Gradient-Map-Guided Adaptive Domain Generalization for Cross Modality MRI Segmentation

Bingnan Li   LIBN@SHANGHAITECH.EDU.CN
School of Information Science and Technology, ShanghaiTech University

Zhitong Gao   GAOZHT@SHANGHAITECH.EDU.CN
School of Information Science and Technology, ShanghaiTech University

Xuming He   HEXM@SHANGHAITECH.EDU.CN
School of Information Science and Technology, ShanghaiTech University
Shanghai Engineering Research Center of Intelligent Vision and Imaging

Abstract

Cross-modal MRI segmentation is of great value for computer-aided medical diagnosis, enabling flexible data acquisition and model generalization. However, most existing methods have difficulty in handling local variations in domain shift and typically require a significant amount of data for training, which hinders their usage in practice. To address these problems, we propose a novel adaptive domain generalization framework, which integrates a learning-free cross-domain representation based on image gradient maps and a class prior-informed test-time adaptation strategy for mitigating local domain shift. We validate our approach on two multi-modal MRI datasets with six cross-modal segmentation tasks. Across all the task settings, our method consistently outperforms competing approaches and shows a stable performance even with limited training data. Our Codes are available now at https://github.com/cuttle-fish-my/GM-Guided-DG.

Keywords: MRI Segmentation, Domain Generalization, Test Time Adaptation

1. Introduction

Semantic segmentation of Magnetic Resonance Imaging (MRI) sequences, a core task in computer-aided medical diagnoses, has achieved remarkable progress due to the powerful representation learning based on deep neural networks (Ronneberger et al., 2015; Isensee et al., 2021). Conventional approaches typically assume the same data distribution and share the same modality setting in both training and test stages (Zhou et al., 2019). Despite their promising
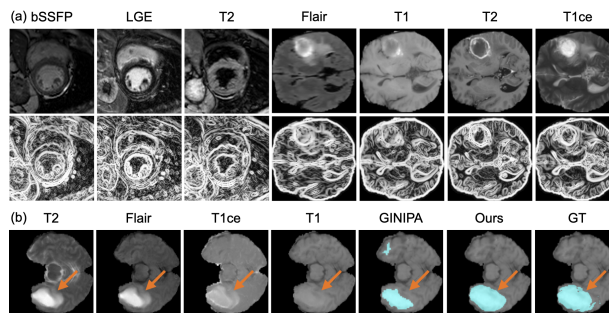

Figure 1: (a) An illustration of original images and Gradient-Map Representation (GMR) which can effectively mitigate the domain gap; (b) Comparison of segmentation outputs (GINIPA vs. our method) when local variations exist for different modalities.

results, these methods either employ multi-modal inputs or learn a separate model for each modality, which induces high time or financial costs. To alleviate this, one promising strategy involves training a model on one specific modality and enabling it to adapt to other modalities during the test phase, which is referred to as Single Domain Generalization (SDG) (Qiao et al., 2020). However, it is challenging to apply the common SDG strategy to MRI segmentation due to the large domain gaps between different modalities and often limited training data in medical applications. There have been a few attempts to tackle the problem of SDG for MRI image segmentation, which mostly focus on learning a domain-invariant feature space by designing specific learning

strategies (Liu et al., 2020; Xu et al., 2022) or domain randomization techniques (Zhang et al., 2020; Ouyang et al., 2022; Liu et al., 2021).

Nonetheless, the efficacy of these learned feature spaces is highly contingent on the amount of training data and their performance tends to degrade rapidly with fewer data (see Figure 4 for details). Moreover, most of these methods focus on generalization with image-level domain variation and may struggle with local variations in different domains, such as the different responses of lesion tissues across modalities. More recent work (Liu et al., 2022) introduces Test-Time Adaptation (TTA) techniques coupled with a shape dictionary representation. However, such a strategy requires spatial alignment of target regions, and hence is limited in handling objects with diverse shapes and locations (e.g., brain tumors (Menze et al., 2014)).

In this work, we aim to address the aforementioned limitations by introducing a novel gradient-map-guided adaptive domain generalization framework for MRI segmentation. In particular, we adopt a learning-free domain-invariant input representation based on the image gradient map, which is capable of mitigating the global image style shift across different modalities. Based on that, we then develop a new test-time adaption strategy that integrates semantic class prior with pseudo-label-based self-training. This allows us to cope with local appearance variations and adapt to image-specific pixel-wise class distributions.

Specifically, we adopt a two-phase domain generalization procedure for cross-modal MRI segmentation. During the training phase, our method encodes input images into their gradient maps (see Figure 1 (a) for an illustration), and trains a semantic segmentation network based on the training data of the source modality. During the test phase, we further fine-tune the segmentation model on each input image from the target modality in an iterative manner using the pseudo-labels of pixels generated with a dynamically-estimated pixel-wise class prior.

We evaluate our method on two multi-modality MRI datasets, including MS-CMRSeg2019 (Zhuang, 2018) and BraTS2018 (Menze et al., 2014), under the standard and limited training data settings. The results show the superiority of our method in comparison with recent domain generalization and TTA methods. We summarize our contributions as follows:

- We propose an adaptive single domain generalization method for cross-modal semantic segmentation of MRI images to better cope with both global and instance-specific domain gaps.

- We introduce an efficient gradient-based domain-invariant representation in MRI image segmentation and develop an effective TTA strategy based on class prior-aware self-training.

- Our method achieves state-of-the-art performance on cross-modal cardiac segmentation and brain tumor segmentation tasks.

2. Related Work

The cross-modality MRI segmentation problem has received significant attention in the literature, and a variety of approaches have been developed under different problem settings. A straightforward solution is to adapt a source-domain-trained model to the target domain during training, which is referred to as Unsupervised Domain Adaptation (UDA) (Chen et al., 2020b; Vesal et al., 2021; Yang et al., 2019). However, the requirement for target data during training poses a practical challenge, as acquiring target data can be arduous or infeasible in many real-world scenarios. This limitation has led to the widespread use and growing importance of Domain Generalization (DG) and Test-Time Adaptation (TTA) methods, which only require access to source domain data during the training stage. In the subsequent sections, we offer an overview of the relevant work within the domains of DG and TTA.

Domain Generalization The goal of Domain Generalization is to learn a model capable of generalizing to any other domain during the test. While some techniques rely on training data from multiple domains (Pereira et al., 2016; Dolz et al., 2018), we focus on a more challenging setting where only one domain is available during training. This is also called Single Domain Generalization (SDG).

Existing methods address the problem by encouraging models to learn a domain-invariant feature space implicitly. For instance, domain randomization methods like random weighted network (Ouyang et al., 2022) and data augmentations (Zhang et al., 2020; Otálora et al., 2019; Chen et al., 2020a) simulate potential target domain information, promoting the development of domain-robust models. In addition to these techniques, special learning strategies like meta-learning (Oliveira et al., 2022; Liu et al.,

2020), disentangle learning (Liu et al., 2021) and adversarial learning (Xu et al., 2022) are used to regularize the model for domain invariance. However, most of these methods focus on generalization on global-level domain shifts, such as different styles, and struggle with local appearance variations across different domains. Besides, the efficacy of these learned feature spaces could also be highly contingent on the size of the training dataset. By contrast, we utilize a learning-free gradient-map-based domain-invariant representation to handle the global style variation. During test time, we introduce a pseudo-label-based self-training strategy to mitigate the local domain shifts that can vary inconsistently across different modalities.

