# Train Once, Adapt Anywhere: Edge-Guided Single-source Domain Generalization for Medical Image Segmentation

**Jun Jiang** [1] JUNJIANG.STEVE@GMAIL.COM
[1] *Shenzhen Institute for Advanced Study, UESTC, Shenzhen, China*
**Shi Gu** [1,2] GUS@UESTC.EDU.CN
[2] *School of Computer Science and Engineering, UESTC, Chengdu, China*

**Editors:** Under Review for MIDL 2024

## Abstract

Multi-source domain generalization in medical image analysis depends on multi-source data to learn domain-invariant representations. This leads to the high cost of data collection. In this work, we take the "Train Once, Adapt Anywhere" as our objective and consider a challenging but practical setting: Single-source Domain Generalization (SDG). Meanwhile, we note that (i) the medical image segmentation applications where generalization errors often come from imprecise predictions at the ambiguous boundary of anatomies and (ii) the edge of the image is domain-invariant, which can reduce the domain shift between the source and target domain in all network layers. Specifically, we borrow the prior knowledge from Digital Image Processing and take the edge of the image as input to enhance the model attention at the boundary of anatomies and improve the generalization performance on unknown target domains. Extensive experiments on three typical medical image segmentation datasets, which cover cross-sequence, cross-center, and cross-modality settings with various anatomical structures, demonstrate our method achieves superior generalization performance compared to the state-of-the-art SDG methods. The code is available at https://github.com/thinkdifferentor/EGSDG.

**Keywords:** Domain Generalization, Transfer Learning, Medical Image Segmentation.

## 1. Introduction

Deep segmentation networks have achieved amazing progress for medical image analysis in recent years (Butoi et al., 2023; Isensee et al., 2021). However, when there is domain shift between the training and testing data, the performance of data-driven deep models degrades dramatically, like scanning technique, acquisition parameters, device manufacturers, etc. Recently, many efforts of Domain Generalization (DG) and Unsupervised Domain Adaptation (UDA) have been made to improve the model's generalization ability on the target domain (Su et al., 2023; Feng et al., 2023). Further, domain generalization can be divided into Multi-source Domain Generalization (MDG) and Single-source Domain Generalization (SDG). On the one hand, MDG models (Dou et al., 2019; Li et al., 2019) are designed with multiple source domains to learn domain-invariant representations, and it may not work when there is only a single-source domain. On the other hand, the high cost of medical image labeling and the strict regulations of privacy protection make it difficult to obtain large amounts of medical data. Besides, previous UDA works (Feng et al., 2023; Chen et al., 2019) require retraining when receiving the samples from a new data distribution, which leads to the high cost of model deployment. To this end, we take the

© 2024 CC-BY 4.0, J.J. & S.G. .

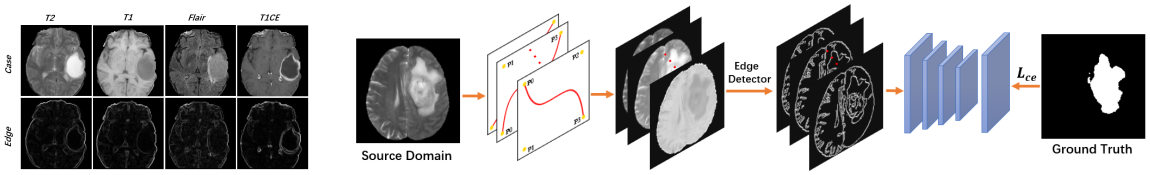


Figure 1: Example case with corresponding Sobel gradient map of BraTS'19 samples, which can be used as domain-invariant information to guide the training process (left). Overview of our proposed *Edge-Guided Single-source Domain Generalization* (EGSDG) for medical image segmentation (right). Firstly, we perform BézierCurve data augmentation on the single-source domain. Then, the edge detector is used to extract edge maps of the augmented samples. Finally, we take the edge maps as input to train the segmentation network with cross-entropy loss $L_{ce}$. Note that we first utilize the edge detector to extract the edge of the test sample and feed it into the model for prediction.

"*Train Once, Adapt Anywhere*" as our objective and consider a challenging but practical setting: single-source domain generalization.

For segmentation tasks, models often make inaccurate predictions at target boundaries. This weakness is more pronounced for domain generalization or adaptation segmentation tasks in medical images due to the domain gap and ambiguous boundary of anatomies. Recently, some works have proposed corresponding solutions and made significant progress (Liu et al., 2022; You et al., 2023; Feng et al., 2023). However, there are several limitations to them. First, they do not directly take the image edge as the model input, which terribly weakens the supervision of edge information during the training process (Details refer to Appendix A). Second, they acquire the image edge or shape priors by learning, which will take lots of training time. Based on these insights, we employ the edge detection algorithm to get image edge maps and use them as input to train the model directly. This simple yet effective way filters domain-specific information, and improving the generalization ability significantly. To the best of our knowledge, we make a first attempt to employ the edge of image as input to train the network. The major contributions of this work are as follows:

- We make a comprehensive analysis to the impact of image edge on the model generalization ability, including the positions of edge supervision signals, such as shallow, deep, or output layers; the fusion strategies of edge map and feature map, such as concatenating it with shallow or deep features and directly as network input.

- For the challenging yet essential SDG problem of medical image segmentation, we propose a novel approach, which significantly improves the generalization performance on unknown target domains.

- We conducted extensive experiments on three typical medical image segmentation datasets that cover various anatomical structures. With only a single-source domain, our method achieves superior generalization performance on the unknown target domain compared to the state-of-the-art SDG methods.

## 2. Related Works

### 2.1. SDG of Nature Image

SDG models of nature images can be divided into two main methods: (1) *the image-level method*, which improve the model generalization by data augmentation with the help of existing large datasets (e.g. ImageNet) (Yue et al., 2019; Lee et al., 2022), and (2) *the feature-level method*, which aims to learn domain-invariant segmentation network by removing the style information of feature maps with normalization or whitening strategy (Choi et al., 2021; Peng et al., 2022). However, these models may not work well on grayscale medical images, because there are significant differences in texture, structure, and data privacy policies between medical and natural images.

### 2.2. SDG of Medical Image

In medical images, there are fewer works on the SDG segmentation task compared to natural images. Most of these models (Liu et al., 2022; Ouyang et al., 2021; Su et al., 2023) conduct data augmentation on the source domain to improve the model's robustness. Different from the previous works, we introduce an edge-guided model, which filters the domain-specific information effectively and improves the generalization ability significantly.

### 2.3. Edge-Guided Methods

Recently, many efforts of edge-guided methods have been made to raise attention to the segmentation boundary and improve the generalization ability of models. Cardace et al. (Cardace et al., 2021) presented a novel low-level adaptation strategy with semantic edges and displacement maps from shallow features to obtain sharp predictions. CIConv (Lengyel et al., 2021) exploited a visual inductive prior derived from physics-based reflection models and cast a number of color-invariant edge detectors as trainable layers for domain adaptation. In our work, we utilize the edge detector to extract edge maps of images and take them as input to train the model directly.

