded objects in the background affect segmentation: Recall in Sec. I, spurious correlations (due to confoundings) between objects in the background and the object-of-

interests in the source domain might be taken by a segmentation network as domain-specific clues for making predictions[2]. These decision rules may break in the target domain, leading to performance downgrade. 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 the object-of-interests.

From the perspective of network architectures, objects in the background often affect predictions of the object-of-interests by the following ways: 1) Background features can affect global feature statistics at normalization layers [56], since feature statistics are usually calculated across all spatial locations. 2) The large receptive fields often make the pixels of the object-of-interests and those of the neighboring background to be inevitably perceived and processed together [57].

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[3]. 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. 1-A using X_f and X_b, and redraw that in Fig. 3-A. We can see:

1) $X_f \to f_\phi(X) \leftarrow X_b$: 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) $f_\phi(X) \leftarrow X_b \leftarrow A \to X_f$: More importantly, the confounding effect of A that correlates X_b and X_f is revealed in the path $X_b \leftarrow A \to X_f$. 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 remain unchanged across domains, we ignore the confoundings caused by C, and omit C in Fig. 3.

Removing confounding by intervention: We propose to mitigate the shifted correlation effect during the training stage, by removing the confounding $X_b \leftarrow A \to X_f$ using the intervention $do(X_f = \mathbf{x}_f)$ [13]. This operation resamples the appearances of correlated objects X_b's, independent of X_f.

---

[2]In our preliminary experiment, we verified the existence of decision rules that are based on the background, in medical image segmentation: We distorted the backgrounds of images by randomly swapping patches of the background (therefore the global image statistics would remain unchanged.). We then tested a segmentation network with these background-distorted images, and have observed substantial performance downgrade compared with results on the original undistorted images. This phenomenon has also been verified in general computer vision and these spurious correlations are sometimes termed as *context bias* [14], [55].

[3]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.
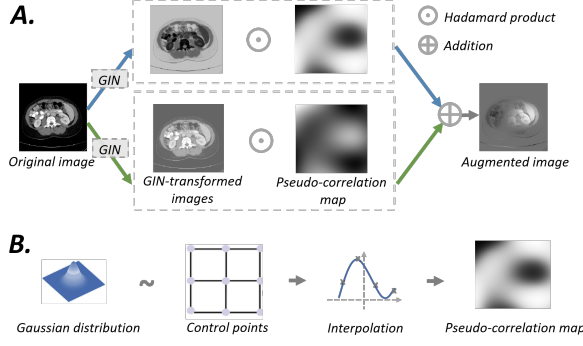
Fig. 4. **A. Implementation of IPA:** In each iteration, a new pseudo-correlation map is generated, and it is used as pixel-wise coefficients for blending two GIN-transformed images. This is equivalent to assigning different appearance transformations to different pixels/patches, according to their corresponding 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.

This intervention in effect removes $A \to X_f$. Formally, we are learning the interventional distribution $p(Y|do(X_f = \mathbf{x}_f))$ based on the intervened causal diagram Fig. 3-B:

$$p(Y|do(X_f = \mathbf{x}_f)) = \sum_{\mathbf{x}_b} p(Y|\mathbf{x}_f, \mathbf{x}_b) p(\mathbf{x}_b)$$
$$= \sum_{\mathbf{x}_b} \sum_{a} p(Y|\mathbf{x}_f, \mathbf{x}_b) p(\mathbf{x}_b|a) p(a). \quad (5)$$

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 objects 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 $\mathbf{x}_f$. Also, this intervention *cannot* be realized by GIN alone: GIN's transformation functions are spatially-invariant[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 [58]. 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

[4]If two objects share the same appearance, their appearances would remain the same after being transformed by GIN.

$\mathbf{y}$, 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. 4-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:

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

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. 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 [59] which is designed for image classification. Different from AugMix, IPA necessities strict spatial correspondence between pixels and labels to ensure accuracy of this pixel-wise prediction. Also, the blending coefficients of IPA are intentionally made spatially variable to simulate our causal intervention.
Pseudo-correlation maps: As shown in Fig. 4-B, we configure pseudo-correlation maps as a field of continuous randomly-valued scalars with low spatial frequency. They are interpolated from a lattice of randomly-valued control points, using cubic B-spline kernels [58], based on the efficient implementation from [60]. 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 [39]. 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.