Test Time Adaptation The objective of Test-Time Adaptation (TTA) is to harness online test data to adapt the model during the testing phase. Several techniques have been explored in this field. Wang et al. (2020) introduce a novel method that employs entropy minimization exclusively on BathNorm (BN) layers to force a confident prediction. He et al. (2021) utilize Auto Encoder (AE) as a domain distance indicator and update the model via the reconstruction loss. Wang et al. (2023b) introduces prototype-based methods into the realm of TTA and harness the feature distance to different prototype as a measure of classification probability. Tang et al. (2023) proposed a BP-free TTA method using Hebbian layers that inspired from Hebbian learning (Hebb, 2005). Nevertheless, TTA methods often rely heavily on the base model performance, and some require special network designs such as AE and Hebbian layers. Recently, Liu et al. (2022) introduced a novel approach that combines DG and TTA by incorporating dictionary learning during training and constraining the consistency of the dictionary coefficient between two different noise-disturbed images in test time. However, the strong assumption about shape-invariant and location-invariant segmentation targets across source and target domains limits the method when handling objects with diverse shapes and locations.

In contrast, our proposed pseudo-label-based TTA strategy utilizes class-prior to prioritize pseudo-label generation of the foreground classes. This approach empowers our models to produce high-quality segmentations, even in cases where the initial predictions are far from ideal. Besides, our method avoids assumptions about the shape and location of the seg-

mentation target, which demonstrates the flexibility and applicability of our proposed method.

3. Methodology

In this section, we introduce our adaptive domain generalization framework for MRI segmentation, which proposes two novel components to cross-modality segmentation network learning: 1) a modality-robust input representation based on gradient maps and 2) a class prior-informed test-time adaptation strategy. An overview of our framework is shown in Figure 2. Below, we first introduce the problem setting in Section 3.1, then describe the design of our gradient-map-based semantic segmentation network and its training on the source domain in Section 3.2. This is followed by our prior-informed test time adaptation in Section 3.3.

3.1. Problem Setting

Consider a training dataset $D = \{\mathbf{x}_n, \mathbf{y}_n\}_{n=1}^N$ with $\{\mathbf{x}_n, \mathbf{y}_n\} \sim P_{XY}^s$, where $\mathbf{x}_n \in R^{H \times W}$ is a 2D MRI image from a source modality and $\mathbf{y}_n \in [1, C]^{H \times W}$ is the corresponding segmentation label. Here H, W are the spatial dimensions and C is the number of semantic classes. The goal of cross-modality semantic segmentation is to learn a segmentation model $\mathcal{M}_\theta$ from D that can be generalized to images from any other modalities during the test. In this paper, we aim to design an adaptive domain generalization framework, which utilizes both the source-modality data and the online unlabeled test data to adapt the network to fit the target modality, where the test image $\mathbf{x} \sim P_{XY}^t$ with $P_X^t \neq P_X^s$.

3.2. Modality-robust Segmentation Network

We adopt a U-Net (Ronneberger et al., 2015) as our segmentation backbone, and first build a modality-robust model using the source-modality dataset D. In particular, as detailed below, we introduce an image gradient map-based representation as the network input and a heavy data augmentation strategy in model training, both of which mitigate the modality bias in the learned segmentation network.

Gradient-Map Representation We first introduce our *Gradient-Map Representation (GMR)* as a modality-invariant input representation, which leverages the image prior to achieve modality generalization with high data efficiency. Our design is inspired
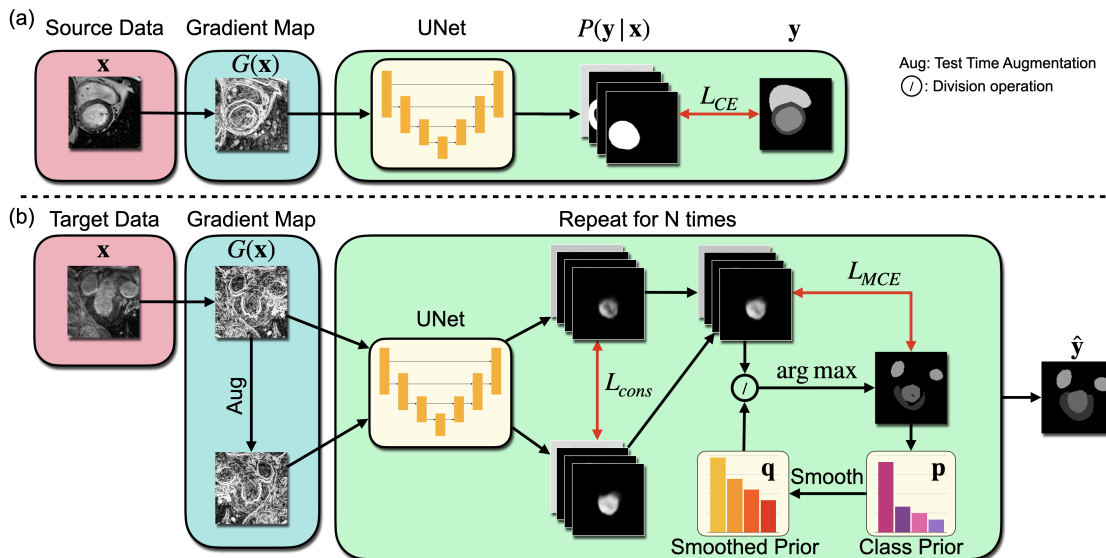
Figure 2: Overview of our method. Subgraph (a) and (b) refer to the training and test stage pipeline respectively. For the training state, we employ the gradient map $G(\mathbf{x})$ as a domain-invariant representation to mitigate image-level domain shift. For test stage, we utilize class prior $\mathbf{p}$ to adapt posterior $P(\mathbf{y}|\mathbf{x})$ for N times, alleviating local discrepancy.

by a key observation that while different modalities may have unique global styles, the most pronounced variations in intensity occur at relatively consistent locations, typically at the boundaries between adjacent tissues. This observation motivates us to use the gradient map, which captures these relative intensity changes (structural information), as a modality-robust representation.

Formally, the gradient map of an input image $\mathbf{x}$, denoted as $G(\mathbf{x})$, is computed as follows:

$$G(\mathbf{x}) = HE\left(\sqrt{corr\left(\mathbf{x}, \mathbf{K}\right)^2 + corr\left(\mathbf{x}, \mathbf{K}^T\right)^2}\right), \quad (1)$$

where $HE(\cdot)$ refers to the histogram equalization, $corr(\cdot, \cdot)$ represents the correlation operation and $\mathbf{K} = [-1\ 0\ 1]$. We visualize GMR examples in Figure 1 (a), which shows that the discrepancy between different modalities has been effectively reduced and the structural information is well preserved.

