

Generalizable Single-Source Cross-modality Medical Image Segmentation via Invariant Causal Mechanisms

Boqi Chen^{1,2,3}
 Reto Sutter^{4,5}

Yuanzhi Zhu³
 Gunnar Rätsch^{1,2}

Yunke Ao^{1,2,4}
 Ender Konukoglu³

Sebastiano Caprara^{4,5}
 Anna Susmelj^{2,3}

¹ Department of Computer Science, ETH Zurich ² ETH AI Center

³ Computer Vision Lab, ETH Zurich ⁴ Balgrist University Hospital ⁵ University of Zurich

Abstract

Single-source domain generalization (SDG) aims to learn a model from a single source domain that can generalize well on unseen target domains. This is an important task in computer vision, particularly relevant to medical imaging where domain shifts are common. In this work, we consider a challenging yet practical setting: SDG for cross-modality medical image segmentation. We combine causality-inspired theoretical insights on learning domain-invariant representations with recent advancements in diffusion-based augmentation to improve generalization across diverse imaging modalities. Guided by the “intervention-augmentation equivariant” principle, we use controlled diffusion models (DMs) to simulate diverse imaging styles while preserving the content, leveraging rich generative priors in large-scale pretrained DMs to comprehensively perturb the multidimensional style variable. Extensive experiments on challenging cross-modality segmentation tasks demonstrate that our approach consistently outperforms state-of-the-art SDG methods across three distinct anatomies and imaging modalities. The source code is available at <https://github.com/ratschlab/ICMSeg>.

1. Introduction

Deep convolutional neural networks (NNs) have recently achieved remarkable performance in medical image segmentation [1, 20, 30, 33, 48, 71]. Despite their success, these models’ performance heavily relies on the assumption that training and testing samples are independent and identically distributed (i.i.d.). In medical applications, however, distribution shifts between source (training) and target (test) data frequently occur due to varying scanning protocols, different device vendors, and diverse imaging modalities, etc. Directly applying models trained on a source domain to an unseen target domain with a different distribution typically results in significant performance degradation [23]. Broadly,

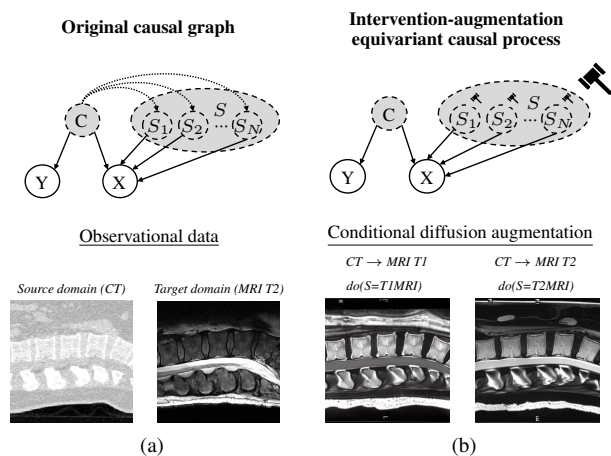


Figure 1. Overview of causal framework. (a) SCM for data generative process and examples of samples from the observational distribution. (b) Causal graph after *do*-intervention on style variables and examples of corresponding equivariant augmentations generated with the conditional diffusion model.

two main types of domain shift exist in medical imaging: cross-center, where the shift primarily stems from different acquisition protocols and devices, and cross-modality, where the shift is caused by different imaging modalities (e.g., CT to MRI). The latter presents a more substantial distribution shift between the source and target domains.

To address the domain shift problem, researchers have proposed two primary solutions: unsupervised domain adaptation (UDA) [5, 13, 17, 37, 67] and domain generalization (DG) [3, 9, 31, 32, 35, 36, 38, 40]. UDA assumes access to unlabeled data from the target domain, enabling joint training on both source and target images to learn domain-invariant features. While effective for known target domains, UDA cannot guarantee high performance in unseen domains and often faces challenges in accessing target domain data due to privacy regulations. In contrast, DG aims to train a generalizable and robust predictive model using only source domain data, thus overcoming the limitations

of UDA and providing a compelling alternative for real-world medical applications. Conventional DG [9, 35, 36] approaches seek to learn a domain-invariant feature space by aligning distributions among multiple source domains. Recent works [2, 32] have established a more formal link between learned domain-invariances and the data generative process from a causal perspective, leading to improved performance in *multi-source* DG (MDG). However, acquiring multi-domain training data for medical applications remains challenging due to privacy issues and data scarcity. A more practical alternative is to train a model on a single source domain and enhance its generalizability to unseen target domains, known as *single-source* DG (SDG). Adapting MDG strategies to SDG, however, is challenging due to the limited diversity in training data.

In this paper, we consider a challenging yet practical setting: cross-modality SDG. While previous studies have explored this problem in medical imaging, most rely on basic augmentation techniques, such as manually-selected image-level transformations [40, 70] or random intensity adjustments [56, 62], which are often insufficient for bridging domain gaps in cross-modality contexts. Recently, diffusion-based augmentation methods have shown significant promise in handling large distribution shifts in the natural image domain [14, 22]. This success can be attributed to two key factors: (1) the rich generative prior from large-scale pretraining and (2) comprehensive style mining through text prompting, resulting in refined, target-specific augmentations that improve generalization. However, their efficacy in medical imaging remains unexplored.

Motivated by this, we propose to combine the diffusion-based augmentations with recent theoretical advancements in causal identifiability [59] and generalization via invariant mechanisms [2]. Specifically, we assume a structural causal model (SCM) for the data generation process, representing the dependency among observed data, labels (*i.e.*, segmentation masks), and unobserved latent variables of content and style (Fig. 1a). We then develop a style-augmentation method based on the “intervention-augmentation equivariant” principle [19] using controlled diffusion models (DMs), which enables simulating diverse imaging styles across modalities via text prompting, while preserving the content. We posit that the style variable is inherently multidimensional, and our strong augmentation strategy comprehensively intervenes on it (Fig. 1b), whereas previous causality-inspired SDG methods [3, 40] rely on basic image-level transformations that may not fully disentangle style from content, potentially limiting model generalizability to unseen domains.

In summary, the main contributions of this paper are summarized as:

- We leverage recent theoretical advances in DG under causal assumptions to design our approach for learning

domain-invariant representations.

- We propose a style intervention method that utilizes rich generative prior of pretrained text-to-image DMs.
- We demonstrate our method’s effectiveness through extensive cross-modality medical image segmentation experiments, consistently outperforming state-of-the-art SDG methods across three distinct anatomies and imaging modalities.

2. Related Work

2.1. Single-source Domain Generalization

SDG is a more challenging and under-explored task compared to MDG. Due to insufficient diversity of the training data, SDG approaches usually first expand the distribution of the source domain via data augmentation techniques such as manually-selected image-level transformations [7, 47, 66] and learning-based augmentation [29, 45, 61, 68]. For instance, in natural image domain, Li et al. [29] leverage random style transfer with adaptive instances normalization [18], followed by contrastive learning. Chen et al. [3] apply predefined image-level transformations such as brightness and contrast adjustments to create auxiliary domains and then conduct counterfactual inference to learn robust features by analyzing the causal effect of simulated domain shifts. While effective for natural images, such augmentations do not introduce sufficient variability for medical imaging data, which contains highly specialized and more nuanced visual patterns.