## 3. Methodology

### 3.1. Preliminaries

Edge detector significantly filters out useless information, while preserving the important structural properties of an image. There are a large number of edge detection algorithms available, each designed to be sensitive to certain types of edges like edge orientation, noise environment, and edge structure. We take the most classic ones for exploring, including Canny (Canny, 1986), AutoCanny (Rong et al., 2014), Roberts (Roberts, 1963), Prewitt (Prewitt et al., 1970), Sobel (Kittler, 1983), and Laplacian (LeCun et al., 1998). For an image, its gradient at $(x, y)$ is defined as vector $\bigtriangledown f(x, y)$, which is composed of the partial derivative of the image in the X and Y directions:

$$\bigtriangledown f(x, y) = [G_x, G_y] = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right] \tag{1}$$

Table 1: Quantitative comparison of different methods on BraTS'19 (left) and Prostate (right) datasets. Note that CIConv* indicates training with BézierCurve augmented dataset and the result of SADN is reported by that method on BraTS'18 dataset.

| Method | Source Domain: T2 | | | |
|---|---|---|---|---|
| | T1 | T1ce | Flair | Average |
| **Upper Bound** | 74.42 | 71.64 | 82.75 | 76.27 |
| **Lower Bound** | 13.82 | 11.58 | 66.61 | 30.67 |
| **IBN-Net** | 34.37 | 48.27 | 42.33 | 41.66 |
| **SW** | 31.83 | 40.48 | 34.95 | 35.75 |
| **RobustNet** | 8.59 | 10.14 | 68.29 | 29.01 |
| **SADN** | 49.36 | 38.09 | 75.87 | 54.44 |
| **CSDG** | 46.76 | 44.99 | 60.20 | 50.65 |
| **CIConv** | 15.36 | 20.83 | 76.07 | 37.42 |
| **CIConv*** | 53.82 | 52.69 | 74.05 | 60.19 |
| **BasicAug** | 17.57 | 27.13 | 74.11 | 39.60 |
| **Cutout** | 20.76 | 24.48 | 76.44 | 40.56 |
| **RandConv** | 16.93 | 29.64 | 65.05 | 37.21 |
| **BezierCurve** | 52.10 | 43.12 | 70.63 | 55.28 |
| **EGSDG (ours)** | **62.59** | **54.68** | **77.07** | **64.78** |

| Method | Source Domain: Site B | | | | | |
|---|---|---|---|---|---|---|
| | Site A | Site C | Site D | Site E | Site F | Average |
| **Upper Bound** | 89.13 | 89.96 | 89.31 | 87.76 | 89.34 | 89.10 |
| **Lower Bound** | 63.62 | 19.42 | 81.06 | 83.89 | 71.17 | 63.83 |
| **IBN-Net** | 67.36 | 46.79 | 65.09 | 71.45 | 76.88 | 65.51 |
| **SW** | 70.83 | 51.71 | 70.89 | 51.96 | 68.97 | 62.87 |
| **RobustNet** | 73.27 | 55.04 | 77.41 | 54.79 | 70.21 | 66.14 |
| **CSDG** | 69.75 | 61.47 | 74.27 | 76.31 | 70.54 | 70.47 |
| **CIConv** | 73.48 | 63.51 | 80.80 | 62.15 | 74.93 | 70.97 |
| **CIConv*** | 76.41 | 59.74 | 76.63 | **78.10** | 77.17 | 73.61 |
| **BasicAug** | 71.32 | 65.75 | 81.71 | 66.90 | 72.44 | 71.62 |
| **Cutout** | 63.25 | 59.78 | **88.57** | 46.82 | 76.87 | 67.06 |
| **RandConv** | 45.96 | 48.16 | 30.95 | 61.35 | 39.70 | 45.22 |
| **BezierCurve** | 75.45 | **67.98** | 69.09 | 69.90 | 60.77 | 68.64 |
| **EGSDG (ours)** | **78.51** | 64.16 | 82.95 | 77.34 | **78.20** | **76.23** |

The modulus and direction of the gradient are defined by:

$$|\bigtriangledown f\,(x,y)| = \sqrt{G_x^2 + G_y^2}, \theta(x,y) = \arctan(\frac{G_y}{G_x}) \tag{2}$$

For Laplacian, the second derivative is defined as:

$$\bigtriangledown^2 f\,(x,y) = \frac{\partial^2 f(x,y)}{\partial x^2} + \frac{\partial^2 f(x,y)}{\partial y^2} \tag{3}$$

Note that the digital images are discrete and different edge detection algorithms differ in the way of $G_x$ and $G_y$ calculation, which are provided in Appendix B. The details of Canny and AutoCanny algorithms can be found in Appendix C and (Rong et al., 2014). Compared to Canny algorithm, AutoCanny does not need to manually set Gaussian smoothing parameters and the double thresholds.

### 3.2. Problem Definition

In single-source domain generalization, we are given a single source domain $D^s = \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where $s$ represents the domain ID, $x_i^s \in \mathbb{R}^{H\times W\times 3}$ is the $i$-th image in the source domain $s$. $y_i^s \in \mathbb{R}^{H\times W}$ is the corresponding ground truth mask, and $N_s$ is the total number of samples. Given unseen target domain $D^t = \left\{x_i^t, y_i^t\right\}_{i=1}^{N_t}$, which is not accessible during the training process, we aim to minimize the error between prediction $\hat{y}_i^t$ and ground truth mask $y_i^t$.

### 3.3. Edge-Guided SDG

Edge or gradient information is one of the most important features of an image. The edge of image is domain-invariant, which can reduce the domain shift between the source and target domain in all network layers (Lengyel et al., 2021). Different from previous works, TASD (Liu et al., 2022) establishes the dictionary learning to extract the explicit shape priors and CIConv (Lengyel et al., 2021) derived from the complex Kubelka-Munk theory to

Table 2: Quantitative comparison of different methods on MMWHS dataset. Note that CIConv* indicates training with the BézierCurve augmented dataset and the result of SADN is reported by that method.

| Method | Source Domain: MRI | | | | | Source Domain: CT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **AA** | **LAC** | **LVC** | **MYO** | **Average** | **AA** | **LAC** | **LVC** | **MYO** | **Average** |
| **Upper Bound** | 89.74 | 84.99 | 87.44 | 83.34 | 86.38 | 80.76 | 82.29 | 92.38 | 78.23 | 83.42 |
| **Lower Bound** | 32.18 | 35.92 | 19.53 | 9.42 | 24.26 | 18.44 | 8.84 | 38.72 | 9.65 | 18.91 |
| **IBN-Net** | 59.04 | 67.63 | 67.34 | 45.49 | 59.88 | 31.23 | 42.36 | 59.91 | 34.63 | 42.03 |
| **SW** | 52.94 | 69.52 | 64.28 | 44.64 | 57.84 | 38.95 | 47.62 | 62.82 | 33.30 | 45.67 |
| **RobustNet** | 68.07 | 74.68 | 62.56 | 46.09 | 62.85 | 52.27 | **60.08** | 67.26 | 32.97 | 53.14 |
| **SADN** | 51.42 | 50.20 | 52.86 | 52.31 | 51.70 | 33.38 | 31.65 | 33.29 | 30.45 | 32.19 |
| **CSDG** | 66.91 | 68.06 | 64.43 | 52.24 | 62.91 | 37.10 | 51.76 | 70.64 | 41.38 | 50.22 |
| **CIConv** | 67.42 | 70.83 | 65.19 | 42.77 | 61.55 | 45.40 | 45.38 | 57.08 | 32.44 | 45.08 |
| **CIConv*** | **78.38** | 75.67 | 69.33 | 55.92 | 69.83 | 45.75 | 50.64 | 71.93 | 35.33 | 50.91 |
| **BasicAug** | 68.03 | 61.91 | 54.68 | 34.75 | 54.84 | 34.68 | 54.46 | 42.79 | 20.75 | 38.17 |
| **Cutout** | 61.02 | 71.10 | 53.17 | 40.21 | 56.38 | 27.91 | 44.90 | 32.11 | 11.85 | 29.19 |
| **RandConv** | 66.64 | 61.26 | 50.80 | 37.24 | 53.98 | 37.67 | 42.70 | 43.22 | 21.53 | 36.28 |
| **BezierCurve** | 64.78 | 63.62 | 67.65 | 41.81 | 59.46 | 43.56 | 46.82 | 51.72 | 23.84 | 41.48 |
| **EGSDG (ours)** | 73.45 | **78.48** | **71.94** | **60.13** | **71.00** | **55.14** | 57.34 | **72.50** | **45.84** | **57.71** |

build a learnable edge detector layer, our model is borrowed from the classical edge detection algorithm with less computational complexity and more stable performance. Visualization comparison of classical edge detection algorithms and CIConv refer to Appendix D.