Model Training We employ the Heavy Augmentation (Vesal et al., 2021) technique in the training process. This method involves augmenting both the source image $\mathbf{x}$ and its corresponding annotation $\mathbf{y}$ using a transformation $T \in \mathcal{T}$ sampled from a distri-

bution $P(T)$. The optimization problem of the training process can be described as follows:

$$\min_{\theta} \mathbb{E}_T \left[L_{ce}\left(\mathcal{M}_\theta \left(T^{(X)}\left(G\left(\mathbf{x}\right)\right)\right), T^{(Y)}(\mathbf{y})\right)\right] \quad (2)$$

where $L_{ce}(\cdot, \cdot)$ is the Cross-Entropy (CE) loss and $T^{(X)/(Y)}(\cdot)$ refer to the data augmentation function applied to input data $\mathbf{x}$ and annotation $\mathbf{y}$.

3.3. Prior-Informed Test-Time Adaptation

While the GRM representation reduces the modality gap in terms of global intensity style shift, it is unable to eliminate modality-specific local appearance variations. For instance, Figure 1 (b) displays a brain MRI with arrows indicating the local intensity variation that is inconsistent across four modalities. Such modality-specific variation hinders the model from generalizing across different modalities.

To tackle this, we propose a test-time adaptation strategy to finetune the segmentation network with unlabeled test data in an online fashion. In particular, we develop a self-training method that gradually expands the regions of certain classes, which typically include the regions with modality-specific variation (e.g., lesion areas), via an iterative self-labeling

scheme. We refer to this process as *Prior-Informed Test Time Adaptation (PITTA)*.

Specifically, given a test image $\mathbf{x}$, we start from the model trained on the source modality as in Section 3.2 and perform N adaptation iterations. For each iteration, we first merge the BatchNorm (BN) statistics from the source and target domain to stabilize prediction (Wang et al., 2023a), and then employ consistency loss and Masked Cross Entropy (MCE) loss to dynamically adjust the model. Below we will introduce the details of each step in PITTA.

Test Time Augmentation Instead of direct prediction, we leverage Test Time Augmentation (Zeng et al., 2017; Ronneberger et al., 2015; Beier et al., 2017) to improve the robustness of model prediction. In detail, for the n-th iteration, we utilize a test-time augmentation function $A_T(\cdot)$[1] generate two logit predictions from the original input and its augmented version, respectively. We denote those two outputs as $\mathbf{S}^{(n)}$ and $\mathbf{S}'^{(n)}$, which are computed as follows:

$$\mathbf{S}^{(n)} = \mathcal{M}_\theta(G(\mathbf{x})) \quad (3)$$

$$\mathbf{S}'^{(n)} = A_T^{-1}(\mathcal{M}_\theta(A_T(G(\mathbf{x})))) \quad (4)$$

where $A_T^{-1}(\cdot)$ refers to the inverse augmentation function, which serves to counteract the initial spatial augmentations. We take the average of the two logit outputs to generate the pixel-wise class probabilities as follows,

$$P^{(n)}(\mathbf{y}|\mathbf{x}) = softmax\left(\left(\mathbf{S}^{(n)} + \mathbf{S}'^{(n)}\right)/2\right). \quad (5)$$

Prior-Informed Self-training We now introduce our self-training strategy to finetune the model on the test data. In each iteration, our training loss consists of two terms: one is a consistency loss L_{cons} and the other is a pseudo-label-based MSE loss L_{mce}. We fine-tune the segmentation network with the following total loss:

$$L_{total} = L_{mce} + \lambda L_{cons} \quad (6)$$

where λ is the weighting coefficient. Below we describe the details of two losses in the n-th iteration.

1) Consistency Loss. Given the two outputs $\mathbf{S}^{(n)}, \mathbf{S}'^{(n)}$, we define the consistency loss as below:

$$L_{cons} = |\mathbf{S}^{(n)} - \mathbf{S}'^{(n)}|. \quad (7)$$

---

1. Note that our test-time augmentation $A_T(\cdot)$ differs from the Heavy Augmentation in model training (Section 3.2). to augment the input MRI, and

2) Pseudo label-based Loss. We design an adaptive pseudo-label generation scheme that utilizes the class prior information to better cope with the local variations across modalities and the distinctive class distribution of each input. Specifically, we first estimate a per-image class distribution $\mathbf{p}^{(n)}$ from the class distribution of the training set and the model prediction on the current test input. Based on this estimation, we then perform a re-weighting on prediction probability to increase the proportion of pseudo labels from smaller classes with less data.

Such a scheme improves the model adaptation from two aspects: First, we observe that cross-modal local variations are typically associated with foreground classes with smaller regions. Our re-weighting enables the model fine-tuning to focus on those classes by generating more training pixels with pseudo labels from them. Second, by gradually shifting from the training class prior to (estimated) per-image class distribution, our model is able to adapt to each test image by properly re-balancing the probability scores of all the classes.

Formally, we initialize $\mathbf{p}$ as the empirical class distribution estimated from the training set in the first iteration (i.e., $n = 1$). Denote the number of classes as C, and at the n-th iteration, we compute a re-weighting probability vector $\mathbf{q}^{(n)}$ from the estimated per-image class distribution $\mathbf{p}^{(n)}$ as follows:

$$\mathbf{q}^{(n)} = \alpha\frac{1}{C}\mathbf{1} + (1-\alpha)\mathbf{p}^{(n)} \quad (8)$$

where α is the coefficient for smoothing and $\mathbf{1}$ is a full-one vector of length C. We then generate our pseudo labels for the input $\mathbf{x}$ using a re-weighted probability output as below:

$$\hat{\mathbf{y}}_{ij}^{(n)} = \arg\max\left(P^{(n)}(\mathbf{y}_{ij}|\mathbf{x})/\mathbf{q}^{(n)}\right) \quad (9)$$

where ij is the pixel index and / denotes element-wise division of two vectors.

Given the generated pseudo labels $\hat{\mathbf{y}}^{(n)}$, we define our pseudo-label-based loss using a masked version of the Cross-Entropy (MCE). This choice is motivated by the fact that cross-modal local variations typically fall within a subset of classes, and the classes that are already well-segmented provide little loss feedback. Specifically, denote the subset as $\mathcal{C}_{lv}$ and the corresponding pixel mask as $\mathbf{M}$, we introduce our MCE loss term as follows,

$$L_{mce} = L_{ce}\left(\mathbf{M} \odot P^{(n)}(\mathbf{y}|\mathbf{x}), \hat{\mathbf{y}}^{(n)}\right) \quad (10)$$

where $\mathbf{M}_{i,j} = \bigcup_{c \in \mathcal{C}_{lv}} \mathbf{1}\left(\hat{\mathbf{y}}_{i,j}^{(n)} = c\right)$, $\mathbf{1}(\cdot)$ is the indicator function and $\odot$ is the element-wise multiplication.