To this end, several approaches tailored to medical imaging have been proposed. Zhang et al. [66] utilize stacked photometric and geometric transformations to simulate changes in image quality, appearance, and spatial configuration. Zhou et al. [70] employed Bézier transformations with switchable batch normalization layers based on style proximity. Ouyang et al. [40] introduce local variability through randomly-sampled kernel augmentations with pseudo-correlation map-based mixing, simulating more diverse domain shifts. Su et al. [56] further advance this concept by integrating global and local Bézier transformations with saliency-based fusion, potentially offering more semantically meaningful augmentations. Despite these advancements, these methods are based on generic image-level manipulations, limiting their ability to capture complex domain variations. In contrast, DMs leverage rich generative priors and enable target-specific augmentations through text prompting, introducing more complex and comprehensive distribution shifts.

2.2. Causality for Domain Generalization

The theory of causation has been extensively explored as a foundation for DG [28, 50, 51, 64]. The idea of invariance

is closely linked to causality [36]. Peters et al. [44] introduce the concept of invariant causal prediction, identifying causal relationships that are stable across different environments. Arjovsky et al. [2] establish a formal link among generalization, causation and invariance. These theoretical advancements in understanding the relationship between causation and invariances have sparked a distinct line of research [3, 31, 32, 34, 40] focused on designing loss functions and regularization techniques based on these principles.

3. Method

In this section, we first present a motivation that obtaining domain-invariant representations using the direct causal parents of segmentation labels in the corresponding SCM leads to an optimal solution for the DG problem. We demonstrate how to achieve such an optimal solution in SDG using the “soft” intervention with contrastive learning. We then introduce a diffusion-based augmentation technique that is based on the “intervention-augmentation equivariance” principle such that it can serve as a surrogate tool for simulating interventions on unobserved variables in the SCM. Finally, we provide an overview of the two training stages of our method.

3.1. Single-source Domain Generalization under a Causal Perspective

Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ represent input images and corresponding segmentation masks. We extend common assumptions [31, 40] on the data generative process from a causal perspective by allowing an additional causal relation from content to style as proposed in [59] (Fig. 1a). We further assume that an expert (or oracle) is able to provide correct segmentation masks Y from observations X alone.

In the context of SDG, we assume that we are given training pairs from a single domain D_0 : $\{(X_i^0, Y_i^0)\}_{i=1}^{n_0}$, and a fixed set of target domains $D_1, \dots, D_N$. Each dataset D_e contains i.i.d. samples from some probability distribution $P(X^e, Y^e)$. We aim to obtain an optimal predictor f to enable out-of-distribution (OOD) generalization. In particular, f is trained on D_0 such that f minimizes the worst case risk $R^e(f) := \mathbb{E}_{X^e, Y^e} [\ell(f(X^e), Y^e)]$ for any given target domain D_e . Arjovsky et al. [2] demonstrate that an invariant predictor would obtain an optimal solution if and only if it uses only the direct causal parents of Y in the corresponding SCM [42].

Under the described SCM assumptions, von Kügelgen et al. [59] prove that if the augmented pairs of views (X, X^+) in contrastive methods are generated under the principle of a “soft” intervention on S , then the InfoNCE [39] objective combined with an encoder function Φ (e.g., NN) identifies

the invariant content C partition:

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_{X \sim p_X} \left[\log \frac{\exp(\text{sim}(\nu, \nu^+)/\tau)}{\sum_{X^- \in \mathcal{N}} \exp(\text{sim}(\nu, \nu^-)/\tau)} \right], \quad (1)$$

where $\nu = \Phi(X)$, $\text{sim}(\cdot, \cdot)$ is some similarity metric (e.g., cosine similarity), τ is a temperature parameter, and X^- are negative pairs.

The notion of *identifiability* is crucial here, as it means that under suitable augmentations, the learned representations are expected to match the true underlying latent content factors - the essential ingredient for the optimal solution of DG problem [2].

To adapt to the segmentation task and retain the spatial structure by patch-wise contrastive learning, we apply the InfoNCE loss after the bottleneck layer of the U-Net and combine it with segmentation loss (such as cross-entropy or Dice loss [8]):

$$\mathcal{L}_{\text{combined}} = \mathcal{L}_{\text{Seg}} + \lambda_{\text{reg}} \mathcal{L}_{\text{InfoNCE}}, \quad (2)$$

where λ_{reg} is a hyperparameter that balances the contribution of the InfoNCE loss.

In real-world applications, direct intervention on the style variable is infeasible as it is unobserved. For instance, in medical imaging, this would require scanning the same patient using different imaging modalities multiple times under controlled conditions—a process that is often impractical for acquiring training data. However, Ilse et al. [19] introduce the concept of *intervention-augmentation equivariance*, formally demonstrating that data augmentation *can* serve as a surrogate tool for simulating interventions.

In the following section, we apply this principle to design augmentation pairs in Eq. (1), enabling the InfoNCE-regularized training to recover content features and obtain the optimal predictor described above.

3.2. Diffusion-based Style Intervention

3.2.1 Diffusion-based Image Generation

Recent advancements in text-to-image latent DMs (LDMs), such as Stable Diffusion (SD) [46], show the ability to generate diverse and realistic images. The rich prior knowledge learned from vast text-image datasets is widely exploited for downstream tasks such as data augmentation [10, 58]. An LDM comprises a Variational Auto-Encoder (VAE) [24] which projects images into a latent space and a U-Net [48] that learns the denoising process within this space. Specifically, the diffusion process gradually introduces random Gaussian noise to the latent representation $\mathbf{z}_0$ of a clean image $\mathbf{x}_0$ over T discrete steps [53, 55]. For any intermediate noisy state, we can express it formally using the property of Gaussian noise as [16]:

$$\mathbf{z}_t = \sqrt{\alpha_t} \mathbf{z}_0 + \sqrt{1 - \alpha_t} \epsilon, \quad (3)$$

