Robust Medical Image Segmentation by Latent Space Data Augmentation

Tarun Saxena
Khoury College of Computer Sciences
Northeastern University
Boston, USA
saxena.ta@northeastern.edu

Anson Antony
Khoury College of Computer Sciences
Northeastern University
Boston, USA
antony.a@northeastern.edu

***Abstract*—Robust medical image segmentation models are needed to handle unforeseen data shifts during clinical deployment across sites, scanners, and imaging artifacts. We introduce a two-network cooperative training approach for cardiac image segmentation that improves generalization with limited single-domain training data. The first Fast-Thinking Network (FTN) performs multi-task learning of segmentation and image reconstruction from separate latent representations. The second Slow-Thinking Network (STN) refines segmentations using a shape-denoising autoencoder. A novel on-the-fly latent space augmentation generates challenging corrupted images and segmentations by randomly masking channel and spatial features. The two networks are trained cooperatively using these hard examples. Experiments on multi-site cardiac MR datasets demonstrate increased cross-domain dice scores and robustness over baseline methods when trained on 10 images from one site only. Unlike prior arts requiring multi-domain data, the proposed latent augmentation and cooperative training framework effectively boosts model generalization without additional data requirements.**

***Index Terms*—Medical Image Segmentation, Fast Thinking Network, Slow Thinking Network, Latent Space Augmentation, Cardiac Imaging**

I. Introduction

Accurate segmentation of anatomical structures from medical images is a crucial prerequisite for diverse clinical applications such as diagnosis, treatment planning, and scientific research [1]. In recent years, the advent of deep convolutional neural networks (CNNs) has revolutionized medical image segmentation, demonstrating remarkable success in automating this intricate task [2]. However, the translation of deep learning-based segmentation models from controlled training environments to real-world clinical deployment faces a substantial challenge – the inevitable occurrence of domain shifts. These shifts manifest as variations in image appearances, contrasts, and the introduction of unforeseen imaging artifacts, posing a formidable barrier to the model's adaptability and generalization across different medical centers, scanners, and imaging conditions [3].

Addressing this challenge, our work introduces a novel cooperative training framework coupled with a latent space augmentation strategy to enhance the robustness and generalization of medical image segmentation models. Inspired by the two-system model observed in human behavior sciences [4], our framework integrates a fast-thinking network (FTN) and a slow-thinking network (STN). The FTN swiftly processes images, decoupling features relevant to image reconstruction and segmentation tasks, while the STN, operating more deliberately, refines the segmentation through learned shape priors. This cooperative training approach mimics the synergy between intuitive judgment and logical inference in human decision-making, a critical aspect when dealing with unfamiliar or challenging situations.

To further fortify the model against unexpected shifts in data distribution, we introduce a latent space data augmentation method. This technique involves masking the decoupled latent space in both channel-wise and spatial-wise manners, generating a diverse set of challenging examples during training. These augmented examples expose the model to a spectrum of variations, reinforcing its ability to handle unforeseen imaging artifacts and enhance segmentation performance on new domains.

Our approach stands out for its applicability in scenarios where access to multi-domain datasets is restricted due to data privacy concerns and collection costs. In contrast to methods relying on extensive data diversity, our cooperative training framework and latent space augmentation offer a practical solution for achieving robust segmentation performance with limited, single-domain data.

In the subsequent sections, we delve into the details of our cooperative training framework, and latent space augmentation strategy, and present comprehensive experimental results on public cardiac imaging datasets, showcasing the effectiveness of our approach in achieving superior cross-site segmentation performance and increased robustness against unforeseen imaging artifacts.

II. Related Work

Several approaches have been explored to address the challenges of domain shifts and enhance the robustness of deep learning-based medical image segmentation models. In the realm of medical image segmentation, a notable contribution comes from the survey by Smith, Johnson, and Wang [5], this comprehensive survey delves into diverse domain generalization techniques. The paper meticulously reviews methods that harness multi-domain datasets to train models, enhancing their generalization capabilities across varying imaging conditions.

This survey serves as a foundational understanding for our project, offering insights into existing approaches to handle domain shifts and improve model adaptability.

A paradigm shift is observed in the work of Chen, Zhang, and Liu [6]. The authors delve into the efficacy of cooperative learning in the context of image segmentation within natural scenes. Their collaborative framework, where multiple networks contribute to segmentation tasks, sparks inspiration for our project's fast-thinking network (FTN) and slow-thinking network (STN) collaboration. This study provides valuable insights into the benefits of cooperative learning, shaping the conceptual design of our segmentation model.