After the model update, we compute the class frequency from the model output $\hat{\mathbf{y}}^{(n)}$ on the test image and update the per-image class distribution $\mathbf{p}$ based on the Exponential Moving Average (EMA) as below:

$$\mathbf{p}^{(n+1)} = \alpha \mathbf{p}^{(n)} + (1-\alpha)\#\left\{\hat{\mathbf{y}}^{(n)}\right\} \qquad (11)$$

where α is the weight coefficient and $\#\{\cdot\}$ indicates the function computing the class frequency from the pseudo labels. After the N-th iteration, the model outputs the final prediction $\hat{\mathbf{y}}$ as follows:

$$\hat{\mathbf{y}}_{ij} = \arg\max\left(P^{(N)}(\mathbf{y}_{ij}|\mathbf{x})/\mathbf{q}^{(N)}\right) \qquad (12)$$

Note that we use re-weighted probability scores in order to take into account per-image class imbalance. The pseudo-code of our Prior-Informed Adaptation process is shown in Appendix A Algorithm 1.

4. Experiments

We evaluate our method on two public cross-modality MRI datasets, MS-CMRSeg2019 (Zhuang, 2018; Qiu et al., 2023) and BraTS2018 (Menze et al., 2014; Lloyd et al., 2017; Bakas et al., 2018, 2017) with various cross-modal segmentation tasks. Below, we first introduce the dataset information in Section 4.1 and experiment setup in Section 4.2. Then we present our experimental results in Section 4.3 and ablation study in Section 4.4.

4.1. Datasets

The MS-CMRSeg2019 Dataset (Zhuang, 2018; Qiu et al., 2023) features cardiac MRIs from 35 patients across three modalities: balanced Steady-State Free Precession (bSSFP), Late Gadolinium Enhancement (LGE), and T2-weighted MRI. Each patient has multiple slices (typically 8-12 for bSSFP, 10-18 for LGE, and 3-7 for T2). The patients have an average age of 56.2 ± 7.92 years and an average weight of 74.4 ± 5.65 kg (Zhuang et al., 2022). The dataset targets the segmentation of three essential cardiac structures: Myocardium (Myo), Left Ventricle (LV), and Right Ventricle (RV). Given that bSSFP offers clearer visibility for these structures and is preferred for clinical annotations, we designate it as the source domain, and treat LGE and T2 as the target domains (Zhuang et al., 2022).

The BraTS2018 Dataset (Menze et al., 2014; Lloyd et al., 2017; Bakas et al., 2018, 2017) features brain MRI scans of 285 patients, captured across four modalities: T1, T2, post-contrast T1-weighted (T1ce), and T2 Fluid Attenuated Inversion Recovery (FLAIR). The dataset includes annotations across four categories, but we follow previous work (Xie et al., 2022; Zou et al., 2020; Han et al., 2021), to focus on the domain generalization of the Whole Tumor (WT) region. Given that the T2 and FLAIR modalities provide clearer visibility of the WT region, we select them as our source domains, and experiment with the domain generalization settings from T2 to T1 and T1ce, as well as from FLAIR to T1 and T1ce.

4.2. Experiment Setup

Baselines We compare our methods with a diverse set of existing techniques to establish a comprehensive performance benchmark. These include single-domain generalization approaches, test-time adaptation techniques, unsupervised domain adaptation methods, and several other baseline models. Below, we outline the specific methods used for comparison: (a) **SrcOnly**, a baseline method that trains models on the source domain without any DG techniques. (b) **SAM++** (Kirillov et al., 2023), a zero-shot large-scale segmentation model. We use bounding box prompts generated from the prediction of our model. For the myocardium class, we also incorporate a negative point at its geometrical center. (c) **MinEnt** (Vu et al., 2019), an Unsupervised Domain Adaptation (UDA) technique that aims to minimize the entropy of target predictions during the training phase. (d) **HA** (Vesal et al., 2021), A DG approach that leverages a variety of pre-defined transformations to augment the training data. (e) **GINIPA** (Ouyang et al., 2022), a DG method that uses a random weighted network and random spatial interpolation to stylize the image. (f) **Tent** (Wang et al., 2020), a TTA method that employs entropy minimization during the test stage. (g) **HA** (Vesal et al., 2021)**+Tent** (Wang et al., 2020), a DG+TTA baseline that integrates HA and Tent for training and test stage.

Evaluation Following the literature (Ouyang et al., 2022; Vesal et al., 2021, 2019), we employ the Volumetric Dice Score as the evaluation metric. we conducted all experiments three times and report the results with mean and standard derivation.

Table 1: Quantitative comparisons in terms of Volumetric Dice scores on MS-CMRSeg2019 dataset.

Setting	bSSFP→T2			
Method	Category	Myo	LV	RV	Average
SrcOnly	-	$0.0712_{\pm\ 0.0386}$	$0.0687_{\pm\ 0.0490}$	$0.0068_{\pm\ 0.0093}$	$0.0489_{\pm\ 0.0302}$
SAM++	-	$0.8364_{\pm\ 0.0004}$	$0.8370_{\pm\ 0.0014}$	$0.7619_{\pm\ 0.0060}$	$0.8118_{\pm\ 0.0017}$
MinEnt	-	$0.0482_{\pm\ 0.0067}$	$0.0513_{\pm\ 0.0051}$	$0.0153_{\pm\ 0.0108}$	$0.0382_{\pm\ 0.0067}$
HA	DG	$0.6600_{\pm\ 0.0064}$	$0.7576_{\pm\ 0.0145}$	$0.6295_{\pm\ 0.0256}$	$0.6824_{\pm\ 0.0126}$
GINIPA	DG	$0.7862_{\pm\ 0.0272}$	$0.8435_{\pm\ 0.0208}$	$0.7676_{\pm\ 0.0122}$	$0.7991_{\pm\ 0.0134}$
Tent	TTA	$0.1952_{\pm\ 0.0963}$	$0.3426_{\pm\ 0.1302}$	$0.1982_{\pm\ 0.0903}$	$0.2453_{\pm\ 0.0806}$
HA&Tent	DG&TTA	$0.6698_{\pm\ 0.0296}$	$0.7568_{\pm\ 0.0229}$	$0.6838_{\pm\ 0.0305}$	$0.7035_{\pm\ 0.0191}$
Ours	DG&TTA	$\mathbf{0.8516}_{\pm\ \mathbf{0.0031}}$	$\mathbf{0.8934}_{\pm\ \mathbf{0.0040}}$	$\mathbf{0.8166}_{\pm\ \mathbf{0.0063}}$	$\mathbf{0.8539}_{\pm\ \mathbf{0.0044}}$
Setting	bSSFP→LGE			
Method	Category	Myo	LV	RV	Average
SrcOnly	-	$0.5625_{\pm\ 0.0316}$	$0.7046_{\pm\ 0.0392}$	$0.6329_{\pm\ 0.0323}$	$0.6333_{\pm\ 0.0341}$
SAM++	-	$0.6997_{\pm\ 0.0089}$	$0.8773_{\pm\ 0.0014}$	$0.8158_{\pm\ 0.0084}$	$0.7976_{\pm\ 0.0031}$
MinEnt	-	$0.6021_{\pm\ 0.0075}$	$0.7486_{\pm\ 0.0051}$	$0.6677_{\pm\ 0.0049}$	$0.6728_{\pm\ 0.0014}$
HA	DG	$0.7389_{\pm\ 0.0065}$	$0.8880_{\pm\ 0.0016}$	$0.8509_{\pm\ 0.0028}$	$0.8259_{\pm\ 0.0023}$
GINIPA	DG	$0.7557_{\pm\ 0.0114}$	$0.9021_{\pm\ 0.0022}$	$0.8437_{\pm\ 0.0139}$	$0.8338_{\pm\ 0.0074}$
Tent	TTA	$0.5739_{\pm\ 0.0359}$	$0.7658_{\pm\ 0.0348}$	$0.6260_{\pm\ 0.0355}$	$0.6552_{\pm\ 0.0351}$
HA&Tent	DG&TTA	$0.7188_{\pm\ 0.0133}$	$0.8765_{\pm\ 0.0050}$	$0.7935_{\pm\ 0.0078}$	$0.7963_{\pm\ 0.0016}$
Ours	DG&TTA	$\mathbf{0.7964}_{\pm\ \mathbf{0.0020}}$	$\mathbf{0.9091}_{\pm\ \mathbf{0.0031}}$	$\mathbf{0.8560}_{\pm\ \mathbf{0.0091}}$	$\mathbf{0.8538}_{\pm\ \mathbf{0.0043}}$