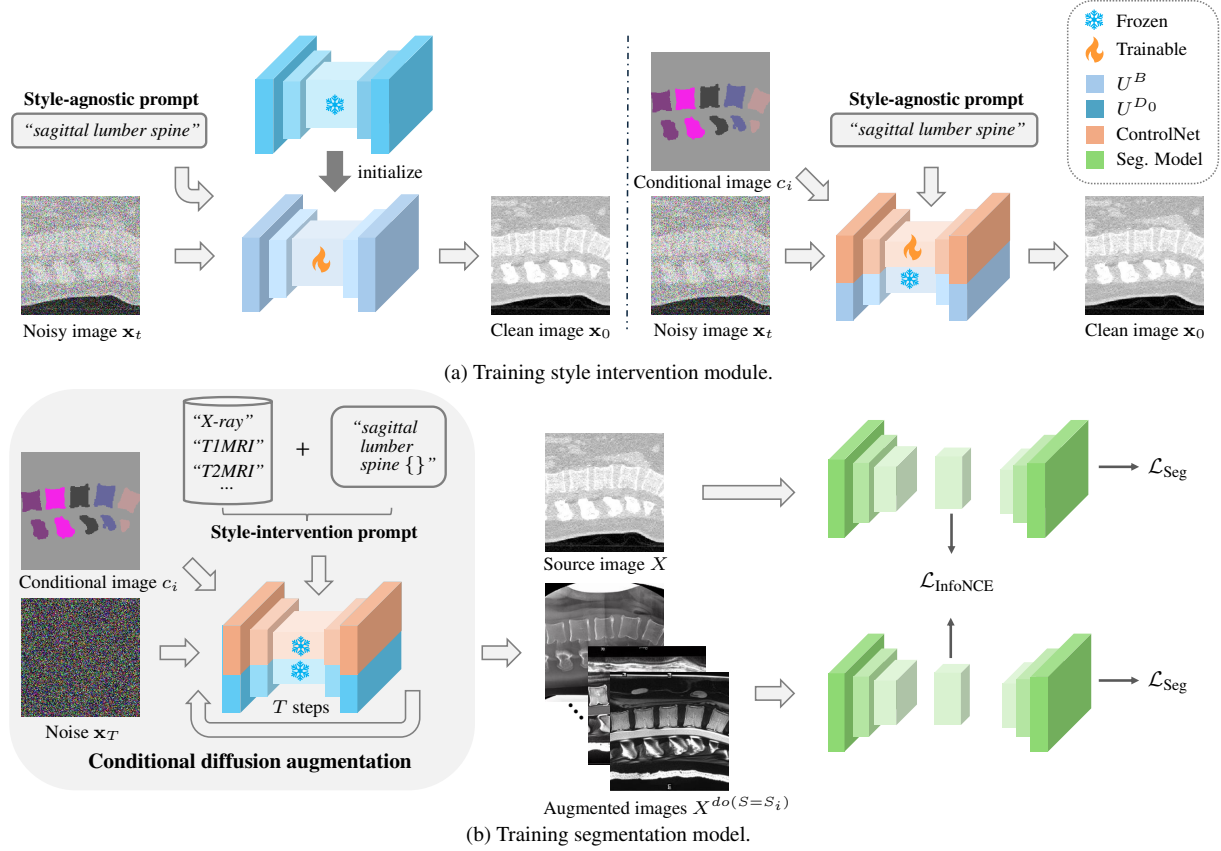


Figure 2. Overview of our method. (a) Fine-tune pretrained SD U-Net (U^B) on the source domain D_0 (left), and then train ControlNet with the fine-tuned SD U-Net (U^{D_0}) to inject image conditions (right), both using style-agnostic prompts. (b) Generate style-intervened images using ControlNet and U^B with style-intervention prompts. The segmentation model is trained on pairs of original and style-intervened images using a segmentation loss and InfoNCE regularization.

where $t \in \{1, \dots, T\}$ represents the current timestep or noise level, ϵ is a Gaussian noise, and α_t is a predefined noise schedule [16, 54]. Each training step estimates the noise added to $\mathbf{z}_t$ at a randomly selected timestep t by optimizing the following regression loss [16, 46]:

$$\mathcal{L}_{SD} = \mathbb{E}_{\mathbf{z}_0, t, c_t, \epsilon} [\gamma_t \|\epsilon - \epsilon_\theta(\mathbf{z}_t, c_t, t)\|_2^2], \quad (4)$$

where t is uniformly sampled from $[0, T]$. The coefficient γ_t adjusts the weight for different noise levels, ϵ_θ is the NN that predicts the added noise, and c_t is the text prompt that describes the image. During inference, we iteratively denoise a randomly sampled Gaussian noise $\mathbf{z}_T \sim \mathcal{N}(0, I)$ to generate $\mathbf{z}_0$ based on the prompt c_t , which is then decoded into an image $\mathbf{x}_0$.

3.2.2 Controllable Generation for Style Intervention

To achieve intervention-augmentation equivariance, we assume the segmentation mask represents the content variable in the causal data generation process describe in Fig. 1a. Since we wish to intervene only on the style variable, the augmentations should share the same segmentation masks

as source domain images. Such a constraint is difficult to satisfy via only prompting on the anatomy. To this end, we add a ControlNet module [65]. The training objective of this module is to teach the model to denoise data given a conditional image (e.g., semantic segmentation mask, etc.), which ensures that the model learns to preserve condition-specific contents during the generation process. In our case, we train the ControlNet module with source domain images and corresponding segmentation masks such that it captures the relationship between generated anatomical structures and given segmentation masks. Then, we can leverage the generative prior of the pretrained SD and the content-preserving capabilities of the ControlNet to generate style-varied images while maintaining the segmentation-relevant content.

Specifically, the ControlNet module is initialized as a copy of the encoding layers of the SD U-Net and connected via zero convolutions. The training of the ControlNet follows the loss defined in Eq. (4) with an additional segmentation mask condition c_i :

$$\mathcal{L}_{ControlNet} = \mathbb{E}_{\mathbf{z}_0, t, c_t, c_i, \epsilon} [\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t, c_t, c_i)\|_2^2], \quad (5)$$

here θ represents the parameters of both the trainable ControlNet module and the frozen pretrained SD backbone U-Net (U^B). After training, the ControlNet module will provide additional control signal, *i.e.*, the segmentation masks, to U^B , ensuring content preservation in the generation process.

Jia et al. [22] empirically show that ControlNet, when trained with U^B on the source domain D_0 directly, learns to inject not only image content conditions but also style information of D_0 . We conjecture this undesired style overfitting is due to a mismatch between the prior information in U^B and the domain knowledge of D_0 used to train the ControlNet, which compromises the style prior stored in U^B . To mitigate this issue, we adopt the Style Swap technique [22], forcing ControlNet to focus on learning the content information. Specifically, instead of training ControlNet directly on D_0 with U^B , we first fine-tune U^B on D_0 using instance-fine-tuning strategies [11, 15, 49, 60] to obtain an instance fine-tuned U-Net (U^{D_0}). Following [22], we employ DreamBooth [49] which adjusts SD U-Net parameters to recognize specific instances by optimizing the following objective:

$$\mathcal{L}_{U^{D_0}} = \mathbb{E}_{\mathbf{z}_0, t, c_t, \epsilon, \epsilon'} \left[\gamma_t \|\hat{\mathbf{z}}_\theta(\mathbf{z}_t, c_t, t) - \mathbf{z}_0\|_2^2 + \lambda_{\text{pr}} \gamma_{t'} \|\hat{\mathbf{z}}_\theta(\mathbf{z}_{t'}^{\text{pr}}, c_{t'}^{\text{pr}}, t') - \mathbf{z}_0^{\text{pr}}\|_2^2 \right], \quad (6)$$

where $\hat{\mathbf{z}}_\theta$ is the clean signal prediction network [46] (which is equivalent to noise prediction ϵ_θ up to an affine transformation), $\mathbf{z}_0 \in D_0$ is the data of new instance, and $\mathbf{z}_0^{\text{pr}}$ is the generated samples of the pre-trained model using text condition c_t^{pr} . The second term is the prior-preservation term and λ_{pr} controls the relative weight of this term.