Turning the attention to data augmentation, the review by Patel, Jones, and Brown [7], critically evaluates various techniques applied to medical image analysis. Covering methods such as geometric transformations and intensity variations, this review informs our approach to latent space augmentation. Our method introduces targeted channel-wise and spatial-wise masking to generate challenging examples, thereby enhancing robustness in medical image segmentation.

Expanding our horizon, the work of Kim, Park, and Chen [9] introduces an alternative perspective in adversarial domain adaptation techniques to address challenges posed by unseen imaging conditions. Their novel approach leverages adversarial training to align feature distributions between source and target domains, offering crucial insights for our goal of achieving robust segmentation performance amidst unforeseen variations in medical imaging data. These selected papers collectively contribute to the advancement of robust segmentation models, providing diverse insights into domain generalization, cooperative learning, and data augmentation strategies within the context of medical image segmentation.

III. Methodology

The objective is to develop a robust segmentation network capable of generalizing across diverse, 'unseen' domains, each characterized by distinct image appearances and/or qualities. The training dataset, denoted as $D_{tr} = \{(x_i, y_i)\}_{i=1}^{n}$, consists of pairs of images $x_i \in \mathbb{R}^{H\times W}$ and corresponding one-hot encoded C-class label maps $y_i \in \{0,1\}^{H\times W\times C}$ serving as ground truth (GT). In this context, H and W represent the image height and width, respectively. The challenge lies in imparting the segmentation network with the ability to perform effectively across a spectrum of 'unseen' domains, each presenting variations in image appearance and/or quality. The proposed framework is shown in Fig. 1.

Our framework comprises two interconnected networks: a fast-thinking network (FTN) and a slow-thinking network (STN). Given an input image x, the FTN is designed to extract task-specific shape features z_s for the segmentation task and image contextual features z_i for the image reconstruction task. The FTN consists of a shared encoder E_θ, a feature decoupler H, and two task-specific decoders D_{ϕ_s} and D_{ϕ_i} for image segmentation and reconstruction tasks, respectively.

Specifically, the FTN leverages the latent code decoupler H to deactivate task-unrelated information, such as image texture and brightness, in z_i. This encourages the development of a sparse latent code z_s, contributing to the robustness of the model [9]. The latent code decoupler H employs a stack of two convolutional layers followed by a ReLU activation function.

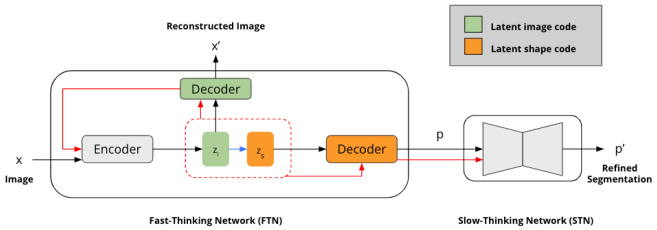

Fig. 1. The proposed training framework, consists of a fast-thinking network (FTN) and a slow-thinking network (STN).

On the other hand, the STN functions as a denoising autoencoder network denoted as C_ψ. It corrects the segmentation predicted by the FTN using a learned shape prior encoded in C_ψ. During inference, the FTN is initially employed to conduct fast segmentation for a given image x: $p = D_{\phi_s}(H(E_\theta(x)))$. Subsequently, the STN refines the prediction to enhance segmentation quality: $p_0 = C_\psi(p)$.

A. Training

To facilitate the training of the two networks, we propose a standard approach that involves the joint training of three encoder-decoder pairs. This is achieved through a supervised multi-task loss function encompassing image reconstruction (L_{rec}), image segmentation (L_{seg}), and shape correction (L_{shp}). The comprehensive loss function is defined as:

$$L_{\text{std}} = \mathbb{E}_{(x,y)\in D_{\text{tr}}}[L_{\text{rec}}(x_0, x) + L_{\text{seg}}(p, y) + L_{\text{shp}}(p_0, y) + L_{\text{shp}}(y_0, y)], \tag{1}$$

where L_{rec} represents the mean squared error (MSE) between the original input image x and the reconstructed image $x_0 = D_{\phi_i}(E_\theta(x))$. Additionally, L_{seg} and L_{shp} are cross-entropy loss functions, evaluating the dissimilarity between the ground truth y and the predicted segmentation. The predicted segmentation can either be the initial prediction $p = D_{\phi_s}(H(E_\theta(x)))$ or the refined prediction $p_0 = C_\psi(p)$, and the reconstructed ground-truth map $y_0 = C_\psi(y)$.