IV. EXPERIMENTS

A. Datasets and evaluation protocols

The proposed approach is evaluated in three cross-domain settings: 1) cross-modality abdominal segmentation from CT to T2-SPIR MRI (Abdominal CT-MR), 2) cross-sequence cardiac segmentation from bSSFP MRI to LGE MRI (Cardiac bSSFP-LGE), and 3) prostate segmentation on MRI across six centers (Cross-center Prostate). Details of the datasets

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 CT-MRI	Liver, L-kidney, R-kidney, Spleen	Axial	Source	Computed tomography (CT)	30	[61]
			Target	T2 spectral presaturation with inversion recovery (SPIR) MRI	20	[62]
Cardiac bSSFP-LGE	L-ventricle, Myocardium, R-ventricle	Short-axis	Source	Balanced steady-state free precession (bSSFP) MRI	45	[63]
			Target	Late gadolinium enhanced (LGE) MRI	45	
Prostate Cross-center	Prostate	Axial	1 Source	Prostate MRI from 6 centers	30, 30, 19	[28]
			5 Targets		13, 12, 12	[64]–[66]

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.

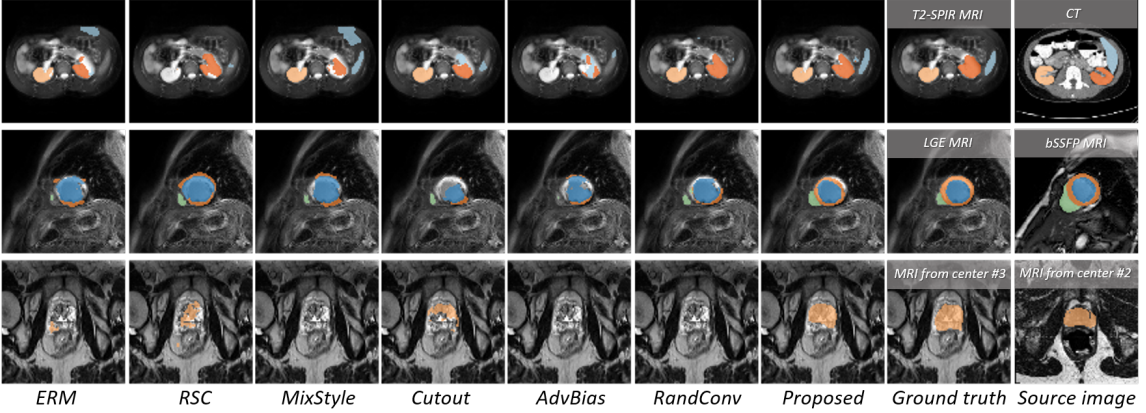| Method	Abdominal CT-MRI					Cardiac bSSFP-LGE				Prostate Cross-center
	Liver	R-Kidney	L-Kidney	Spleen	Average	L-ventricle	Myocardium	R-ventricle	Average	Prostate
Upper bound	91.30	92.43	89.86	89.83	90.85	92.04	83.11	89.30	88.15	86.23
ERM	78.03	78.11	78.45	74.65	77.31	86.06	66.98	74.94	75.99	56.59
Cutout [35]	79.80	82.32	82.14	76.24	80.12	88.35	69.06	79.19	78.87	66.56
RSC [31]	76.40	75.79	76.60	67.56	74.09	87.06	69.77	75.69	77.51	60.78
MixStyle [24]	77.63	78.41	78.03	77.12	77.80	85.78	64.23	75.61	75.21	57.06
AdvBias [48]	78.54	81.70	80.69	79.76	80.17	88.23	70.29	80.32	79.62	61.98
RandConv [39]	73.63	79.69	85.89	83.43	80.66	89.88	75.60	85.70	83.73	57.74
Proposed	86.62	87.48	86.88	84.27	86.31	90.35	77.82	86.87	85.01	70.37

Fig. 5. Qualitative results on cross-domain segmentation under three scenarios: Abdominal CT-MRI (top row), Cardiac bSSFP-LGE (middle row) and Prostate Cross-center (bottom row). Examples of source domain (training dataset) images are shown in the rightmost column.

and the source-target splits are summarized in Table I. All 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 [-275, 125] [67] 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 [28]. For cross-center 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 [63].