Implementation Details We employ U-Net architecture (Ronneberger et al., 2015) as the segmentation model. During the training phase, the batch size is 24, and the learning rate is 0.0001. All models undergo training for 10000 iterations. In the training stage, we utilize an online imaging library (imgAug[2]) to implement Heavy Augmentation (Vesal et al., 2021). In the test stage, we adopt horizontal flipping as the Test Time Augmentation function. The parameter ρ remains constant at 0.4 for all experiments, while λ is set to 1. Additionally, α is configured as 0.9 for MS-CMRSeg2019 and 0.5 for BraTS2018. The adaptation learning rate is set at 0.01, and each image undergoes two adaptation iterations. Moreover, $\mathcal{C}_{lv}$ used for MCE loss is set to be $\{Myo\}$ and $\{WT\}$ for MS-CMRSeg2019 and BraTS2018 respectively. We implemented the code framework via PyTorch on a single NVIDIA A40 (48GB). More implementation details are described in Appendix C.

4.3. Results

MS-CMRSeg2019 Results As demonstrated in Table 1, our approach consistently surpasses previous methods across all classes and settings. Its performance is particularly noteworthy in the T2 target domain, where the domain shift from the source domain bSSFP is significant, as visualized in Figure 1 (a). Specifically, our method improves the baseline average class Dice score from a mere 4.9% to an impressive 85%. Furthermore, it outperforms the previous state-of-the-art (SOTA) method, GINIPA (Ouyang et al., 2022), by an average of 5.4% in Dice score.

BraTS2018 Results Table 2 showcases the quantitative performance comparison on the BraTS2018 dataset, where our method achieves the best performance for all settings. Interestingly, we observe that most previous methods fall short on this dataset. One possible reason for this could be the pronounced local variations among different modalities, as shown in Figure 1 (b). These local variations seem to

---

Table 2: Quantitative comparisons in terms of Dice scores on BraTS2018 dataset.

Setting	T2→T1	T1→T1ce	Flair→T1	Flair→T1ce
SrcOnly	0.0805 ± 0.0032	0.1528 ± 0.0134	0.0616 ± 0.0026	0.3423 ± 0.0061
SAM++	0.6475 ± 0.0024	0.6224 ± 0.0188	0.4048 ± 0.0113	0.5492 ± 0.0079
MinEnt	0.0657 ± 0.0058	0.1157 ± 0.0156	0.0755 ± 0.0022	0.3647 ± 0.0102
HA	0.0778 ± 0.0017	0.1724 ± 0.0244	0.0699 ± 0.0034	0.3630 ± 0.0027
GINIPA	0.2302 ± 0.0107	0.2840 ± 0.0093	0.1865 ± 0.0118	0.3607 ± 0.0068
Tent	0.0327 ± 0.0034	0.1117 ± 0.0075	0.0463 ± 0.0045	0.2790 ± 0.0167
HA&Tent	0.0585 ± 0.0078	0.1371 ± 0.0183	0.1012 ± 0.0159	0.3688 ± 0.0143
Ours	$\mathbf{0.6793} \pm \mathbf{0.0028}$	$\mathbf{0.6876} \pm \mathbf{0.0035}$	$\mathbf{0.4260} \pm \mathbf{0.0074}$	$\mathbf{0.5958} \pm \mathbf{0.0038}$

present difficulties for previous techniques that exclusively depend on image-level augmentations, such as HA(Vesal et al., 2021) and GINIPA(Ouyang et al., 2022).

Result Visualization As evidenced by Figure 3, our method generates segmentation masks that closely resemble the ground truth. Notably, it accurately segments regions impacted by local appearance discrepancies, as indicated by the arrows in the graph. This underscores the effectiveness of our proposed PITTA component in mitigating the effects of local domain shifts. Additional visual results are provided in Appendix E, Figure 5.

Table 3: Ablation study on our model components. Performed on MS-CMRSeg2019 Dataset.

HA	GMR	PITTA	bSSFP→T2	bSSFP→LGE
✓	✗	✗	0.6824 ± 0.0126	0.8259 ± 0.0023
✓	✓	✗	0.8430 ± 0.0066	0.8383 ± 0.0044
✓	✗	✓	0.7393 ± 0.0257	0.8400 ± 0.0012
✓	✓	✓	$\mathbf{0.8539} \pm \mathbf{0.0044}$	$\mathbf{0.8538} \pm \mathbf{0.0043}$

4.4. Ablation Study

We assess the effectiveness of our proposed architecture by conducting ablation studies on its two main components: GMR and PITTA. Table 3 presents a performance comparison of our method's variants on the MS-CMRSeg2019 dataset under two different settings. In the 2nd row, incorporating the GMR component improves the baseline method (HA) by increasing the average volumetric Dice score by 16%

and 1.2% on the two target datasets, respectively. Similarly, in the 3rd row, adding the PITTA component boosts the baseline performance, leading to an average Dice score improvement of 5.7% and 1.4%. The final row shows that our complete method, which integrates both components, achieves the best performance, thereby confirming the effectiveness of our design choices.

Data Efficiency Given the challenges of obtaining sufficient data and annotations, data-efficiency becomes crucial in medical settings. To evaluate this, we examine the stability of our method under varying training dataset sizes, with the results shown in Figure 4. Specifically, in the MS-CMRSeg2019 dataset, we adjust the number of training patients from the original 35 down to subsets of 15, 20, 25, and 30 and compare our performance to the existing state-of-the-art, GINIPA. Our findings indicate that our model is less sensitive to reductions in dataset size. For instance, when the dataset is trimmed from 35 to 15 patients, the performance of our model declines by only 0.44%, compared to a 2.89% drop for GINIPA. This underscores the robustness of our method when data is limited.

For other analysis experiments, please refer to Appendix D for details.