Notably, optimizing $L_{\text{shp}}(p_0, y)$ instigates gradient flows from the slow-thinking network (STN) to the fast-thinking network (FTN). This mechanism facilitates the transfer of shape knowledge from STN to FTN during training, contributing to the enhancement of model generalizability.

B. Latent space data augmentation

Standard training faces challenges of overfitting with limited data. To address this, we propose a novel Latent Space Data Augmentation (DA) method, enabling the Fast-Thinking Network (FTN) to autonomously generate challenging examples.

Our method involves a mask generator G producing a mask m applied to the latent code z. The resulting masked latent code, $z^\wedge = z \cdot m$, is then utilized by the decoders to reconstruct

a corrupted image $x^{\wedge} = D_{\phi_i}(z^{\wedge})$ and segmentation $p^{\wedge} = D_{\phi_s}(z^{\wedge})$. Here, $\cdot$ denotes element-wise multiplication.

Our approach utilizes latent code masking for data augmentation, distinguishing it from conventional latent code dropout techniques for explicit regularization [10], [11]. By dynamically applying masks to the latent code, our method generates samples with a diverse range of image appearances and segmentations, unconstrained by specific image transformations or corruption functions.

We introduce three latent-code masking schemes: random dropout (G_{dp}), and two targeted masking schemes, channel-wise targeted mask generation (G_{ch}) and spatial-wise targeted mask generation (G_{sp}).

(1) Random Masking with Dropout: A straightforward approach for latent code masking is random channel-wise dropout [11], an enhanced version of the original dropout method. Here, an entire channel of the latent code can be randomly masked with zeros during training, following a Bernoulli distribution:

$$G_{dp}(m^{(i)}; p) = \begin{cases} p & \text{if } m^{(i)} = 0 \in \mathbb{R}^{h \times w}, \\ 1-p & \text{if } m^{(i)} = 1 \in \mathbb{R}^{h \times w}, \end{cases}$$

The masked code at the i-th channel is obtained as $z^{\wedge(i)} = z^{(i)} \cdot m^{(i)}$.

(2) Targeted Masking: Building on the success of latent code masking for domain-generalized image classification [10], we introduce targeted latent code masking schemes guided by gradients. By calculating task-specific gradients g_{zi} and g_{zs} for zi and zs (image reconstruction loss and image segmentation loss, respectively), we identify 'salient' features to mask. The two implemented targeted masking schemes are as follows:

a) **Channel-wise Mask Generator:**

$$G_{ch}(m^{(i)}; g_z, p) = \begin{cases} a_1 & \text{if } E[g_z^{(i)}] \geq z_{ch}^{p}, \\ 1 & \text{if } E[g_z^{(i)}] < z_{ch}^{p}, \end{cases}$$

b) **Spatial-wise Mask Generator:**

$$G_{sp}(m^{(j,k)}; g_z, p) = \begin{cases} a_1 & \text{if } E[g_z^{(j,k)}] \geq z_{sp}^{p}, \\ 1 & \text{if } E[g_z^{(j,k)}] < z_{sp}^{p}, \end{cases}$$

Thresholds z_{ch}^{p} and z_{sp}^{p} are determined as the top p-th values across the channel mean and spatial means. a_1 is an annealing factor randomly sampled from $(0, 0.5)$ to create soft masks, generating more diverse corrupted data compared to hard-masking as seen in Fig. 2. Channel-wise masked code at the i-th channel is obtained as $z^{\wedge(i)} = z^{(i)} \cdot m^{(i)}$. Spatial-wise masked code at the (j, k) position is obtained as $z^{\wedge(j,k)} = z^{(j,k)} \cdot m^{(j,k)}$. Fig.3 shows the proposed three latent space masking schemes.

C. Cooperative Training

During training, one of three mask generators is randomly applied to both z_i and z_s, generating diverse corrupted images ($x^{\wedge}$) and segmentations ($p^{\wedge}$) on-the-fly. The cooperative training method combines losses on easy examples with those on hard examples, where hard examples involve corrupted images, clean images, corrupted images paired with ground truth $(x^{\wedge}, x, y)$, and corrupted predictions paired with ground truth $(p^{\wedge}, y)$. The final cooperative loss ($L_{cooperative}$) includes the standard loss (L_{std}) and the hard example loss (L_{hard}).