In addition, data augmentation can enrich the gradient information of the training samples, which will bring huge performance gains to our edge-guided model's generalization ability (Details refer to Appendix E and F). For medical images, we expect to map the source image to diverse grayscale value distribution and keep the appearance of the anatomic structures perceivable at the same time. Motivated by Model Genesis (Zhou et al., 2019), we employ the Bézier Curve (Mortenson, 1999) as our data augmentation method, which is generated from two end points ($P_0$ and $P_3$) and two control points ($P_1$ and $P_2$), defined as:

$$B(t) = \sum_{i=0}^{n} \binom{n}{i} P_i (1-t)^{n-i} t^i, n = 3, t \in [0, 1] \tag{4}$$

where $t$ is a fractional value along the length of the line.

The learning process of our EGSDG is shown in Figure 1. Firstly, we perform the BézierCurve data augmentation on source samples ($x_i^s \in \mathbb{R}^{H \times W \times 3}$) before the training stage. Then, the edge detector is exploited to extract the edge maps ($e_i^s \in \mathbb{R}^{H \times W}$) of the augmented samples. Finally, we take the edge maps $e_i^s$ as input to train the segmentation network $\phi^w$ with parameters $w$ by minimizing cross-entropy loss:

$$\mathcal{L}_{ce}(\phi^w; D^s) = -\sum_i [y_i^s, log(\phi^w(e_i^s))] \tag{5}$$

We use the edge detector to compress the image into a single-channel edge map. It effectively filters domain-specific information and trains a model with high generalization performance. The network locates the segmentation object by the gradient change (Roberts, Prewitt, Sobel, and Laplacian) or the anatomy contour (Canny and AutoCanny) of the target area.

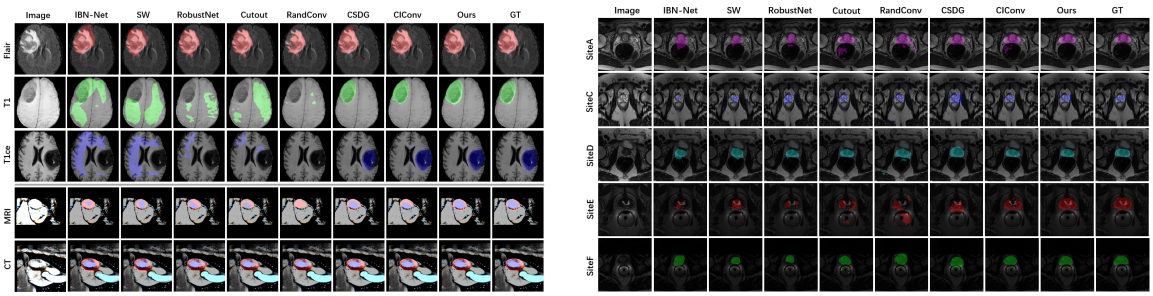


Figure 2: Qualitative comparison of BraTS'19 (top) and MMWHS (bottom) samples segmentation (left) and Prostate samples segmentation (right). MRI means CT→MRI domain generalization and CT means MRI→CT domain generalization.

## 4. Experiments and Results

### 4.1. Experimental Setup

**Datasets and Preprocessing** In our experiments, we employ three datasets, the cross-sequence brain tumor segmentation dataset (BraTS'19, T2 as source domain) (Menze et al., 2015), the cross-center prostate dataset (Prostate, Site B as source domain) (Liu et al., 2020), and the cross-modality cardiac dataset (MMWHS, CT and MRI as source domain respectively) (Zhuang and Shen, 2016), for evaluation. Details are given in Appendix G.

**Network and Training Details** Following CSDG (Ouyang et al., 2021), we utilize U-Net (Ronneberger et al., 2015) with an EffcientNet-b2 (Tan and Le, 2019) backbone as our segmentation model and implement our model by PyTorch framework on one NVIDIA TITAN XP GPU. We use Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of $3 \times 10^{-4}$ and batch size of 8 to train the model. For all experiments, the learning rate is decayed according to the polynomial rule for stable training.

**Evaluation Metrics** We take the Dice coefficient (Dice) as our evaluation metric, which measures the overlapping ratio between prediction and ground truth. The higher the Dice value, the better the segmentation performance.

### 4.2. Comparison Experiments

We compare our method with SOTA single-source domain generalization methods (IBN-Net (Pan et al., 2018), SW (Pan et al., 2019), RobustNet (Choi et al., 2021), SADN (Zhou et al., 2022), CSDG (Ouyang et al., 2021), and CIConv (Lengyel et al., 2021)) and the data augmentation methods (BasicAug, Cutout (Devries and Taylor, 2017), RandConv (Xu et al., 2020), and BézierCurve (Zhou et al., 2019). Note that BasicAug indicates training the model with ColorJitter, GaussianBlur, RandomGrayscale, etc. basic photometric data augmentations. For a fair comparison, we employ the same segmentation network to train the CIConv model. Besides, we also provide the results without domain generalization by directly applying the model learned in the source domain to unknown target domains (Lower Bound) and with supervised training on the target domain (Upper Bound). In addition, the comprehensive comparison between CIConv and ours is given in Appendix H.

Table 3: Ablation study of edge detectors with original (first group) and BézierCurve augmented (second group†) sample on the BraTS'19 (left) and the Prostate (right) datasets.

| Edge Detector | Source Domain: T2 | | | |
|---|---|---|---|---|
| | T1 | T1ce | Flair | Average |
| **Canny** | 40.30 | 50.18 | 61.96 | 50.81 |
| **AutoCanny** | 48.67 | 56.25 | 66.07 | 57.0 |
| **Roberts** | 43.70 | 48.93 | 67.40 | 53.34 |
| **Prewitt** | 50.28 | 50.31 | 72.38 | 57.66 |
| **Sobel** | 51.38 | 50.35 | 71.63 | **57.79** |
| **Laplacian** | 31.39 | 43.39 | 61.86 | 45.55 |
| **Canny†** | 56.58 | 53.48 | 62.39 | 57.48 |
| **AutoCanny†** | 56.28 | 53.76 | 63.84 | 57.96 |
| **Roberts†** | 58.66 | 55.60 | 68.84 | 61.03 |
| **Prewitt†** | 55.26 | 55.82 | 72.59 | 61.22 |
| **Sobel†** | 62.59 | 54.68 | 77.07 | **64.78** |
| **Laplacian†** | 54.55 | 55.78 | 58.68 | 56.34 |

| Edge Detector | Source Domain: Site B | | | | | |
|---|---|---|---|---|---|---|
| | Site A | Site C | Site D | Site E | Site F | Average |
| **Canny** | 72.12 | 46.13 | 64.82 | 63.93 | 62.74 | 61.95 |
| **AutoCanny** | 72.70 | 59.54 | 83.00 | 70.36 | 81.11 | **73.34** |
| **Roberts** | 68.04 | 49.73 | 75.93 | 71.82 | 78.79 | 68.86 |
| **Prewitt** | 73.39 | 48.83 | 81.08 | 80.27 | 69.05 | 70.52 |
| **Sobel** | 73.35 | 48.36 | 84.13 | 79.95 | 71.07 | 71.37 |
| **Laplacian** | 73.48 | 50.19 | 81.20 | 79.92 | 81.74 | 73.31 |
| **Canny†** | 66.28 | 56.55 | 58.59 | 70.29 | 66.43 | 63.63 |
| **AutoCanny†** | 78.51 | 64.16 | 82.95 | 77.34 | 78.20 | **76.23** |
| **Roberts†** | 75.23 | 57.62 | 80.43 | 79.45 | 71.26 | 72.80 |
| **Prewitt†** | 71.40 | 56.59 | 76.44 | 79.50 | 77.95 | 72.38 |
| **Sobel†** | 75.48 | 55.87 | 75.53 | 84.10 | 70.60 | 72.32 |
| **Laplacian†** | 75.66 | 42.04 | 85.28 | 83.30 | 82.89 | 73.83 |