5. Conclusion

In this paper, we propose a novel adaptive domain generalization framework to address the cross-modal MRI segmentation problem. Our method integrates a learning-free cross-domain representation based on image gradient maps for coping with global domain shift and a class prior-informed test-time adaptation
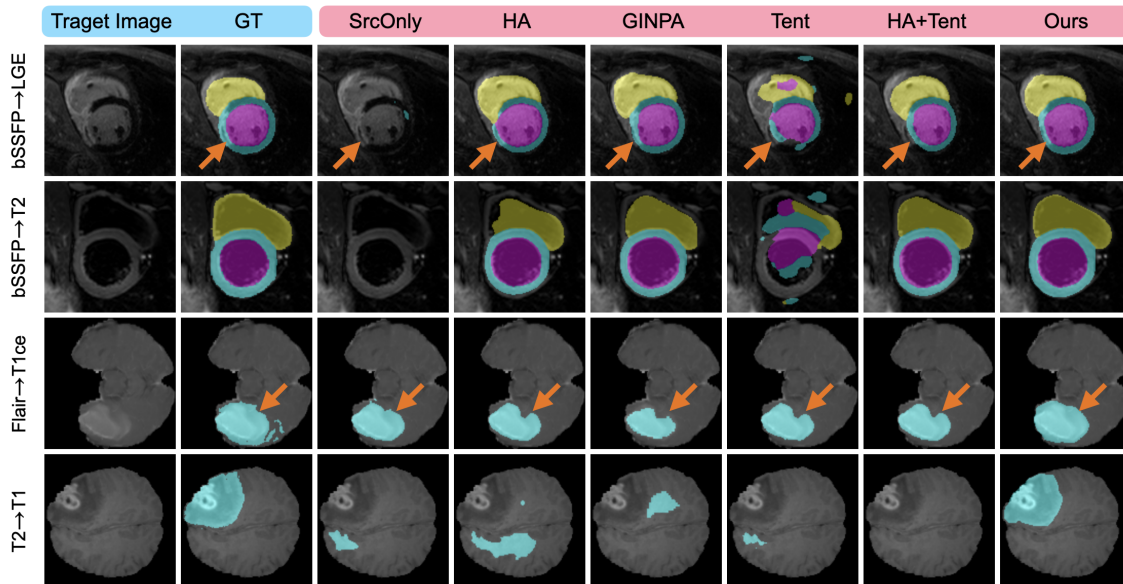
Figure 3: Comparison of cross-modality segmentation results between baselines and our method under four task settings. Our method achieves superior outcomes in these challenging cases.
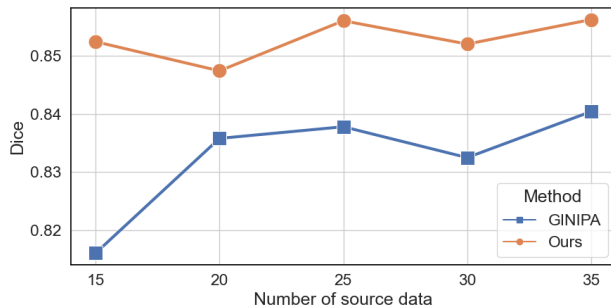

Figure 4: Comparisons on data efficiency of our model with GINIPA under bSSFP → LGE setting on MS-CMRSeg2019 dataset.

strategy for mitigating local domain shift. We extensively evaluate our method on two multi-modality MRI datasets with six cross-modal segmentation tasks. The results show the superior performance and efficiency of our method in comparison with previous domain generalization and TTA methods.

6. Acknowledgements

References

Amir Atapour-Abarghouei and Toby P Breckon. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2800–2810, 2018.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *The cancer imaging archive*, 286, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Thorsten Beier, Constantin Pape, Nasim Rahaman, Timo Prange, Stuart Berg, Davi D Bock, Albert Cardona, Graham W Knott, Stephen M Plaza, Louis K Scheffer, et al. Multicut brings automated neurite segmentation closer to human performance. *Nature methods*, 14(2):101–102, 2017.

Chen Chen, Wenjia Bai, Rhodri H Davies, Anish N Bhuva, Charlotte H Manisty, Joao B Augusto, James C Moon, Nay Aung, Aaron M Lee, Mihir M Sanghvi, et al. Improving the generalizability of convolutional neural network-based segmentation on cmr images. *Frontiers in cardiovascular medicine*, 7:105, 2020a.

Chen Chen, Cheng Ouyang, Giacomo Tarroni, Jo Schlemper, Huaqi Qiu, Wenjia Bai, and Daniel Rueckert. Unsupervised multi-modal style transfer for cardiac mr segmentation. In *Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmentation, CRT-EPiggy and LV Full Quantification Challenges: 10th International Workshop, STACOM 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Revised Selected Papers 10*, pages 209–219. Springer, 2020b.

Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. *IEEE transactions on medical imaging*, 38(5):1116–1126, 2018.

Xiaoting Han, Lei Qi, Qian Yu, Ziqi Zhou, Yefeng Zheng, Yinghuan Shi, and Yang Gao. Deep symmetric adaptation network for cross-modality medical image segmentation. *IEEE transactions on medical imaging*, 41(1):121–132, 2021.

Yufan He, Aaron Carass, Lianrui Zuo, Blake E Dewey, and Jerry L Prince. Autoencoder based self-supervised test-time adaptation for medical image analysis. *Medical image analysis*, 72:102136, 2021.

Donald Olding Hebb. *The organization of behavior: A neuropsychological theory.* Psychology press, 2005.

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.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. *arXiv preprint arXiv:1701.01036*, 2017.

Quande Liu, Qi Dou, and Pheng-Ann Heng. Shape-aware meta-learning for generalizing prostate mri segmentation to unseen domains. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part II 23*, pages 475–485. Springer, 2020.

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.

Xiao Liu, Spyridon Thermos, Agisilaos Chartsias, Alison O'Neil, and Sotirios A Tsaftaris. Disentangled representations for domain-generalized cardiac

segmentation. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11*, pages 187–195. Springer, 2021.

Christopher T Lloyd, Alessandro Sorichetta, and Andrew J Tatem. High resolution global gridded data for use in population studies. *Scientific data*, 4(1): 1–17, 2017.

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 (10):1993–2024, 2014.

Hugo Oliveira, Roberto M Cesar, Pedro HT Gama, and Jefersson A Dos Santos. Domain generalization in medical image segmentation via meta-learners. In *2022 35th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, volume 1, pages 288–293. IEEE, 2022.

Sebastian Otálora, Manfredo Atzori, Vincent Andrearczyk, Amjad Khan, and Henning Müller. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Frontiers in bioengineering and biotechnology*, page 198, 2019.

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.

Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.

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.

Junyi Qiu, Lei Li, Sihan Wang, Ke Zhang, Yinyin Chen, Shan Yang, and Xiahai Zhuang. Myops-net: Myocardial pathology segmentation with flexible

combination of multi-sequence cmr images. *Medical Image Analysis*, 84:102694, 2023.

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.