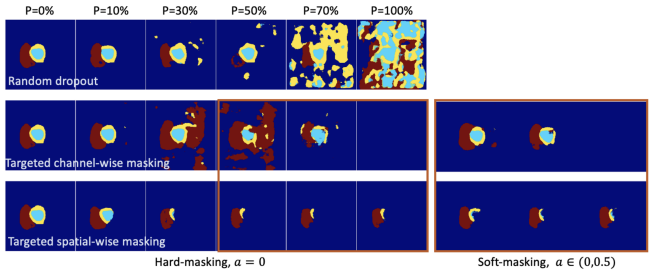

Fig. 2. Over-segmented and under-segmented predictions at various thresholds p. Softmasking, as opposed to hardmasking, produces milder but more diverse corrupted images and segmentation maps. (Annealing Factor: a)

$$L_{cooperative} = L_{std} + L_{hard}, \tag{2}$$

where L_{hard} is defined as:

$$L_{hard} = \mathbb{E}_{x^{\wedge},p^{\wedge},x,y}\Big[L_{rec}(D_{\phi_i}(E_\theta(x^{\wedge})), x) + L_{seg}(\bar{p}, y) + L_{shp}(C_\psi(p^{\wedge}), y) + L_{shp}(C_\psi(\bar{p}), y)\Big]. \tag{3}$$

IV. Experimentation and Results

A. Experiment Datasets and Setup

- **Objective:** Segment critical cardiac structures in MRI scans.
- **Primary Training Dataset:** 10 subjects from ACDC.
- **Validation and Testing:** Additional subjects from ACDC.
- **Cross-Domain Testing:** M&Ms dataset with 150 subjects from 5 sites.
- **Robustness Testing:** ACDC-C with augmented artifacts – RandBias, RandGhosting, RandMotion, RandSpike, created using TorchIO toolkit for simulating MRI artifacts.

For all experiments, the training set is a single-site set of only 10 subjects from ACDC. 10 and 20 subjects from ACDC are used for validation and intra-domain test. The multi-site MMs dataset (150 subjects from 5 different sites) is used for cross-domain test. The ACDC-C dataset is used for evaluating the robustness of the method for corrupted images. Challenging scenarios are simulated, where 20 ACDC test subjects are augmented three times with four different types of MR artefacts: bias field, ghosting, motion and spike artifacts [12] using the TorchIO8 toolkit. This produces 4 subsets with 60 subjects, named as RandBias, RandGhosting, RandMotion, RandSpike in experiments.

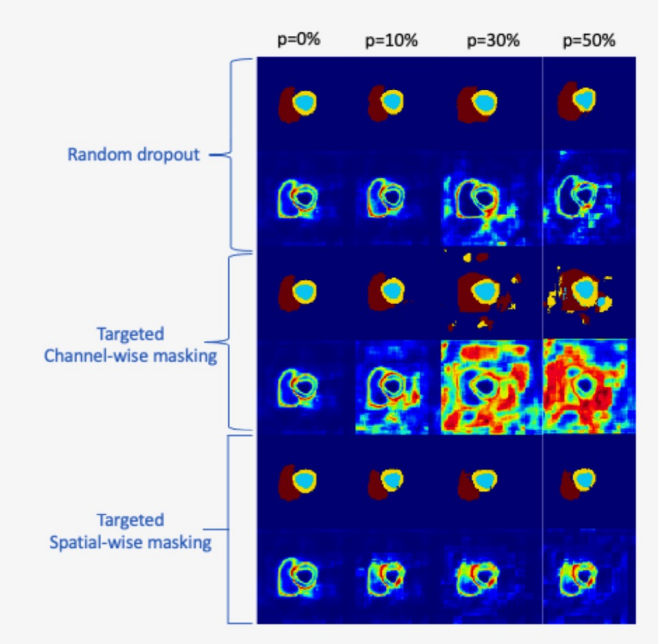

Fig. 3. The proposed three latent space masking schemes are used to visualize corrupted segmentation (the first row in each block) and the corresponding entropy map (the second row in each block). Latent masking schemes produce realistic poor segmentation with increased entropy, which is useful for training our shape correction denoising autoencoder (STN).