3.3. Overview of Methods

Fig. 2 illustrates the overview of our method. It consists of two stages: training the style intervention module and the segmentation model. In the first stage (Fig. 2a), we fine-tune a pretrained SD U-Net (U^B) on the source domain D_0 using style-agnostic prompts. These prompts guide image generation with information on anatomy but not imaging modality, resulting in an instance-tuned U-Net (U^{D_0}) that overfits to the style of D_0 . We then train the ControlNet alongside U^{D_0} on D_0 , focusing on learning to inject image conditions into the generation process. In the second stage (Fig. 2b), we integrate the original U^B with ControlNet to leverage its rich style prior and use segmentation labels from D_0 as image conditions. To construct the style-intervention prompt, we concatenate unique class names of the anatomies present in the segmentation masks with a style prompt randomly sampled from a predefined style pool. This prompt serves as the text condition. This

approach enables us to generate training pairs of original and style-intervened images. Finally, these pairs are used to train the segmentation model, which is optimized using the loss function defined in Eq. (2).

4. Experiments

4.1. Experimental Setting

Datasets and Preprocessing. We evaluate our method on three tasks: abdominal segmentation (AS), lumbar spine segmentation (LSS) and lung segmentation (LS) with three imaging modalities: CT, MRI and X-Ray. For the AS task, we segment four abdominal organs: liver, right kidney (R-Kidney), left kidney (L-Kidney), and spleen. We use the CT and MRI data from [27] and [21], following the split in [40] and [21], respectively. For the LSS task, we perform instance segmentation of the vertebrae from L1 to L5. We construct a cross-modality dataset using CT scans from [52], T2-MRI from [41] with 30 volumes of in-house T1/T2 MRI scans, and X-Ray scans from [25]. For the LS task, we segment left lung (L-Lung) and right lung (R-Lung) with CT data from [63] and X-Ray data from [6]. All 3D volumes are reformatted to 2D and resized to 256×256 pixels during training. During inference, we obtain a 3D volume by stacking 2D slice predictions and compare it with the ground-truth 3D volume to calculate the per-volume Dice score. The final result is the average across all test volumes. Common augmentations proposed in [66] is applied for all methods.

Implementation Details. We use U-Net [48] as our segmentation model and Dice loss [8] as the segmentation loss. For all datasets, we train the model for 50 epochs with a batch size of 64. We utilize the Adam optimizer with an initial learning rate of 1×10^{-3} . To ensure stable training, we decay the learning rate following the cosine annealing schedule. For the style intervention module, we fine-tune the SD model trained on the ROCO dataset [43] using DreamBooth for 15,000 iterations with a learning rate of 1×10^{-6} and batch size of 1, and train the ControlNet for 72,000 iterations with a learning rate of 1×10^{-5} and batch size of 2. All slices are resized to 512×512 pixels to meet the training resolution requirements of SD. All experiments are conducted using PyTorch framework on a single NVIDIA RTX A6000 GPU with 48 GB of memory. Further details are provided in the supplemental material.

4.2. Results and Discussion

We evaluate our method against multiple baselines: a weak baseline consisting of a segmentation model trained on the source domain with only basic augmentations; a supervised benchmark where the segmentation model is trained on the target domain; and several DG methods. In particular, MixStyle [69] enhances generalization by mix-

Table 1. Comparison of methods on the cross-modality abdominal segmentation (AS) task. Results are mean $\pm$ standard deviation over 3 random initializations. Best and second-best per column (excluding Baseline and Benchmark) are bold and underlined.

Method	Abdominal CT $\rightarrow$ Abdominal MRI					Abdominal MRI $\rightarrow$ Abdominal CT				
	Liver	R-Kidney	L-Kidney	Spleen	Average	Liver	R-Kidney	L-Kidney	Spleen	Average
BASLINE	52.57 $\pm$ 2.8	32.53 $\pm$ 6.5	57.23 $\pm$ 1.3	37.07 $\pm$ 4.9	44.87 $\pm$ 2.9	91.27 $\pm$ 0.6	67.77 $\pm$ 1.9	77.70 $\pm$ 2.0	88.13 $\pm$ 1.8	81.20 $\pm$ 0.4
MIXSTYLE [69]	89.83 $\pm$ 1.9	85.97 $\pm$ 3.4	81.30 $\pm$ 7.9	67.73 $\pm$ 8.7	81.23 $\pm$ 1.7	87.03 $\pm$ 1.7	81.03 $\pm$ 2.2	72.60 $\pm$ 4.9	86.23 $\pm$ 1.3	81.73 $\pm$ 1.1
RANDCONV [62]	82.50 $\pm$ 2.9	78.37 $\pm$ 0.2	82.27 $\pm$ 2.9	72.00 $\pm$ 2.8	78.77 $\pm$ 1.0	90.13 $\pm$ 1.7	84.10 $\pm$ 0.7	85.60 $\pm$ 1.9	87.57 $\pm$ 0.9	86.87 $\pm$ 0.8
DUALNORM [70]	82.90 $\pm$ 1.7	80.17 $\pm$ 3.1	73.33 $\pm$ 0.8	55.10 $\pm$ 2.1	72.87 $\pm$ 1.0	84.47 $\pm$ 1.3	80.80 $\pm$ 6.6	79.97 $\pm$ 6.7	80.17 $\pm$ 5.7	81.60 $\pm$ 3.5
CSDG [40]	87.27 $\pm$ 0.7	84.93 $\pm$ 0.0	87.00 $\pm$ 0.9	73.67 $\pm$ 1.9	83.20 $\pm$ 0.7	91.77 $\pm$ 0.1	86.63$\pm$0.7	89.00 $\pm$ 0.4	88.93 $\pm$ 0.6	89.07 $\pm$ 0.1
SLAUG [56]	86.73 $\pm$ 0.8	86.60 $\pm$ 2.2	89.97$\pm$0.3	79.27 $\pm$ 0.8	85.67 $\pm$ 0.9	92.97 $\pm$ 0.4	86.07 $\pm$ 0.9	90.30$\pm$0.7	87.50 $\pm$ 0.7	89.17 $\pm$ 0.5
OURS	89.57$\pm$1.2	87.60$\pm$0.9	87.30 $\pm$ 0.6	80.33$\pm$0.6	86.20$\pm$0.3	92.97$\pm$0.2	86.27$\pm$1.1	90.20$\pm$1.1	91.33$\pm$0.7	90.17$\pm$0.7
BENCHMARK	96.27 $\pm$ 0.2	95.83 $\pm$ 0.2	95.50 $\pm$ 0.2	94.97 $\pm$ 0.2	95.63 $\pm$ 0.1	95.50 $\pm$ 0.2	85.20 $\pm$ 1.1	91.20 $\pm$ 1.2	93.17 $\pm$ 0.1	91.30 $\pm$ 0.5

ing styles of training instances, while RandConv [62] improves robustness via randomly initialized convolution layers. We also compare with three SDG methods for medical image segmentation. DualNorm [70] employs separate normalization layers for domain-similar and domain-dissimilar augmentations generated via nonlinear Bézier transformations [72]. CSDG [40] uses a shallow network to generate novel training samples, mixes them via pseudo-correlation maps, and learns domain-invariant features using a Kullback–Leibler divergence (KL-Div) loss-based regularizer to remove spurious correlations. SLAUG [56] enforces class-level representation invariance using constrained Bézier transformations on both global and class-level regions, fusing the locally and globally transformed augmentations using a saliency-balancing approach. Implementation details are provided in the supplementary material. Preprocessing and augmentation are consistent across all methods, as described in Sec. 4.1.