Yushun Tang, Ce Zhang, Heng Xu, Shuoshuo Chen, Jie Cheng, Luziwei Leng, Qinghai Guo, and Zhihai He. Neuro-modulated hebbian learning for fully test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3728–3738, 2023.

Vu-Hoang Tran and Ching-Chun Huang. Domain adaptation meets disentangled representation learning and style transfer. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pages 2998–3005. IEEE, 2019.

Sulaiman Vesal, Nishant Ravikumar, and Andreas Maier. Automated multi-sequence cardiac mri segmentation using supervised domain adaptation. In *International workshop on statistical atlases and computational models of the heart*, pages 300–308. Springer, 2019.

Sulaiman Vesal, Mingxuan Gu, Ronak Kosti, Andreas Maier, and Nishant Ravikumar. Adapt everywhere: unsupervised adaptation of point-clouds and entropy minimization for multi-modal cardiac image segmentation. *IEEE Transactions on Medical Imaging*, 40(7):1838–1851, 2021.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Wei Wang, Zhun Zhong, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, and Nicu Sebe. Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 24090–24099, June 2023a.

Wei Wang, Zhun Zhong, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, and Nicu Sebe. Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24090–24099, 2023b.

Qingsong Xie, Yuexiang Li, Nanjun He, Munan Ning, Kai Ma, Guoxing Wang, Yong Lian, and Yefeng Zheng. Unsupervised domain adaptation for medical image segmentation by disentanglement learning and self-training. *IEEE Transactions on Medical Imaging*, 2022.

Yanwu Xu, Shaoan Xie, Maxwell Reynolds, Matthew Ragoza, Mingming Gong, and Kayhan Batmanghelich. Adversarial consistency for single domain generalization in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 671–681. Springer, 2022.

Junlin Yang, Nicha C Dvornek, Fan Zhang, Julius Chapiro, MingDe Lin, and James S Duncan. Unsupervised domain adaptation via disentangled representations: Application to cross-modality liver segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 255–263. Springer, 2019.

Tao Zeng, Bian Wu, and Shuiwang Ji. Deepem3d: approaching human-level performance on 3d anisotropic em image segmentation. *Bioinformatics*, 33(16):2555–2562, 2017.

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.

Tongxue Zhou, Su Ruan, and Stéphane Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3:100004, 2019.

Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2933–2946, 2018.

Xiahai Zhuang, Jiahang Xu, Xinzhe Luo, Chen Chen, Cheng Ouyang, Daniel Rueckert, Victor M Campello, Karim Lekadir, Sulaiman Vesal, Nishant RaviKumar, et al. Cardiac segmentation on late gadolinium enhancement mri: a benchmark study from multi-sequence cardiac mr segmentation challenge. *Medical Image Analysis*, 81:102528, 2022.

Danbing Zou, Qikui Zhu, and Pingkun Yan. Unsupervised domain adaptation with dual-scheme fusion network for medical image segmentation. In *IJCAI*, pages 3291–3298, 2020.

Appendix A. PITTA Pseudo Code

Algorithm 1: Test Procedure of our method

Input: input image $\mathbf{x}$, class prior $\mathbf{p}$, α, λ, $\mathcal{C}_{lv}$
Output: Segmentation mask $\hat{\mathbf{y}}$
$\mathbf{q}^{(0)} \leftarrow \frac{\alpha}{C}\mathbf{1} + (1-\alpha)\mathbf{p}^{(0)}$
for $n \leftarrow 1, \mathbf{to} N$ **do**
    $\mathbf{S}^{(n)} \leftarrow \mathcal{M}_\theta(G(\mathbf{x}))$
    $\mathbf{S}'^{(n)} \leftarrow A_T^{-1}(\mathcal{M}_\theta(A_T(G(\mathbf{x}))))$
    $L_{cons} \leftarrow |\mathbf{S}^{(n)} - \mathbf{S}'^{(n)}|$
    $P^{(n)}(\mathbf{y}|\mathbf{x}) \leftarrow softmax((\mathbf{S}^{(n)} + \mathbf{S}'^{(n)})/2)$
    $\hat{\mathbf{y}}_{i,j}^{(n)} \leftarrow \arg\max\left(P^{(n)}(\mathbf{y}_{i,j}|\mathbf{x})/\mathbf{q}^{(n)}\right)$
    $\mathbf{M}_{i,j} \leftarrow \bigcup_{c \in \mathcal{C}_{lv}} \mathbf{1}(\hat{\mathbf{y}}_{i,j}^{(n)} = c)$
    $L_{mce} \leftarrow L_{ce}(P^{(n)}(\mathbf{y}|\mathbf{x}) \odot \mathbf{M}, \hat{\mathbf{y}}^{(n)})$
    $L_{total} \leftarrow L_{mce} + \lambda L_{cons}$
    Update θ using L_{total}
    $\mathbf{p}^{(n+1)} \leftarrow \alpha\mathbf{p}^{(n)} + (1-\alpha)\#\left\{\hat{\mathbf{y}}^{(n)}\right\}$
    $\mathbf{q}^{(n+1)} \leftarrow \frac{\alpha}{C}\mathbf{1} + (1-\alpha)\mathbf{p}^{(n+1)}$
end
$\hat{\mathbf{y}} \leftarrow \arg\max\left(P^{(N)}(\mathbf{y}|\mathbf{x})/\mathbf{q}^{(N)}\right)$

Appendix B. More discussion on related work

B.1. Our method vs. style-transfer-based techniques

Most style-transfer-based methods (Tran and Huang, 2019; Li et al., 2017; Atapour-Abarghouei and Breckon, 2018) require target-domain images to train the style-transfer function. In contrast, our approach is designed for domain generalization, where no target-domain information is available during training. Essentially, our method could be seen as defining a training-free transfer function via the gradient map, as in Eq. (1), to transfer MRI images to a domain-agnostic feature space.

Appendix C. More Implementation Details

C.1. Data Processing

For MS-CMRSeg2019 dataset, we resize the LGE and T2 modality to match the shape of bSSFP images and crop them into 128×128. For BraTS2018 dataset, we crop the images via MRI volume bounding box

(where intensity > 0) and resize them into 128×128. Unlike previous protocols (Xie et al., 2022; Zou et al., 2020; Han et al., 2021) which use random data partitioning, we designate the last 10% of HGG and LGG data in BraTS2018 as test set to facilitate a fair comparison for the subsequent work.

C.2. Model Architecture

We utilize the U-Net (Ronneberger et al., 2015) as our segmentation backbone and adapt the implementation from OpenAI [3] in our experiment. Our segmentation model comprises 15 encoder blocks, one middle block, and 15 decoder blocks with different resolutions (from 128×128 to 8×8). The detailed model structure can be found in the official implementation of OpenAI.

Appendix D. Additional Quantitative Results

D.1. Effect of Averaging Two logits

In order to demonstrate the effect of taking an average on two logits in Eq.(5), we empirically compared our method to using outputs from normal images alone. As demonstrated in Table 4, the averaged outputs yield a slight performance improvement.

Table 4: A comparison between using only the normal output (w/o prediction averaging) and averaging the outputs from both augmented and normal images (Ours). The experiments were conducted on the MS-CMRSeg2019 dataset, under the bSSFP→LGE setting.