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 [68]. For our proposed method, we trained the segmentation network using an Adam optimizer [69] 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

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.

1: **for** $t = 1 ... $ N **do**
2:     Sample an image-label pair $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)})$ from $\{(\mathbf{x}, \mathbf{y})\}$.
3:     Sample intensity/texture transformations $g_{\theta_1}^{(t)}(\cdot), g_{\theta_2}^{(t)}(\cdot)$ using GIN.
4:     Apply $g_{\theta_1}^{(t)}(\cdot), g_{\theta_2}^{(t)}(\cdot)$ as described in Eq. 4 to $\mathbf{x}^{(t)}$ and obtain $g_{\theta_1}^{(t)}(\mathbf{x}^{(t)}), g_{\theta_2}^{(t)}(\mathbf{x}^{(t)})$.
5:     Compute a pseudo-correlation map $\mathbf{b}^{(t)}$ by interpolating along randomly valued control points as in Fig. 4-B.
6:     Compute augmented images $\mathcal{T}_1^{(t)}(\mathbf{x}^{(t)}), \mathcal{T}_2^{(t)}(\mathbf{x}^{(t)})$ using IPA, as described in Eq. 6.
7:     Compute training loss $\mathcal{L}^{(t)}(\phi)$ using Eq. 3.
8:     Update parameters of the segmentation network $f_\phi(\cdot)$: $\phi^{(t+1)} := \phi^{(t)} - l_r^{(t)} \frac{\partial \mathcal{L}^{(t)}(\phi)}{\partial \phi}$.
9: **end for**

---

generalization methods. Among them Cutout [35] enforces the model to be robust to corruptions by deliberately removing patches from training images. RSC [31] defines features that lead to the largest gradients as non-robust features and removes them in training. MixStyle [24] synthesize novel domains by mixing instance-level feature statistics [33]. AdvBias [48] is designed for medical images. It augments images using adversarial perturbations. Closely related to our work is RandConv [39], 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 consistently outperforms peer methods. In particular, the performance gains of our method compared with the closely-related RandConv [39] 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. 5.

To visualize the feature spaces, in Fig. 6 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 conducted ablation studies on two key design-of-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
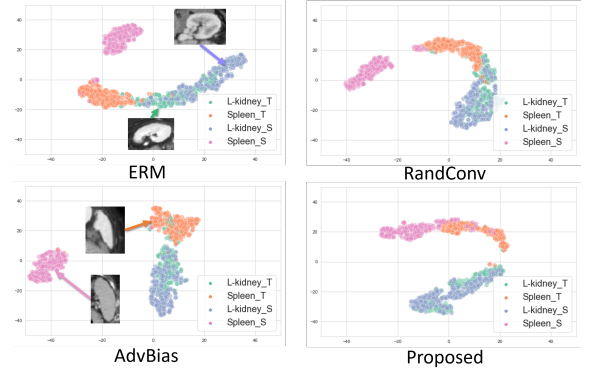

Fig. 6. t-SNE visualizations of the last hidden layer features of segmentation networks, trained using different domain generalization techniques, for the Abdominal CT-MR segmentation scenario. Suffixes _S or _T denote the source (CT) or the target domain (MRI).

TABLE III
EFFECTS OF NUMBER OF TOTAL 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 (reported)	86.04
4 (reported)	86.04	4	85.49
8	85.66	8	83.21
16	79.17	16	81.90

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

2) Configurations of IPA: To examine the effect of interventional pseudo-correlation augmentation, we performed ablation study by removing IPA from the proposed approach. The results in the first two rows of Table IV validate the benefit of mitigating the shifted-correlation effect using IPA, especially for the cardiac and the prostate settings, where p-values < 0.001 under Wilcoxon signed-rank tests.
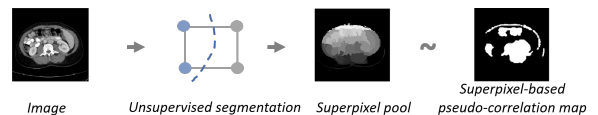

Fig. 7. An alternative configuration of pseudo-correlation maps: randomly sampled superpixels.