4.2.1 Comparison with State-of-the-art Methods

The quantitative results for cross-modality AS, LSS and LS are summarized in Tab. 1, Tab. 2, and Tab. 3, respectively. Overall, our approach consistently outperforms state-of-the-art methods across all tasks. Specifically, in the AS task, our method achieves the highest average Dice scores of 86.20 percentage points (pp) and 90.17 pp in the CT to MRI and MRI to CT directions, respectively. Although the improvement over the most competitive baseline, SLAUG, is modest (0.53 pp and 1.00 pp), our approach consistently approaches the supervised benchmark, with only 1.13 pp difference for source domain MRI. In the LSS task, our method surpasses the previous best-performing methods by 3.13 pp (CT to MRI) and 4.03 pp (MRI to CT). The more substantial improvements compared with the AS task showcase our method’s superiority in generalizing across large domain gaps. In the LS task, our method demonstrates an improvement of over 10 pp in Dice score compared to the baseline, whereas other methods show only marginal

(less than 1 pp) or even negative improvements. Specifically, we achieve an average Dice score of 78.79 pp, which is 10.78 pp higher than the second-best method (SLAUG, 68.01 pp). This substantial gain highlights the robustness of our approach, particularly in scenarios with complex domain shifts where other DG techniques struggle. These consistent performance gains across various tasks and modalities demonstrate that our approach effectively handles both small and large domain gaps, offering superior generalizability across diverse anatomical structures and imaging modalities in medical image segmentation.

We also visualize the segmentation results of our method and other DG approaches for all tasks in Fig. 3, Fig. 4 and Fig. 5, respectively. The first two columns show source and target domain images, illustrating the domain shift. We can see that the segmentation masks in the unseen target domain produced by our method exhibit greater accuracy and improved spatial continuity of the foreground compared to the baselines.

4.2.2 Ablation Study

Efficacy of InfoNCE-based Regularization. We evaluate the effectiveness of using patch-wise InfoNCE loss for regularization through ablation studies with three variants: (1) training the segmentation model on original and style-intervened images without regularization (w/o Reg.); (2) applying KL-Div loss to the output logits from original and style-intervened images following [40] (w/ KL-Div); and (3) using standard InfoNCE loss [4] on average-pooled features (w/ S. InforNCE). Results in Tab. 3 show training on augmented data without regularization results in significantly lower performance, with a Dice score 7.46 pp below ours, highlighting the importance of domain-invariance regularization. Our approach outperforms KL-Div-based regularization by 3.63 pp, demonstrating InfoNCE loss’s superiority in extracting domain-invariant content features. The 4.56 pp improvement over non-patch-wise InfoNCE loss suggests the value of preserving spatial structure for seg-

Table 2. Comparison of methods on the cross-modality lumbar spine segmentation (LSS) task. Results are mean $\pm$ standard deviation over 3 random initializations. Best and second-best per column (excluding Baseline and Benchmark) are bold and underlined.

Method	Lumbar Spine CT $\rightarrow$ Lumbar Spine MRI						Lumbar Spine MRI $\rightarrow$ Lumbar Spine CT					
	L1	L2	L3	L4	L5	Avg	L1	L2	L3	L4	L5	Avg
BASELINE	0.00 $\pm$ 0.0	0.00 $\pm$ 0.0	0.00 $\pm$ 0.0	0.00 $\pm$ 0.0	0.25 $\pm$ 0.0	0.05 $\pm$ 0.4	0.05 $\pm$ 0.1	0.60 $\pm$ 0.7	1.57 $\pm$ 1.4	1.63 $\pm$ 1.8	0.00 $\pm$ 0.0	0.80 $\pm$ 0.7
MIXSTYLE [69]	41.90 $\pm$ 6.6	37.20 $\pm$ 2.0	39.60 $\pm$ 7.9	48.57 $\pm$ 6.7	47.33 $\pm$ 6.7	42.93 $\pm$ 2.6	42.83 $\pm$ 3.7	48.77 $\pm$ 5.4	50.43 $\pm$ 7.5	43.63 $\pm$ 1.7	36.87 $\pm$ 2.9	44.47 $\pm$ 3.9
RANDCONV [62]	65.93 $\pm$ 0.8	66.07 $\pm$ 0.6	65.00 $\pm$ 1.2	66.63 $\pm$ 1.5	70.27 $\pm$ 1.6	66.80 $\pm$ 1.0	52.17 $\pm$ 1.5	51.43 $\pm$ 0.6	51.67 $\pm$ 0.6	41.23 $\pm$ 1.1	36.17 $\pm$ 2.7	46.53 $\pm$ 1.0
DUALNORM [70]	40.00 $\pm$ 4.6	46.20 $\pm$ 1.3	45.63 $\pm$ 2.2	42.70 $\pm$ 4.0	46.77 $\pm$ 3.0	44.27 $\pm$ 2.8	28.73 $\pm$ 0.7	30.37 $\pm$ 0.5	29.63 $\pm$ 3.8	27.33 $\pm$ 3.4	22.30 $\pm$ 2.4	27.70 $\pm$ 1.8
CSDG [40]	69.70 $\pm$ 2.1	70.47 $\pm$ 1.1	70.23 $\pm$ 0.8	73.10 $\pm$ 0.7	75.53 $\pm$ 0.4	71.80 $\pm$ 0.9	46.10 $\pm$ 3.9	52.70 $\pm$ 1.0	54.70 $\pm$ 1.7	50.23 $\pm$ 3.0	39.20 $\pm$ 0.3	48.57 $\pm$ 0.3
SLAUG [56]	43.80 $\pm$ 5.5	36.90 $\pm$ 3.7	35.60 $\pm$ 3.0	44.10 $\pm$ 8.2	44.40 $\pm$ 11.6	40.97 $\pm$ 6.0	73.90 $\pm$ 1.6	72.73 $\pm$ 2.3	75.77 $\pm$ 2.1	69.57 $\pm$ 2.1	54.43$\pm$0.9	69.27 $\pm$ 1.8
OURS	72.23$\pm$1.6	73.00$\pm$2.4	74.20$\pm$1.2	76.57$\pm$0.6	78.70$\pm$0.4	74.93$\pm$1.2	80.50$\pm$0.4	81.20$\pm$0.4	77.70$\pm$0.6	73.70$\pm$1.1	<u>53.60$\pm$0.7</u>	73.30$\pm$0.1
BENCHMARK	89.33 $\pm$ 0.8	89.97 $\pm$ 0.2	88.43 $\pm$ 0.1	89.13 $\pm$ 0.8	80.33 $\pm$ 13.8	87.43 $\pm$ 2.4	90.70 $\pm$ 1.8	89.77 $\pm$ 2.8	88.23 $\pm$ 1.8	88.80 $\pm$ 1.3	72.10 $\pm$ 14.4	85.90 $\pm$ 2.8