	Myo	LV	RV	Average
Normal Output	0.7961	**0.9102**	0.8594	0.8552
Averaged (Ours)	**0.7987**	0.9100	**0.8600**	**0.8562**

D.2. Effect of Transformation Type on Test Time Adaptation

We assessed the performance impact of various transformation functions in test-time augmentation. As

---

3. https://github.com/openai/guided-diffusion

shown in Table 5, Horizontal Flip is the most efficient for performance improvement in our specific task.

Table 5: Transformation type comparison conducted on the MS-CMRSeg2019 dataset, under the bSSFP→LGE setting.

Transformation types	Myo	LV	RV	Average
Rotation 90°	0.7722	0.9040	0.8537	0.8433
ColorJitter	0.7963	0.9108	0.8473	0.8515
Gaussian Blur	0.7852	0.9089	0.8468	0.8470
Horizontal Flip	**0.7987**	**0.9100**	**0.8600**	**0.8562**
Vertical Flip	0.7916	0.9074	0.8565	0.8518
Horizontal Flip +ColorJitter	0.7980	0.9101	0.8521	0.8534
Horizontal Flip + Vertical Flip	0.7956	0.9090	0.8622	0.8556
Horizontal Flip +Rotation 90° +ColorJitter	0.7734	0.9039	0.8442	0.8405
Horizontal Flip +Rotation 90° +ColorJitter +Gaussian Blur	0.7697	0.9060	0.8415	0.8391

D.3. Effect of Pixel Mask

The pixel mask is selected based on the classes significantly impacted by local variations between domains. In our case, myocardium tissue (Myo) is notably affected due to variations like scar lesions being more visible in LGE scans compared to bSSFP. So we focus on Myo during test time to amplify the effects of these local variations. While domain expertise can be beneficial for pinpointing these key classes, it's not a restriction. If such expertise is lacking, practitioners can opt to include all classes by setting the pixel mask to 'true' for all. As shown in Table 6, this approach results in only a minor performance drop of approximately 0.1% in the Dice score.

D.4. Effect of Regularization Strength

We assessed the effect of the regularization strength parameter, λ, on model performance on MS-CMRSeg2019 dataset under bSSFP→T2 setting in Table 7, revealing that our method is relatively robust to changes in this hyperparameter.

Table 6: Performance comparison between models using pixel masks (Ours) and without pixel masks.

Dataset	BraTS				MS-CMRSeg2019	
Setting	T2→T1	T2→T1ce	Flair→T1	Flair→T1ce	bSSFP→LGE	bSSFP→2
Category	WT	WT	WT	WT	Avg	Avg
Ours w/o Mask	0.6803	0.6911	0.4091	0.5978	0.8500	0.8433
Ours	**0.6813**	**0.6914**	**0.4189**	**0.5986**	**0.8562**	**0.8555**

Table 7: Effect of λ on model performance. Results are averaged over all classes.

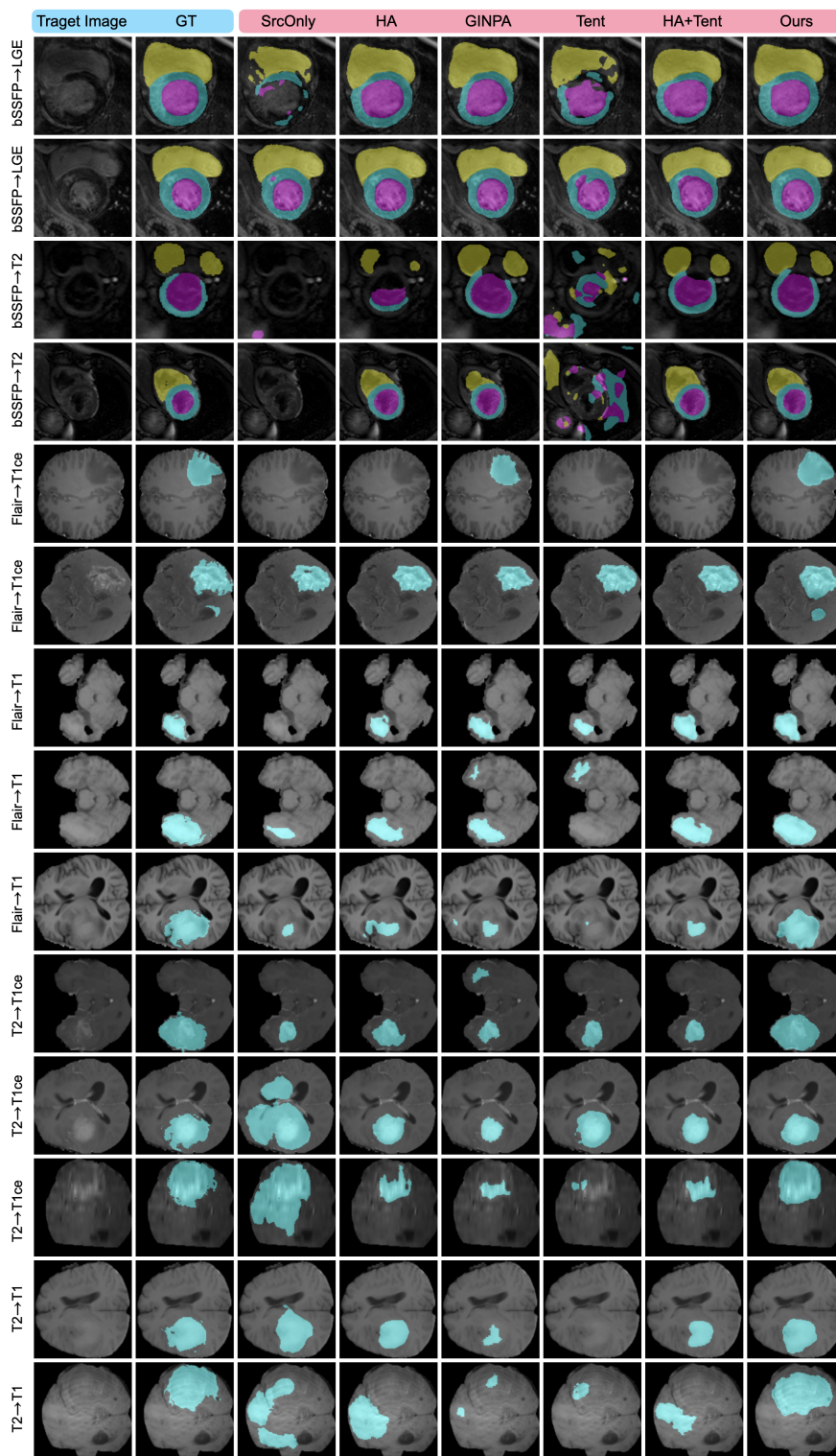| λ	0.1	0.3	0.5	1	1.5	2
Avg Dice	0.7671	**0.7738**	0.7724	0.7663	0.7617	0.7588

Appendix E. More Result Visualization

Figure 5: More segmentation visualization between different methods under various settings