Table 1 and Table 2 report the comparison results on the BraTS'19, Prostate, and MMWHS datasets respectively. Overall, our model outperforms others, especially in the large distribution shift dataset (BraTS'19 and MMWHS). For the results of normalization and whitening-based models (IBN-Net, SW, and RobustNet), which are designed for nature image, their performance is significantly lower than our model in each evaluation dataset. For the results of data augmentation-based methods (SADN, CSDG, Cutout, and RandConv), their performance is unstable on different datasets. The distribution of grayscale values varies across different datasets and the level of domain shift varies among different SDG segmentation settings. However, the augmented samples fail to cover the unseen target domain distribution on the specific task and lead to terrible generalization performance. In addition, the performance of CIConv is lower than our model on all three datasets, and it takes longer training time compared to ours. Qualitative examples are shown in Figure 2. As we can see, our model can produce accurate and sharp predictions at the boundary of anatomies.

### 4.3. Ablation Experiments

**Data Augmentation** For edge-guided models, data augmentation is supposed to simulate the edge or gradient information of the unseen target samples to train a model with great generalization ability. To this end, we explore Cutout (Devries and Taylor, 2017), RandConv (Xu et al., 2020), and BézierCurve (Zhou et al., 2019), which are widely used data augmentation methods in SDG settings. Besides, we introduce BasicAug as a comparison. At the results of the data augmentation methods group in Table 1, we can find that BézierCurve gets better results than others on BraTS'19 and MMWHS datasets and achieves superior performance on the BraTS'19. Besides, compared with other data augmentation methods, BézierCurve can conduct augmentation in data preprocessing. It significantly reduces the training time compared to others (Cutout and RandConv), which makes augmentation at the training stage. To further validate the effectiveness of data augmentation to edge-guided models, the results of edge-guided with RandConv are provided in Appendix I.

Table 4: Ablation study of edge detectors with original (top) and BézierCurve augmented (bottom†) sample on the MMWHS dataset.

| Edge Detector | Source Domain: MRI | | | | | Source Domain: CT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AA | LAC | LVC | MYO | Average | AA | LAC | LVC | MYO | Average |
| **Canny** | 67.91 | 69.87 | 67.63 | 50.15 | 63.89 | 54.11 | 53.41 | 62.74 | 32.86 | **50.78** |
| **AutoCanny** | 65.22 | 71.51 | 64.22 | 51.58 | 63.13 | 44.96 | 56.68 | 58.79 | 34.93 | 48.84 |
| **Roberts** | 72.17 | 70.72 | 59.21 | 55.41 | 64.38 | 32.21 | 49.15 | 52.98 | 18.96 | 38.33 |
| **Prewitt** | 68.76 | 70.82 | 65.87 | 51.67 | 64.28 | 37.98 | 49.31 | 61.04 | 21.92 | 42.56 |
| **Sobel** | 73.67 | 72.45 | 57.31 | 57.42 | **65.21** | 40.80 | 54.55 | 64.92 | 21.94 | 45.55 |
| **Laplacian** | 67.43 | 72.02 | 62.72 | 56.52 | 64.67 | 36.27 | 48.94 | 74.07 | 30.79 | 47.52 |
| **Canny†** | 66.03 | 73.65 | 71.11 | 52.69 | 65.87 | 55.14 | 57.34 | 72.50 | 45.84 | **57.71** |
| **AutoCanny†** | 70.63 | 69.81 | 67.15 | 52.39 | 65.00 | 52.89 | 62.96 | 65.15 | 34.48 | 53.87 |
| **Roberts†** | 72.04 | 73.90 | 65.43 | 54.89 | 66.57 | 46.38 | 46.66 | 65.67 | 29.47 | 47.04 |
| **Prewitt†** | 71.05 | 75.04 | 68.31 | 55.07 | 67.37 | 49.40 | 54.18 | 60.27 | 32.12 | 48.99 |
| **Sobel†** | 73.45 | 78.48 | 71.94 | 60.13 | **71.00** | 35.57 | 52.98 | 60.45 | 31.96 | 45.24 |
| **Laplacian†** | 70.74 | 70.23 | 64.51 | 54.79 | 65.07 | 49.99 | 49.18 | 67.71 | 32.65 | 49.88 |

**Edge Detector** Different edge detectors will extract distinct image edges or gradients for the same image, which affects the training process and testing performance of the model. Accordingly, we conducted comprehensive comparison experiments on classical edge detectors on each dataset, including Canny (Canny, 1986), AutoCanny (Rong et al., 2014), Roberts (Roberts, 1963), Prewitt (Prewitt et al., 1970), Sobel (Kittler, 1983), and Laplacian (LeCun et al., 1998). In addition, we also explore the above edge-guided models on the BézierCurve augmented samples, which can enrich the gradient information of images. Table 3 and Table 4 show the results. As we can see, taking the edge as input promotes the generalization ability of the model remarkably and the BézierCurve can further improve its performance. We note that the optimal edge detection algorithm varies depending on the data distribution and this limitation will bring great challenges to model deployment.

## 5. Conclusion and Discussion

In medical applications, the SDG problem is critical and valuable. On the one hand, due to privacy protection policies, medical data is often quite limited at the training stage. On the other hand, the cost of retraining during the deployment phase is too expensive, and not every medical institution has skillful deep-learning engineers. In this work, we use the edge of image as input to train a network. It improves the model's generalization performance significantly on unknown target domains. Extensive experiments on three typical medical image segmentation datasets demonstrate our approach achieves superior generalization performance compared to the state-of-the-art SDG methods. Moreover, taking the image edge as model inputs can effectively protect the privacy of medical data. However, this compression method makes the grayscale value information completely lost, which is valuable for medical diagnosis. We have tried to combine the image or its features with the edge map in different strategies but the performance is lower than our model. Therefore, an effective integration method needs to be further explored. In addition, on different data sets, the optimal edge detection algorithm is different, which will bring great challenges to model deployment. We will consider employing the deep learning-based edge detector in an end-to-end fashion in future work.

## References

Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21438–21451, October 2023.

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

Adriano Cardace, Pierluigi Zama Ramirez, Samuele Salti, and Luigi di Stefano. Shallow features guide unsupervised domain adaptation for semantic segmentation at class boundaries. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2010–2020, 2021. URL https://api.semanticscholar.org/CorpusID:238407791.

Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 865–872, 2019.

Sungha Choi, Sanghun Jung, Huiwon Yun, Joanne Taery Kim, Seungryong Kim, and Jaegul Choo. Robustnet: Improving domain generalization in urban-scene segmentation via instance selective whitening. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11575–11585, 2021. URL https://api.semanticscholar.org/CorpusID:232404762.

Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *ArXiv*, abs/1708.04552, 2017. URL https://api.semanticscholar.org/CorpusID:23714201.

Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. Domain generalization via model-agnostic learning of semantic features. *Advances in neural information processing systems*, 32, 2019.

Wei Feng, Lie Ju, Lin Wang, Kaimin Song, Xin Zhao, and Zongyuan Ge. Unsupervised domain adaptation for medical image segmentation by selective entropy constraints and adaptive semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 623–631, 2023.

Jan-Mark Geusebroek, Rein van den Boomgaard, Arnold W. M. Smeulders, and Hugo Geerts. Color invariance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23:1338–1350, 2001. URL https://api.semanticscholar.org/CorpusID:14569112.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Josef Kittler. On the accuracy of the sobel edge detector. *Image and Vision Computing*, 1 (1):37–42, 1983.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Suhyeon Lee, Hongje Seong, Seongwon Lee, and Euntai Kim. Wildnet: Learning domain generalized semantic segmentation from the wild. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9926–9936, 2022. URL https://api.semanticscholar.org/CorpusID:247940228.

Attila Lengyel, Sourav Garg, Michael Milford, and Jan C. van Gemert. Zero-shot day-night domain adaptation with a physics prior. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4379–4389, 2021. URL https://api.semanticscholar.org/CorpusID:237248535.

Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019.

Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Medical Imaging*, 2020.

Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1013–1023, 2021.

Quande Liu, Cheng Chen, Qi Dou, and Pheng-Ann Heng. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1756–1764, 2022.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34:1993–2024, 2015. URL https://api.semanticscholar.org/CorpusID:1739295.

Michael E. Mortenson. Mathematics for computer graphics applications. 1999. URL https://api.semanticscholar.org/CorpusID:60038893.

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*, 42:1095–1106, 2021. URL https://api.semanticscholar.org/CorpusID:244527518.

Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *European Conference on Computer Vision*, 2018. URL https://api.semanticscholar.org/CorpusID:50781790.

Xingang Pan, Xiaohang Zhan, Jianping Shi, Xiaoou Tang, and Ping Luo. Switchable whitening for deep representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1863–1871, 2019. URL https://api.semanticscholar.org/CorpusID:128252253.

Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. Semantic-aware domain generalized segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2595, 2022. URL https://api.semanticscholar.org/CorpusID:247939688.

Judith MS Prewitt et al. Object enhancement and extraction. *Picture processing and Psychopictorics*, 10(1):15–19, 1970.

Lawrence G. Roberts. Machine perception of three-dimensional solids. In *Outstanding Dissertations in the Computer Sciences*, 1963. URL https://api.semanticscholar.org/CorpusID:19309211.

Weibin Rong, Zhanjing Li, Wei Zhang, and Lining Sun. An improved canny edge detection algorithm. In *2014 IEEE international conference on mechatronics and automation*, pages 577–582. IEEE, 2014.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. URL https://api.semanticscholar.org/CorpusID:3719281.

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.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *ArXiv*, abs/1905.11946, 2019. URL https://api.semanticscholar.org/CorpusID:167217261.

Zhenlin Xu, Deyi Liu, Junlin Yang, and Marc Niethammer. Robust and generalizable visual representation learning via random convolutions. *ArXiv*, abs/2007.13003, 2020. URL https://api.semanticscholar.org/CorpusID:220793552.

Xin You, Junjun He, Jie Yang, and Yun Gu. Learning with explicit shape priors for medical image segmentation. *arXiv preprint arXiv:2303.17967*, 2023.

Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto L. Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2100–2110, 2019. URL https://api.semanticscholar.org/CorpusID:202540251.

Ziqi Zhou, Lei Qi, Xin Yang, Dong Ni, and Yinghuan Shi. Generalizable cross-modality medical image segmentation via style augmentation and dual normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20856–20865, 2022.

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. *Medical image computing and computer-assisted intervention : MICCAI ... International Conference on Medical Image Computing and Computer-Assisted Intervention*, 11767:384–393, 2019. URL https://api.semanticscholar.org/CorpusID:201070166.

Xiahai Zhuang and Juan Shen. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical image analysis*, 31:77–87, 2016. URL https://api.semanticscholar.org/CorpusID:34028323.

## Appendix A. Comparison of edge-guided strategies

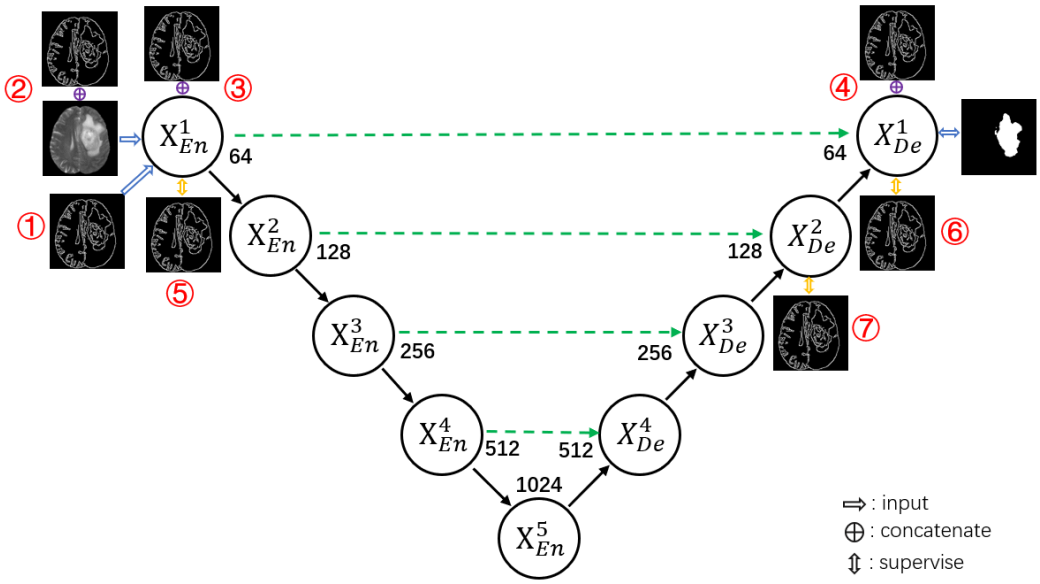


Figure 3: Visualization of different edge-guided strategies with vanilla U-Net framework. Exp. 1 means taking the image edge as input. Exp. 2 means concatenating the image and edge map as input. Exp. 3 means concatenating the feature map $X^1_{En}$ and the edge map as the next layer's input. Exp. 4 means concatenating the feature map $X^1_{De}$ and the edge map as the classifier layer's input. Exp. 5 means employing the image edge as the supervisor of feature map $X^1_{En}$ with a single Conv2d classifier layer. Exp. 6 means employing the image edge as the supervisor of feature map $X^1_{De}$ with a single Conv2d classifier layer. Exp. 7 means employing the image edge as the supervisor of feature map $X^2_{De}$ with a single Conv2d classifier layer. We employ AutoCanny as edge detector in all experiments.

To enhance the model's attention at the boundary of targets, previous works (Lengyel et al., 2021; Liu et al., 2021, 2022; You et al., 2023) have tried different strategies. Here, we conducted a comprehensive analysis to the impact of image edge on the model generalization ability, including the positions of edge supervision signals, such as shallow, deep, or output layers; the fusion strategies of edge feature and feature map, such as concatenating it with shallow or deep features and directly as network input.