Method	Lumbar Spine CT $\rightarrow$ Lumbar Spine X-Ray						Lumbar Spine MRI $\rightarrow$ Lumbar Spine X-Ray					
	L1	L2	L3	L4	L5	Avg	L1	L2	L3	L4	L5	Avg
BASELINE	3.27 $\pm$ 1.6	8.32 $\pm$ 4.4	11.08 $\pm$ 4.3	12.17 $\pm$ 4.6	13.45 $\pm$ 8.4	9.66 $\pm$ 3.4	0.03 $\pm$ 0.06	1.50 $\pm$ 1.7	1.47 $\pm$ 1.3	5.00 $\pm$ 4.0	2.17 $\pm$ 0.6	2.03 $\pm$ 1.0
MIXSTYLE [69]	35.87 $\pm$ 14.6	50.20 $\pm$ 9.0	52.37 $\pm$ 4.8	47.53 $\pm$ 5.9	42.03 $\pm$ 7.3	45.60 $\pm$ 7.1	24.33 $\pm$ 7.3	42.63 $\pm$ 11.8	46.47 $\pm$ 5.9	45.20 $\pm$ 5.4	44.57 $\pm$ 13.3	40.64 $\pm$ 8.4
RANDCONV [62]	67.17 $\pm$ 1.5	72.53 $\pm$ 1.3	72.03 $\pm$ 1.2	71.43 $\pm$ 1.4	72.13 $\pm$ 0.8	71.06 $\pm$ 0.9	49.83 $\pm$ 2.8	64.40 $\pm$ 3.0	65.87 $\pm$ 1.3	72.57 $\pm$ 0.1	63.23 $\pm$ 1.2	63.18 $\pm$ 1.6
DUALNORM [70]	19.74 $\pm$ 4.2	23.25 $\pm$ 2.9	25.99 $\pm$ 3.1	23.24 $\pm$ 4.8	31.29 $\pm$ 2.7	24.70 $\pm$ 1.4	45.34 $\pm$ 2.3	53.61 $\pm$ 2.1	54.94 $\pm$ 2.1	55.95 $\pm$ 1.9	48.56 $\pm$ 2.7	51.68 $\pm$ 0.2
CSDG [40]	66.20 $\pm$ 0.7	70.50 $\pm$ 0.4	70.67 $\pm$ 0.6	71.17 $\pm$ 0.7	73.40$\pm$2.0	70.39 $\pm$ 0.5	49.20 $\pm$ 0.4	65.00 $\pm$ 1.0	68.47 $\pm$ 1.5	73.80$\pm$0.4	65.53 $\pm$ 0.9	64.40 $\pm$ 0.1
SLAUG [56]	62.80 $\pm$ 4.0	69.13 $\pm$ 2.2	67.70 $\pm$ 0.6	67.80 $\pm$ 0.6	72.10 $\pm$ 0.7	67.91 $\pm$ 1.3	42.87 $\pm$ 3.0	48.50 $\pm$ 5.8	51.73 $\pm$ 5.6	55.67 $\pm$ 7.8	55.33 $\pm$ 5.7	50.82 $\pm$ 5.1
OURS	69.97$\pm$2.1	75.23$\pm$0.9	73.97$\pm$0.1	74.63$\pm$0.7	71.93 $\pm$ 0.3	73.15$\pm$0.3	63.40$\pm$1.6	72.95$\pm$2.1	75.00$\pm$1.1	<u>73.75$\pm$3.0</u>	67.65$\pm$0.6	70.55$\pm$0.2
BENCHMARK	93.1 $\pm$ 0.4	93.9 $\pm$ 0.3	94.3 $\pm$ 0.3	94.3 $\pm$ 0.2	92.3 $\pm$ 0.3	93.6 $\pm$ 0.3	93.1 $\pm$ 0.4	93.9 $\pm$ 0.3	94.3 $\pm$ 0.3	94.3 $\pm$ 0.2	92.3 $\pm$ 0.3	93.6 $\pm$ 0.3

Table 3. Comparison of methods on the cross-modality lung segmentation (LS) task. Results are mean $\pm$ standard deviation over 3 random initializations. Best and second-best per column (excluding Baseline and Benchmark) are bold and underlined.

Method	Lung CT $\rightarrow$ Lung X-Ray		
	L-Lung	R-Lung	Average
BASELINE	59.05 $\pm$ 0.6	75.60 $\pm$ 0.1	67.33 $\pm$ 0.3
MIXSTYLE [69]	55.63 $\pm$ 7.1	75.83 $\pm$ 1.5	65.73 $\pm$ 4.2
RANDCONV [62]	50.03 $\pm$ 0.3	71.30 $\pm$ 3.8	60.67 $\pm$ 2.0
DUALNORM [70]	36.82 $\pm$ 0.3	19.24 $\pm$ 0.7	28.03 $\pm$ 2.4
CSDG [40]	52.14 $\pm$ 1.6	66.71 $\pm$ 9.1	59.43 $\pm$ 3.7
SLAUG [56]	57.06 $\pm$ 0.2	78.97$\pm$0.2	68.01 $\pm$ 0.0
OURS	79.84$\pm$0.5	<u>77.75$\pm$0.5</u>	78.79$\pm$0.3
BENCHMARK	94.46 $\pm$ 0.8	96.32 $\pm$ 0.7	96.39 $\pm$ 0.8

mentation tasks within contrastive learning frameworks.

Table 4. Ablations on the cross-modality lumbar spine segmentation (LSS) task (CT $\rightarrow$ MRI). All results are given as mean over 3 random initializations. The best result per column is in bold.

Method	Lumbar Spine CT $\rightarrow$ Lumbar Spine MRI					
	L1	L2	L3	L4	L5	Avg
BASELINE	0.00	0.00	0.00	0.00	0.25	0.05
w/o REG.	63.00	60.57	62.73	73.00	78.07	67.47
w/ KL-DIV	67.17	67.80	70.20	74.53	76.73	71.30
w/ S. INFO NCE	64.80	66.63	68.40	73.90	78.07	70.37
OURS	72.23	73.00	74.20	76.57	78.70	74.93

Efficacy of Diffusion-based Style Intervention. We conjecture the comprehensive style intervention is crucial for

Table 5. Ablations on the cross-modality lumbar spine segmentation (LSS) task (CT $\rightarrow$ MRI). All results are given as mean over 3 random initializations. The best result per column is in bold.

Method	Lumbar Spine CT $\rightarrow$ Lumbar Spine MRI					
	L1	L2	L3	L4	L5	Avg
BASELINE	0.00	0.00	0.00	0.00	0.25	0.05
BIG AUG [66]	0.23	0.60	0.40	0.60	0.20	0.43
GIN+IPA [40]	71.47	72.70	72.93	74.63	74.40	73.23
OURS	72.23	73.00	74.20	76.57	78.70	74.93

Table 6. Ablations on the cross-modality lumbar spine segmentation (LSS) task (CT $\rightarrow$ MRI). All results are given as mean over 3 random initializations. The best result per column is in bold.