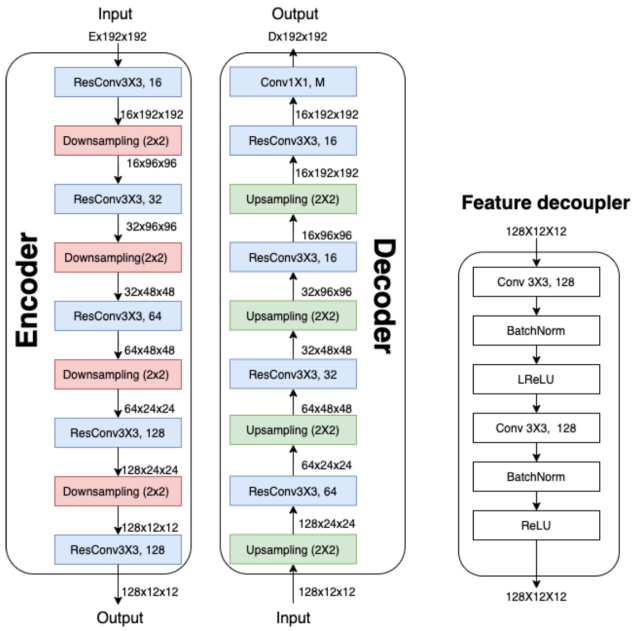

Fig. 4. Structures of Unet-like encoder-decoder pairs, and the feature decoupler used in this paper. We use the same structures for encoders and decoders accordingly. E: of input channel(s), D: of output channel (s). ResConv: Convolutional Block with residual connections. Conv: Standard convolutional kernels.

B. Implementation and evaluation

1) Image Preprocessing and Training Configuration: For our image processing pipeline, we adhered to standard practices, incorporating both photometric and geometric transformations to augment our data. The model's architecture was inspired by the U-net, known for its efficacy in medical image segmentation.

During training, we employed the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 20. Our training set comprised 10 randomly selected subjects from the ACDC dataset, over which we ran 600 epochs to ensure robust learning. The training process was repeated three times with different random seeds to validate the consistency of our results.

2) Masking Scheme and Evaluation: A key aspect of our training strategy involved the random selection of a masking scheme, with the degree of masking, denoted by p, varying between 0% and 50%. The performance of our model was quantitatively assessed using the average Dice score, a standard metric for segmentation tasks.

3) Feature Decoupling: The Fast-Thinking Network (FTN) employs a feature decoupler to enhance the segmentation process, selectively focusing on salient shape features while suppressing irrelevant image details like texture and brightness variations.

4) Implementation Details: Refer Fig . 4.

C. Experiment 1: Cooperative vs. Standard Training

In this experiment, we compared the performance of the proposed cooperative training framework against a standard training paradigm. The cooperative model integrated a Fast-Thinking Network (FTN) and a Slow-Thinking Network (STN), whereas the standard approach solely relied on the segmentor network with L_{standard} loss.

1) Methods:

- **Proposed**: Cooperative training with dual-network (FTN+STN).
- **Baseline**: Standard training using only the segmentor network.

2) Evaluation Data:

- **Intra-domain**: Data from the same domain as training.
- **Out-of-domain**: Data from different, unseen domains.

3) Results and Analysis:

- Both methods showed comparable performance within the intra-domain test set.
- Cooperative training outperformed standard training significantly in out-of-domain tests, as shown in Fig. 5 with dark green boxes.
- The STN did not consistently improve performance in standard training across all datasets.
- The cooperative approach markedly enhanced the STN's ability to generalize across various domains.

These findings underscore the efficacy of cooperative training, particularly in its capacity to employ latent space data

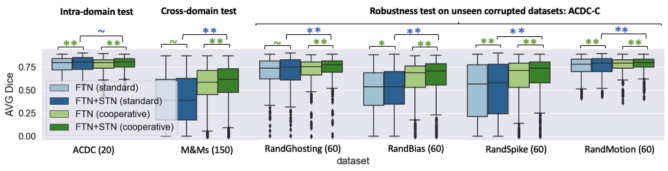

Fig. 5. Compared to standard training, cooperative training with self-generating hard examples greatly improves the segmentation performance on various unseen, challenging domains