We employ the vanilla U-Net (Ronneberger et al., 2015) as a segmentation network which is borrowed from https://github.com/milesial/Pytorch-UNet. Other experiment configurations are the same as the main experiment. Details refer to section 4.1 in the body. The visualization of different edge-guided strategies and corresponding explanations are provided in Figure 3. These edge-guided strategies can be divided into two categories: (i) using the edge map as a supervision signal to increase the model's attention at the boundary of segmentation targets. (ii) concatenating the feature map and edge map to force the model learning domain invariant representation and enhance the generalization ability. Note that

we take the same process at the testing stage for the first category experiments (Exp. 1, 2, 3, and 4), and AutoCanny (Rong et al., 2014) is used as the edge detector in all experiments.

Table 5 and Table 6 report the comparison results on BraTS'19, Prostate, and MMWHS datasets respectively. Overall, adopting the edge map as an additional guided signal can improve model generalization performance, compared to the Lower Bound (without domain generalization). Notably, employing the image edge as input to train the model directly brings tremendous generalization ability gains on the three datasets. As mentioned in the main text, these strategies which do not directly take the image edge as the model input terribly weaken the supervision of the edge information at the training stage.

Table 5: Ablation study of different edge-guided strategies on the BraTS'19 (left) and Prostate (right) datasets with vanilla U-Net.

| Experiments | Source Domain: T2 | | | |
|---|---|---|---|---|
| | T1 | T1ce | Flair | Average |
| **Lower Bound** | 10.72 | 5.86 | 58.34 | 24.97 |
| **Exp. 1 (Ours)** | 37.86 | 39.95 | 51.10 | **42.97** |
| **Exp. 2** | 11.65 | 11.42 | 61.02 | 28.03 |
| **Exp. 3** | 13.90 | 11.37 | 58.88 | 28.05 |
| **Exp. 4** | 11.58 | 9.22 | 59.48 | 26.76 |
| **Exp. 5** | 10.14 | 9.35 | 64.84 | 28.11 |
| **Exp. 6** | 13.28 | 10.57 | 58.93 | 27.59 |
| **Exp. 7** | 10.60 | 8.92 | 64.48 | 28.00 |

| Experiments | Source Domain: Site B | | | | | |
|---|---|---|---|---|---|---|
| | Site A | Site C | Site D | Site E | Site F | Average |
| **Lower Bound** | 42.25 | 25.79 | 59.91 | 14.88 | 37.12 | 35.99 |
| **Exp. 1 (Ours)** | 47.79 | 31.39 | 48.68 | 52.01 | 51.34 | **46.24** |
| **Exp. 2** | 58.65 | 42.94 | 46.88 | 12.81 | 31.87 | 38.63 |
| **Exp. 3** | 38.31 | 17.44 | 61.78 | 21.68 | 41.35 | 36.11 |
| **Exp. 4** | 32.33 | 24.74 | 43.97 | 32.14 | 36.14 | 33.86 |
| **Exp. 5** | 48.3 | 37.14 | 66.96 | 20.48 | 35.02 | 41.58 |
| **Exp. 6** | 43.56 | 25.73 | 78.45 | 20.3 | 55.23 | 44.65 |
| **Exp. 7** | 51.87 | 27.15 | 61.72 | 18.16 | 49.23 | 41.63 |

Table 6: Ablation study of different edge-guided strategies on the MMWHS dataset with vanilla U-Net.

| Experiments | Source Domain: MRI | | | | | Source Domain: CT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AA | LAC | LVC | MYO | Average | AA | LAC | LVC | MYO | Average |
| **Lower Bound** | 22.60 | 36.71 | 23.51 | 13.08 | 23.98 | 12.84 | 37.94 | 23.44 | 5.45 | 19.92 |
| **Exp. 1 (Ours)** | 63.37 | 64.05 | 39.76 | 38.5 | **51.42** | 43.74 | 52.79 | 60.39 | 34.59 | **47.88** |
| **Exp. 2** | 53.95 | 51.45 | 42.73 | 31.1 | 44.81 | 23.78 | 38.40 | 36.22 | 14.02 | 28.10 |
| **Exp. 3** | 49.67 | 54.07 | 39.19 | 25.58 | 42.13 | 21.70 | 38.73 | 43.90 | 13.28 | 29.40 |
| **Exp. 4** | 50.63 | 53.15 | 45.02 | 25.9 | 43.68 | 29.92 | 41.07 | 39.01 | 12.59 | 30.65 |
| **Exp. 5** | 50.26 | 52.04 | 44.76 | 29.86 | 44.23 | 30.11 | 48.25 | 41.46 | 13.62 | 33.36 |
| **Exp. 6** | 50.14 | 49.72 | 34.54 | 21.74 | 39.03 | 25.52 | 38.94 | 44.98 | 11.54 | 30.24 |
| **Exp. 7** | 57.50 | 52.66 | 36.08 | 26.36 | 43.15 | 19.80 | 46.68 | 25.99 | 10.41 | 25.72 |

## Appendix B. Edge detector operators

Digital images are discrete and different edge detection algorithms differ in the way of $G_x$ and $G_y$ calculation. Different edge detector operators are given in Figure 4.

| Edge Detector | Definition | Operator |
|---|---|---|
| Roberts | $G_x = f(x,y) - f(x+1,y+1)$<br>$G_y = f(x,y+1) - f(x+1,y)$ | $G_x = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ $G_y = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$ |
| Sobel | $G_x = \{f(x+1,y-1) + 2f(x+1,y) + f(x+1,y+1)\} - \{f(x-1,y-1) + 2f(x-1,y) + f(x-1,y+1)\}$<br>$G_y = \{f(x-1,y+1) + 2f(x,y+1) + f(x+1,y+1)\} - \{f(x-1,y-1) + 2f(x,y-1) + f(x+1,y-1)\}$ | $G_x = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}$<br>$G_y = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$ |
| Prewitt | $G_x = \{f(x+1,y-1) + f(x+1,y) + f(x+1,y+1)\} - \{f(x-1,y-1) + f(x-1,y) + f(x-1,y+1)\}$<br>$G_y = \{f(x-1,y+1) + f(x,y+1) + f(x+1,y+1)\} - \{f(x-1,y-1) + f(x,y-1) + f(x+1,y-1)\}$ | $G_x = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}$<br>$G_y = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$ |
| Laplacian | $\nabla^2 f(x,y) = f(x+1,y) + f(x-1,y) + f(x,y+1) + f(x,y-1) - 4f(x,y)$ | $\nabla^2 f(x,y) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ |

Figure 4: Overview of classic edge detection algorithms' definitions and corresponding operators of the partial derivative of the image in the X and Y directions.

## Appendix C. Details and configuration of Canny algorithm

### C.1. Detail of Canny algorithm

The steps of the Canny algorithm (Canny, 1986) include Image Smoothing, Gradient Calculation, Non-maximum Suppression, and Edges Checking.

**Image Smoothing** Gaussian filter is used to smooth images and get rid of the noise which is defined as :

$$G(x, y) = \frac{1}{2\pi\sigma^2} exp(-\frac{x^2 + y^2}{2\sigma^2}) \tag{6}$$

where $\sigma$ stands for the size of the Gaussian Kernel, which controls the extent of smoothing the image. This critical parameter needs to be set manually based on experience.

**Gradient Calculation** The traditional Canny algorithm adopts a limited difference of $2 \times 2$ neighboring area to calculate the magnitude and direction of the image gradient. The operator of the partial derivative of the image in the X and Y directions is defined by:

$$G_x = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, G_y = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} \tag{7}$$

**Non-maximum Suppression** After acquiring the gradient magnitude image, it's needed to perform non-maximum suppression (NMS) on the image to accurately locate edges. The process of NMS can help guarantee that each edge is one-pixel width.