Method	Lumbar Spine CT $\rightarrow$ Lumbar Spine MRI					
	L1	L2	L3	L4	L5	Avg
BASELINE	0.00	0.00	0.00	0.00	0.25	0.05
w/o STYLE SWAP	45.53	47.63	42.67	47.43	40.83	44.83
OURS	72.23	73.00	74.20	76.57	78.70	74.93

extracting content features. To validate this, we compare our diffusion-based style intervention with two alternatives: (1) basic photometric and geometric augmentations [66] discussed in Sec. 4.1 (BigAug), and (2) the global intensity non-linear augmentation and interventional pseudo-correlation augmentation (GIN+IPA) method proposed in [40]. Results in Tab. 4 demonstrate that basic augmentations alone are insufficient for cross-modality style changes, with InfoNCE loss providing only a 0.38 pp improvement in average Dice score. Our method surpasses

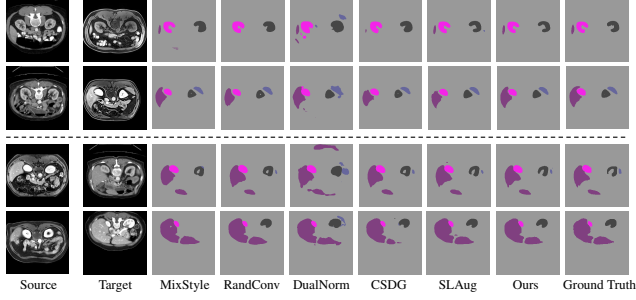


Figure 3. Visualization of segmentation results on the AS task.. First two rows: “CT to MRI” task; Last two rows: “MRI to CT”.

the augmentation strategies proposed in CSDG [40], the second-best method for the LSS task (CT→MRI), by 1.7 pp. This improvement demonstrates the effectiveness of using a controlled SD model for style intervention.

We also visualize our diffusion-generated augmentations in Fig. 6a, showcasing their close resemblance to target domain styles. The augmentations generated with the style prompt “MRI” accurately replicate the high soft-tissue contrast characteristic of lumbar spine MRI, distinguishing vertebral bodies, intervertebral discs, and the spinal cord. The spinal cord’s visibility, particularly its bright appearance on T2-weighted MRI, is difficult to simulate with generic image-level transformations or random intensity changes on CT scans [40, 56, 70]. Conversely, the X-Ray augmentations effectively capture the high-contrast appearance and edge enhancement typical of radiographic imaging, particularly along vertebral body contours. However, our method inherits ControlNet’s limitations in handling fine-grained structures. Fig. 6b illustrates this for the LSS task (source domain CT), where generated augmentations satisfy conditions on large segmentation regions-of-interests (*i.e.*, vertebral body) but struggle with smaller details like the posterior vertebral arch. This unintended alteration of content (*i.e.*, segmentation masks) could have a potential negative impact on the generalization performance.

Efficacy of Style Swap. To study the efficacy of the Style Swap technique [22], we directly train an additional ControlNet with U^B , omitting the instance fine-tuning stage with DreamBooth [49]. Results in Tab. 6 show that bypassing the fine-tuning stage leads to average Dice score drop of 30.10 pp. These findings support our hypothesis that the style prior stored in U^B may be compromised due to a mismatch between its prior information and the domain knowledge of D_0 , which leads to insufficient simulation of cross-modal domain shifts and, consequently, worse generalization performance.

5. Conclusion

In this paper, we address SDG for cross-modality medical image segmentation from a causal perspective and intro-

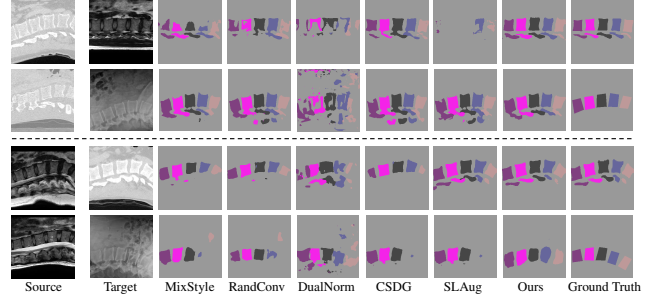


Figure 4. Visualization of segmentation results on the LSS task. First row: “CT to MRI”; second row: “CT to X-Ray”; third row: “MRI to CT”; and last row: “MRI to X-Ray”.

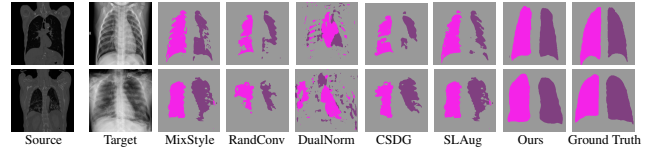


Figure 5. Visualization of segmentation results on the LS task. First two rows: “CT to X-Ray”.

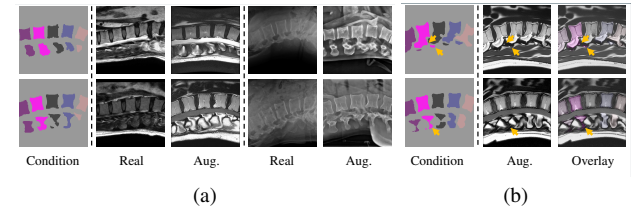


Figure 6. Visualization of diffusion-based style intervention for the LSS task. (a) Augmentation (Aug.) generated with style prompt “MRI” and “X-Ray”. (b) Examples showing limitations.

duce a style intervention method that simulates appearances from various imaging modalities utilizing large-scale pre-trained SD model with ControlNet. Experiments on three segmentation tasks across different anatomies and imaging modalities show that our approach consistently outperforms state-of-the-art SDG methods for medical imaging.

However, our method inherits the limitations of the ControlNet, which struggles to generate images corresponding to fine-grained details in the conditional segmentation masks. Thus, a more refined control on anatomical structure such as [26] is desirable for the exact correspondence between generated anatomical structures and given segmentation masks. Alternatively, we can incorporate uncertainty estimation methods [12, 57] into our framework to address the noisy labels resulting from the coarse control.

Acknowledgements

We gratefully acknowledge funding from the ETH AI Center, the Swiss Federal Institutes of Technology strategic focus area of personalized health and related technologies project 2022-643, and the ETH core funding.