As a further exploration, we also examined an alternative design of pseudo-correlation maps: superpixels that are randomly sampled at each iteration [70], as depicted in Fig. 7. We present its results in the last row of Table IV. We conjecture the

TABLE IV

ABLATION STUDY ON INTERVENTIONAL PSEUDO-CORRELATION AUGMENTATION. THE HIGHEST SCORES ARE IN RED, THE SECOND-HIGHEST SCORES ARE IN BLUE.

Configuration	Abdominal CT-MRI					Cardiac bSSFP-LGE				Prostate Cross-center
	Liver	R-Kidney	L-Kidney	Spleen	Average	L-ventricle	Myocardium	R-ventricle	Average	Prostate
GIN-only	87.34	86.40	87.18	83.25	86.04	90.04	75.74	85.71	83.83	65.10
GIN + IPA (reported)	86.62	87.48	86.88	84.27	86.31	90.35	77.82	86.87	85.01	70.37
GIN + IPA (superpixel-based)	87.26	85.66	85.14	82.30	85.09	91.07	79.63	87.19	85.97	66.53

sub-optimal performance of the superpixel-based maps is due to the fact that superpixels often coincides with real ground-truth masks $\mathbf{y}$'s, making $\mathbf{x}_f$'s unnaturally stand out and thus become shortcuts for the network.

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 for single-source domain generalization. From a methodological perspective, while previous multi-source domain generalization (MDG) and unsupervised domain adapation (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.

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 domain-specific information is suppressed during training, our method experiences slight performance downgrades in source-domain testing sets on abdominal CT ($\sim$ -2.0/100) and prostate MRI's ($\sim$ -1.0/100), compared to the ERM counterparts. Potential solutions might reside in network architecture side, for example, a dynamic architecture which adaptively balances domain-specific and domain-invariant features.

Starting from the current methodology, several potential extensions arise: For example, 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 [22], [28]. 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 *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[2] D. Shen, G. Wu, and H.-I. Suk, "Deep learning in medical image analysis," *Annual review of biomedical engineering*, vol. 19, pp. 221–248, 2017.

[3] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert *et al.*, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *arXiv preprint arXiv:1809.10486*, 2018.

[4] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu *et al.*, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2531–2540, 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*. Springer, 2018, pp. 3–11.

[6] M. Antonelli, A. Reinke, S. Bakas, K. Farahani, B. A. Landman, G. Litjens, B. Menze, O. Ronneberger, R. M. Summers, B. van Ginneken *et al.*, "The medical segmentation decathlon," *arXiv preprint arXiv:2106.05735*, 2021.

[7] K. Kamnitsas, C. Baumgartner, C. Ledig, V. Newcombe, J. Simpson, A. Kane, D. Menon, A. Nori, A. Criminisi, D. Rueckert *et al.*, "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks," in *International conference on information processing in medical imaging*. 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," *arXiv preprint arXiv:1804.10916*, 2018.

[9] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1, pp. 151–175, 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," *arXiv preprint arXiv:1910.04597*, 2019.

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

[12] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.

[13] ——, *Causality*. Cambridge university press, 2009.

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

[15] 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 medicine*, vol. 15, no. 11, p. e1002683, 2018.

[16] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, 2020.

[17] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by back-propagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[18] K. Muandet, D. Balduzzi, and B. Schölkopf, "Domain generalization via invariant feature representation," in *International Conference on Machine Learning*. PMLR, 2013, pp. 10–18.

[19] M. Ilse, J. Tomczak, P. Forré *et al.*, "Selecting data augmentation for simulating interventions," 2020.

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

[21] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 289–305.

[22] Q. Dou, D. Coelho de Castro, K. Kamnitsas, and B. Glocker, "Domain generalization via model-agnostic learning of semantic features," *Advances in Neural Information Processing Systems*, vol. 32, pp. 6450–6461, 2019.