**Edges Checking** Canny adopts a double-threshold method to select edge points after carrying on non-maximum suppression. The pixels whose gradient magnitude is above the high threshold will be marked as edge points, and those whose gradient magnitude is under the low threshold will be marked as non-edge points, and the rest will be marked as candidate edge points. Those candidate edge points that are connected with edge points will be marked as edge points. This method reduces the influence of noise on the edge of the final edge image. The low and high thresholds need to be set manually based on experience.

### C.2. Configuration of Canny algorithm

We have manually designed the corresponding Cany detector parameters for each domain of each dataset. For all scenarios, the size of the Gaussian kernel $3 \times 3$. The configuration of the double-threshold is provided in Table 7.

Table 7: The configuration of Canny algorithm's double-threshold.

| Dataset | Domain | Low | High |
|---|---|---|---|
| **BraTS'19** | T2 | 40 | 80 |
| | T1 | 20 | 60 |
| | Flair | 40 | 100 |
| | T1ce | 20 | 50 |
| **Prostate** | Site A | 50 | 200 |
| | Site B | 100 | 200 |
| | Site C | 50 | 150 |
| | Site D | 50 | 140 |
| | Site E | 20 | 40 |
| | Site F | 30 | 70 |
| **MMWHS** | MRI | 30 | 80 |
| | CT | 70 | 120 |

## Appendix D. Visualization of edge detectors and CIConv

The visual comparison of different classic edge detection algorithms (Canny, 1986; Rong et al., 2014; Roberts, 1963; Prewitt et al., 1970; Kittler, 1983; LeCun et al., 1998) and CIConv (Lengyel et al., 2021) is shown in Figure 5. As we can see, there are large differences in the edge or gradient map extracted by different edge detection algorithms with the same image. Compared with CIConv, the classic edge detection algorithms can filter more useless information with less computation.

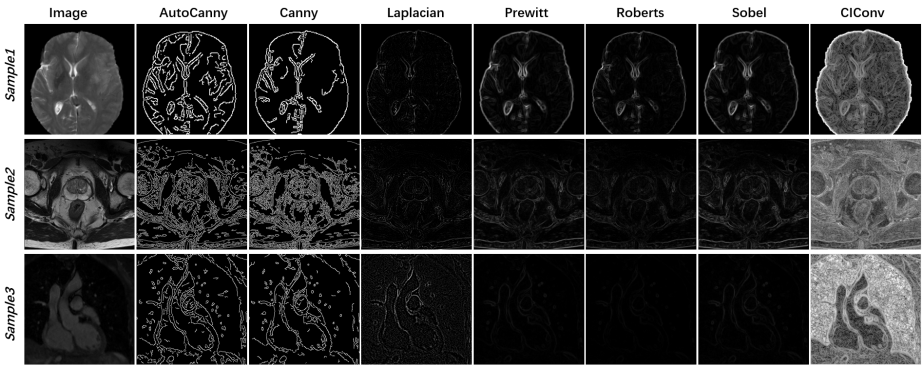


Figure 5: Visualization comparison of different edge detection algorithms and CIConv.

## Appendix E. Visualization of BézierCurve augmentation

The visualization examples of BézierCurve augmentation are shown in Figure 6. As introduced in the main text, this augmentation method maps the source image to diverse grayscale value distribution and keeps the appearance of the anatomic structures perceivable at the same time.

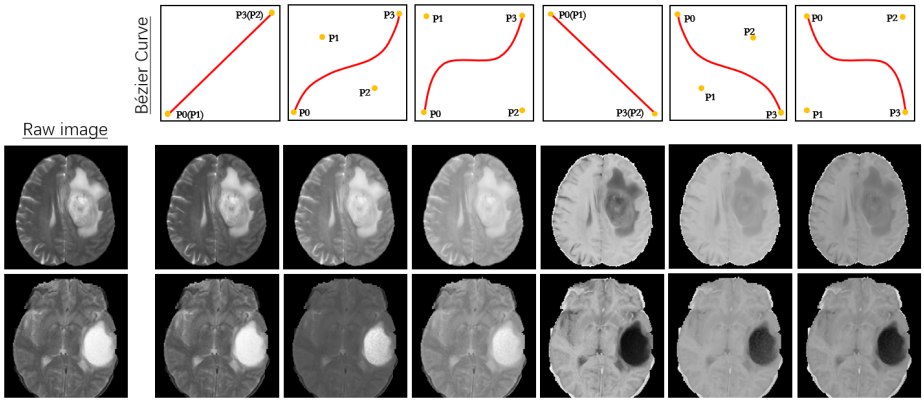


Figure 6: Visualization of generated Bezier Curve and corresponding augmented image on the BraTS'19 samples.

## Appendix F. Edge and gradient map of augmented image

We realized that for the same case, the edge and gradient map in each domain are different by the same edge detector. Therefore, we use data augmentation to simulate the data distribution of unknown target domain before edge extraction to enrich the gradient information of edge-guided training. The visualization of the augmented image and corresponding edge and gradient map is shown in Figure 7.

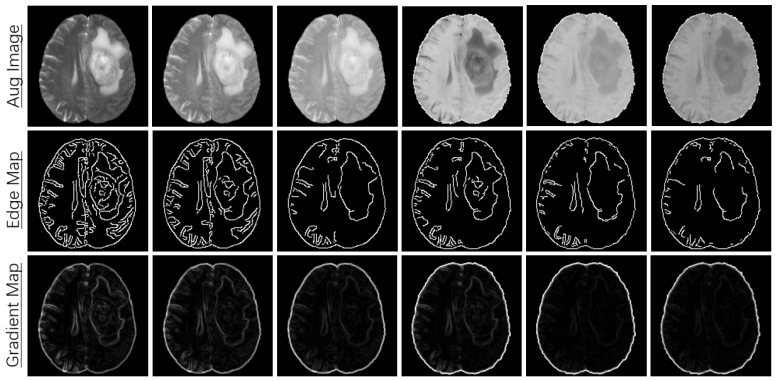


Figure 7: Visualization of augmented image and corresponding edge (AutoCanny) and gradient (Sobel) map.

## Appendix G. Details of datasets and preprocessing

BraTS'19 contains 335 cases which were acquired with different clinical protocols and various scanners from multiple institutions. Each case is composed of four sequences of MR images (T2, T1, Flair, and T1CE). Due to experts always annotating the whole tumor on T2, we use T2 as the source domain and others as unknown target domains. Prostate contains prostate T2-weighted MRI data collected from six different data sources. We follow the previous work (Liu et al., 2020) to partition the data into six datasets A to F, according to the clinical centers that the datasets collected. Consistent with our previous work, we take Site B as the source domain and others as unknown target domains. MMWHS dataset consists of unpaired 20 MRI and 20 CT volumes collected at different clinical sites, which contains the ground truth mask of four cardiac structures, including the ascending aorta (AA), the left atrium blood cavity (LAC), the left ventricle blood cavity (LVC), and the myocardium of the left ventricle (MYO). We make domain generalizations in both directions.

For data preprocessing, each volume was normalized to zero mean and unit variance. Then, we get the slices from each volume in the axial (BraTS'19 and Prostate) or coronal (MMWHS) plane and normalize the image to [-1, 1] before feeding it to the network. For BraTS'19 and MMWHS, we make the center crop and then resize it to $256 \times 256$. For Prostate, the size of the image is $384 \times 384$. Each domain was randomly split with 80% samples for training and 20% samples for testing. Specially, we shuffle all the volumes and divide them into four equal parts for each sequence firstly to prevent the ground truth leakage problem because the mask of each case is shared with four sequences in BraTS'19.