References

- [1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. [1](#)
- [2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. [2](#), [3](#)
- [3] Jin Chen, Zhi Gao, Xinxiao Wu, and Jiebo Luo. Meta-causal learning for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7683–7692, 2023. [1](#), [2](#), [3](#)
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. [6](#)
- [5] Yun-Chun Chen, Yen-Yu Lin, Ming-Hsuan Yang, and Jia-Bin Huang. Crdoco: Pixel-level domain transfer with cross-domain consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1791–1800, 2019. [1](#)
- [6] Viacheslav V Danilov, Diana Litmanovich, Alex Proutski, Alexander Kirpich, Dato Nefaridze, Alex Karpovsky, and Yuriy Gankin. Automatic scoring of covid-19 severity in x-ray imaging based on a novel deep learning workflow. *Scientific reports*, 12(1):12791, 2022. [5](#)
- [7] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. [2](#)
- [8] Michal Drozdal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *International workshop on deep learning in medical image analysis, international workshop on large-scale annotation of biomedical data and expert label synthesis*, pages 179–187. Springer, 2016. [3](#), [5](#)
- [9] Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. Learning to learn with variational information bottleneck for domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 200–216. Springer, 2020. [1](#), [2](#)
- [10] Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, and Wen-Ming Ye. Data augmentation for object detection via controllable diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1257–1266, 2024. [3](#)
- [11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. [5](#)
- [12] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016. [8](#)
- [13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. [1](#)
- [14] Rui Gong, Martin Danelljan, Han Sun, Julio Delgado Mangas, and Luc Van Gool. Prompting diffusion representations for cross-domain semantic segmentation. *arXiv preprint arXiv:2307.02138*, 2023. [2](#)
- [15] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdif: Compact parameter space for diffusion fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7323–7334, 2023. [5](#)
- [16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. [3](#), [4](#)
- [17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. [1](#)
- [18] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. [2](#)
- [19] Maximilian Ilse, Jakub M Tomczak, and Patrick Forré. Selecting data augmentation for simulating interventions. In *International conference on machine learning*, pages 4555–4562. PMLR, 2021. [2](#), [3](#)
- [20] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018. [1](#)
- [21] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. [5](#)
- [22] Yuru Jia, Lukas Hoyer, Shengyu Huang, Tianfu Wang, Luc Van Gool, Konrad Schindler, and Anton Obukhov. Dgin-style: Domain-generalizable semantic segmentation with image diffusion models and stylized semantic control. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2023. [2](#), [5](#), [8](#)
- [23] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In *Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25*, pages 597–609. Springer, 2017. [1](#)
- [24] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. [3](#)
- [25] Podchara Klinwichit, Watcharaphong Yookwan, Sornsupha Limchareon, Krisana Chinnasarn, Jun-Su Jang, and Athita

- Onuean. Buu-lspine: A thai open lumbar spine dataset for spondylolisthesis detection. *Applied Sciences*, 13(15):8646, 2023. [5](#)
- [26] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. *arXiv preprint arXiv:2402.05210*, 2024. [8](#)
- [27] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015. [5](#)
- [28] TASET SHIFT IN MACHINE LEARNING. Dataset shift in machine learning. [2](#)
- [29] Lei Li, Ke Gao, Juan Cao, Ziyao Huang, Yepeng Weng, Xiaoyue Mi, Zhengze Yu, Xiaoya Li, and Boyang Xia. Progressive domain expansion network for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 224–233, 2021. [2](#)
- [30] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018. [1](#)
- [31] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8046–8056, 2022. [1](#), [3](#)
- [32] Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International conference on machine learning*, pages 7313–7324. PMLR, 2021. [1](#), [2](#), [3](#)
- [33] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. [1](#)
- [34] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *arXiv preprint arXiv:2010.07922*, 2020. [3](#)
- [35] Saeid Motiian, Marco Piccirilli, Donald A Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5715–5725, 2017. [1](#), [2](#)
- [36] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International conference on machine learning*, pages 10–18. PMLR, 2013. [1](#), [2](#), [3](#)
- [37] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyunghyun Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4500–4509, 2018. [1](#)
- [38] Toan Nguyen, Kien Do, Duc Thanh Nguyen, Bao Duong, and Thin Nguyen. Causal inference via style transfer for out-of-distribution generalisation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1746–1757, 2023. [1](#)
- [39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. [3](#)
- [40] Cheng Ouyang, Chen Chen, Surui Li, Zeju Li, Chen Qin, Wenjia Bai, and Daniel Rueckert. Causality-inspired single-source domain generalization for medical image segmentation. *IEEE Transactions on Medical Imaging*, 42(4):1095–1106, 2022. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)
- [41] Shumao Pang, Chunlan Pang, Lei Zhao, Yangfan Chen, Zhihai Su, Yujia Zhou, Meiyang Huang, Wei Yang, Hai Lu, and Qianjin Feng. Spineparsenet: spine parsing for volumetric mr image by a two-stage segmentation framework with semantic image representation. *IEEE Transactions on Medical Imaging*, 40(1):262–273, 2020. [5](#)
- [42] Judea Pearl. *Causality*. Cambridge university press, 2009. [3](#)
- [43] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018. [5](#)
- [44] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016. [3](#)
- [45] Fengchun Qiao, Long Zhao, and Xi Peng. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12556–12565, 2020. [2](#)
- [46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. [3](#), [4](#), [5](#)
- [47] Eduardo Romera, Luis M Bergasa, Jose M Alvarez, and Mohan Trivedi. Train here, deploy there: Robust segmentation in unseen domains. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1828–1833. IEEE, 2018. [2](#)
- [48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. [1](#), [3](#), [5](#)
- [49] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference*

- on computer vision and pattern recognition, pages 22500–22510, 2023. 5, 8
- [50] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anti-causal learning. *arXiv preprint arXiv:1206.6471*, 2012. 2
 - [51] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 2
 - [52] Anjany Sekuboyina, Markus Rempfler, Alexander Valentinsch, Bjoern H Menze, and Jan S Kirschke. Labeling vertebrae with two-dimensional reformations of multidetector ct images: an adversarial approach for incorporating prior knowledge of spine anatomy. *Radiology: Artificial Intelligence*, 2(2):e190074, 2020. 5
 - [53] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3
 - [54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4
 - [55] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3
 - [56] Zixian Su, Kai Yao, Xi Yang, Kaizhu Huang, Qiufeng Wang, and Jie Sun. Rethinking data augmentation for single-source domain generalization in medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2366–2374, 2023. 2, 6, 7, 8
 - [57] Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. *Advances in neural information processing systems*, 32, 2019. 8
 - [58] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 3
 - [59] Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021. 2, 3
 - [60] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 5
 - [61] Zijian Wang, Yadan Luo, Ruihong Qiu, Zi Huang, and Mahsa Baktashmotlagh. Learning to diversify for single domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 834–843, 2021. 2
 - [62] Zhenlin Xu, Deyi Liu, Junlin Yang, Colin Raffel, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *arXiv preprint arXiv:2007.13003*, 2020. 2, 6, 7
 - [63] Jinzhong Yang, Greg Sharp, Harini Veeraraghavan, Wouter Van Elmpt, Andre Dekker, Tim Lustberg, and Mark Gooding. Data from lung ct segmentation challenge. *The cancer imaging archive*, 2017. 5
 - [64] Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. *Advances in Neural Information Processing Systems*, 33:289–301, 2020. 2
 - [65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4
 - [66] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Bradford J Wood, Holger Roth, Andriy Myronenko, Daguang Xu, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging*, 39(7):2531–2540, 2020. 2, 5, 7
 - [67] Yiheng Zhang, Zhaofan Qiu, Ting Yao, Dong Liu, and Tao Mei. Fully convolutional adaptation networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6810–6818, 2018. 1
 - [68] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *Advances in Neural Information Processing Systems*, 33:14435–14447, 2020. 2
 - [69] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021. 5, 6, 7
 - [70] Ziqi Zhou, Lei Qi, Xin Yang, Dong Ni, and Yinghuan Shi. Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In *CVPR*, 2022. 2, 6, 7, 8
 - [71] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming 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: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018. 1
 - [72] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. Models genesis: Generic autodidactic models for 3d medical image analysis. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22*, pages 384–393. Springer, 2019. 6