[23] D. Li, J. Zhang, Y. Yang, C. Liu, Y.-Z. Song, and T. M. Hospedales, "Episodic training for domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1446–1455.

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

[25] A. Chartsias, T. Joyce, M. V. Giuffrida, and S. A. Tsaftaris, "Multi-modal mr synthesis via modality-invariant latent representation," *IEEE transactions on medical imaging*, vol. 37, no. 3, pp. 803–814, 2017.

[26] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2229–2238.

[27] P. Chattopadhyay, Y. Balaji, and J. Hoffman, "Learning to balance specificity and invariance for in and out of domain generalization," in *European Conference on Computer Vision*. Springer, 2020, pp. 301–318.

[28] Q. Liu, Q. Dou, and P. A. Heng, "Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains," in *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2020.

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

[30] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.

[31] Z. Huang, H. Wang, E. P. Xing, and D. Huang, "Self-challenging improves cross-domain generalization," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 124–140.

[32] Z. Liu, X. Huang, X. Yang, R. Gao, R. Li, Y. Zhang, Y. Huang, G. Zhou, Y. Xiong, A. F. Frangi *et al.*, "Generalize ultrasound image segmentation via instant and plug & play style transfer," in *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, pp. 419–423.

[33] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

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

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

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

[38] Z. Wang, Y. Luo, R. Qiu, Z. Huang, and M. Baktashmotlagh, "Learning to diversify for single domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 834–843.

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

[40] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," *arXiv preprint arXiv:1808.06670*, 2018.

[41] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *arXiv preprint arXiv:2004.11362*, 2020.

[42] S. Wu, H. Zhang, G. Valiant, and C. Ré, "On the generalization effects of linear transformations in data augmentation," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 410–10 420.

[43] V. Verma, A. Lamb, C. Beckham, A. Najafi, I. Mitliagkas, D. Lopez-Paz, and Y. Bengio, "Manifold mixup: Better representations by interpolating hidden states," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6438–6447.

[44] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[45] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 556–12 565.

[46] L. Li, K. Gao, J. Cao, Z. Huang, Y. Weng, X. Mi, Z. Yu, X. Li, and B. Xia, "Progressive domain expansion network for single domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 224–233.

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

[48] C. Chen, C. Qin, H. Qiu, C. Ouyang, S. Wang, L. Chen, G. Tarroni, W. Bai, and D. Rueckert, "Realistic adversarial data augmentation for mr image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 667–677.

[49] 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," *arXiv preprint arXiv:2107.01079*, 2021.

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

[51] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7313–7324.

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

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

[54] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Advances in neural information processing systems*, vol. 29, pp. 1163–1171, 2016.

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

[56] C. Burns and J. Steinhardt, "Limitations of post-hoc feature alignment for robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2525–2533.

[57] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[58] C. De Boor and C. De Boor, *A practical guide to splines*. springer-verlag New York, 1978, vol. 27.

[59] 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," *arXiv preprint arXiv:1912.02781*, 2019.

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

[61] 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. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015.

[62] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan *et al.*, "Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation," *Medical Image Analysis*, vol. 69, p. 101950, 2021.

[63] X. Zhuang, J. Xu, X. Luo, C. Chen, C. Ouyang, D. Rueckert, V. M. Campello, K. Lekadir, S. Vesal, N. RaviKumar *et al.*, "Cardiac segmentation on late gadolinium enhancement mri: a benchmark study from multi-sequence cardiac mr segmentation challenge," *arXiv preprint arXiv:2006.12434*, 2020.

[64] N. Bloch, A. Madabhushi, H. Huisman, J. Freymann, J. Kirby, M. Grauer, A. Enquobahrie, C. Jaffe, L. Clarke, and K. Farahani, "Nci-isbi 2013 challenge: automated segmentation of prostate structures," *The Cancer Imaging Archive*, vol. 370, 2015.

[65] 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," *Computers in biology and medicine*, vol. 60, pp. 8–31, 2015.

[66] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang *et al.*, "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[67] 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 *European Conference on Computer Vision*. Springer, 2020, pp. 762–780.

[68] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

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

[70] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *International journal of computer vision*, vol. 59, no. 2, pp. 167–181, 2004.