## Appendix H. Comparison between ours and CIConv

In essence, Color Invariant Convolution (CIConv) (Lengyel et al., 2021) is a learnable edge detector that is derived from the physics-based reflection models (Geusebroek et al., 2001). On the one hand, the computational process of Color Invariant theory (Geusebroek et al., 2001) is very complicated, which greatly increases the training and inference time. On the other hand, the performance of different variants of CIConv is unstable. Therefore, we compare the performance of CIConv and edge detection algorithms with the RefineNet (Lin et al., 2017) which is utilized in the CIConv model. Table 8 reports the comparison results, where we can see that the performance of different variants in CIConv varies greatly, while the performance of all edge detectors is stable and superior to the CIConv.

Table 8: The comparison result of CIConv (first group) and Edge-guided model (second group) on the BraTS'19 (left), Prostate (middle), and MMWHS (right) datasets with RefineNet. The best performance of CIConv and Edge-guided model is underlined and bolded respectively.

| Experiment | Source Domain: T2 | | | |
|---|---|---|---|---|
| | T1 | T1ce | Flair | Average |
| **invariant-E** | 34.95 | 41.08 | 62.06 | 46.03 |
| **invariant-W** | 44.79 | 31.15 | 57.10 | 44.35 |
| **invariant-C** | 10.93 | 18.98 | 25.00 | 18.30 |
| **invariant-N** | 0.00 | 0.00 | 0.00 | 0.00 |
| **invariant-H** | 7.98 | 10.78 | 10.80 | 9.85 |
| **Canny** | 42.82 | 45.62 | 51.34 | 46.59 |
| **AutoCanny** | 39.87 | 45.45 | 51.69 | 45.67 |
| **Roberts** | 38.21 | 42.95 | 51.71 | 44.29 |
| **Prewitt** | 36.62 | 36.73 | 60.55 | 44.63 |
| **Sobel** | 38.41 | 44.14 | 59.36 | **47.30** |
| **Laplacian** | 19.60 | 22.54 | 49.32 | 30.49 |

| Experiment | Source Domain: Site B | | | | | |
|---|---|---|---|---|---|---|
| | Site A | Site C | Site D | Site E | Site F | Average |
| **invariant-E** | 58.61 | 55.81 | 68.07 | 46.46 | 58.25 | 57.44 |
| **invariant-W** | 57.71 | 42.28 | 70.78 | 37.61 | 56.84 | 53.04 |
| **invariant-C** | 52.59 | 40.85 | 68.56 | 63.49 | 64.06 | 57.91 |
| **invariant-N** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **invariant-H** | 55.06 | 45.26 | 56.62 | 57.28 | 63.96 | 55.64 |
| **Canny** | 60.02 | 52.19 | 63.12 | 74.62 | 64.56 | 62.9 |
| **AutoCanny** | 77.10 | 62.35 | 71.05 | 71.00 | 66.54 | **69.61** |
| **Roberts** | 63.31 | 40.56 | 66.35 | 39.74 | 42.55 | 50.50 |
| **Prewitt** | 62.25 | 47.78 | 71.78 | 60.48 | 64.69 | 61.40 |
| **Sobel** | 61.25 | 36.48 | 69.08 | 51.35 | 42.87 | 52.21 |
| **Laplacian** | 64.70 | 36.69 | 69.26 | 53.29 | 51.85 | 55.16 |

| Experiment | Source Domain: MRI | | | | |
|---|---|---|---|---|---|
| | AA | LAC | LVC | MYO | Average |
| **invariant-E** | 55.55 | 57.07 | 52.77 | 26.79 | 48.04 |
| **invariant-W** | 70.35 | 76.32 | 58.81 | 42.45 | 61.98 |
| **invariant-C** | 60.17 | 60.09 | 45.79 | 30.00 | 49.01 |
| **invariant-N** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| **invariant-H** | 60.52 | 63.39 | 53.77 | 31.59 | 52.32 |
| **Canny** | 70.10 | 73.30 | 63.79 | 43.81 | 62.75 |
| **AutoCanny** | 67.56 | 71.43 | 54.27 | 44.45 | 59.43 |
| **Roberts** | 71.93 | 70.81 | 64.53 | 54.83 | 65.53 |
| **Prewitt** | 66.75 | 72.95 | 65.67 | 54.29 | 64.92 |
| **Sobel** | 64.85 | 75.37 | 66.26 | 56.15 | **65.66** |
| **Laplacian** | 71.99 | 72.48 | 64.33 | 51.75 | 65.14 |

## Appendix I. Results of Edge-guided with RandConv

RandConv (Xu et al., 2020) employs transformation via randomly initializing the weight of the first convolution layer, which enriches the gradient information of the image during training. We combined it with our edge-guided models to make further exploration. Table 9 reports the results on three datasets, which shows that the performance has improved compared to using only the edge-guided model. However, it's generally lower than training edge-guided models with BézierCurve augmented samples (Results refer to section 4.3 in the body).

Table 9: The result of edge-guided methods with RandConv augmented sample on the BraTS'19 (left), Prostate (middle), and MMWHS (right) datasets by U-Net with the EfficientNet-b2 backbone.

| Edge-Guided with RandConv | Source Domain: T2 | | | |
|---|---|---|---|---|
| | T1 | T1ce | Flair | Average |
| **Canny** | 53.48 | 53.00 | 57.80 | 54.76 |
| **AutoCanny** | 38.85 | 40.13 | 65.22 | 48.07 |
| **Roberts** | 51.43 | 46.49 | 67.82 | 55.25 |
| **Prewitt** | 46.81 | 51.46 | 68.37 | 55.55 |
| **Sobel** | 56.32 | 51.90 | 67.68 | **58.63** |
| **Laplacian** | 51.30 | 53.67 | 68.32 | 57.76 |

| Edge-Guided with RandConv | Source Domain: Site B | | | | | |
|---|---|---|---|---|---|---|
| | Site A | Site C | Site D | Site E | Site F | Average |
| **Canny** | 57.30 | 52.65 | 58.57 | 58.73 | 45.65 | 54.58 |
| **AutoCanny** | 76.27 | 59.08 | 80.09 | 74.93 | 72.56 | **72.59** |
| **Roberts** | 70.75 | 50.53 | 76.69 | 52.93 | 65.61 | 63.30 |
| **Prewitt** | 62.70 | 52.44 | 54.53 | 54.46 | 46.41 | 54.11 |
| **Sobel** | 73.72 | 53.78 | 75.41 | 64.95 | 54.90 | 64.55 |
| **Laplacian** | 65.99 | 60.79 | 61.20 | 56.81 | 32.22 | 55.40 |

| Edge-Guided with RandConv | Source Domain: MRI | | | | |
|---|---|---|---|---|---|
| | AA | LAC | LVC | MYO | Average |
| **Canny** | 63.99 | 70.52 | 58.85 | 52.10 | 61.37 |
| **AutoCanny** | 68.18 | 68.96 | 59.79 | 48.04 | 61.24 |
| **Roberts** | 69.44 | 69.65 | 57.90 | 47.74 | 61.18 |
| **Prewitt** | 65.77 | 65.04 | 49.12 | 55.01 | 58.74 |
| **Sobel** | 72.58 | 69.04 | 66.96 | 57.96 | **66.64** |
| **Laplacian** | 75.93 | 68.83 | 65.15 | 55.59 | 66.38 |