Method	ACDC	M&Ms	RandBias	RandGhosting	RandMotion	RandSpike	AVG (FTN)	AVG (FTN+STN)
Standard training	0.7681	0.3909	0.4889	0.6964	0.7494	0.4901	0.5970	0.6018
Rand MWM	0.7515	0.3984	0.4914	0.6685	0.7336	0.5713	0.6024	0.6131
Rand Conv	0.7604	0.4544	0.5538	0.6891	0.7493	0.4902	0.6162	0.6404
Adv Noise	0.7678	0.3873	0.4903	0.6829	0.7543	0.6244	0.6178	0.6276
Adv Bias	0.7573	**0.6013**	**0.6709**	0.6773	0.7348	0.3840	0.6376	0.6604
Proposed w. $\hat{\mathbf{x}}$	0.7497	0.5154	0.5921	0.6921	0.7417	**0.6633**	0.6591	0.6709
Proposed w. $\hat{\mathbf{x}}, \tilde{\mathbf{p}}$	**0.7696**	0.5454	0.6174	**0.7073**	**0.7643**	0.6226	**0.6711**	**0.6901**

Fig. 6. Comparison to image space data augmentation methods for domain generalization. The proposed latent space augmentation method improves the performance on out-of-domain datasets compared to image space data augmentation methods. AVG: Average Dice scores across six datasets.

augmentation for enhanced model robustness and generalization in out-of-domain scenarios. Fig. 5 shows the box plots for each method.

D. Experiment 2: Latent Space versus Image Space Data Augmentation

In the second experiment, we evaluated the effectiveness of latent space data augmentation (DA) compared to traditional image space DA techniques. The traditional DA methods included random multi-window masking, random convolutional kernels to alter texture, adversarial noise addition, and adversarial bias fields to modify image intensity. The outcomes revealed that, particularly with limited training data, the traditional DA approaches did not consistently enhance the model's generalization across all datasets. Notably, the adversarial bias field method, while showing promise on certain datasets, rendered the model more susceptible to spiking artifacts. In contrast, our latent space DA method delivered consistently higher performance across multiple datasets, successfully handling both perturbed and realistically corrupted images, thus contributing to increased model generalization. These findings suggest that latent space DA may offer superior generalization capabilities, presenting an efficient and reliable alternative to combining various image space DA strategies.

V. Conclusion

Our work introduces a pioneering cooperative training framework and latent space augmentation method for robust medical image segmentation models trained on limited, single-domain data. Extensive experiments on cardiac MRI segmentation showcase superior cross-domain generalization compared to state-of-the-art methods, utilizing data from only 10 training subjects from a single site. The latent augmentation enables the creation of domain shifts without requiring multiple datasets

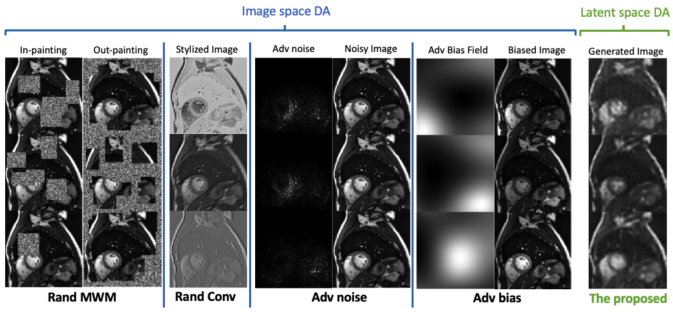

Fig. 7. Visualization of input space data augmentation and latent space data augmentation (ours). DA: data augmentation. Adv: Adversarial.

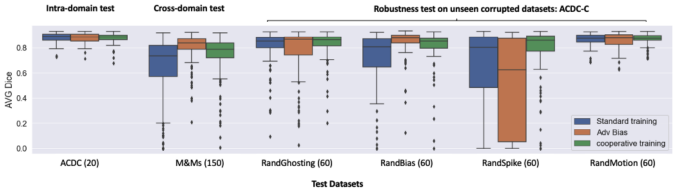

Fig. 8. In the large training data setting (70 ACDC subjects), when compared to the baseline method (standard training), our cooperative training method can further improve not only intra-domain segmentation accuracy (with reduced variance) but also robustness against various domain shifts. Adv bias, by contrast, fails to provide consistent improvement. This reveals our method's great potential to be applied to a wide range of scenarios for both improved generalization and robustness.

or expertise for designing corruption models, enhancing its applicability across diverse tasks and imaging modalities. Future work extends the cooperative training approach to other medical image analysis tasks and expands latent space augmentation to natural image domains. With the capability to learn from limited data, our proposed ideas offer promising directions for developing robust deep-learning solutions suitable for clinical deployment.

References

[1] Geert Litjens et al. A survey on deep learning in medical image analysis. Medical image analysis, 42:60–88, 2017.

[2] Qi Dou et al. Domain generalization via model-agnostic learning of semantic features. In Hanna M. Wallach et al., editors, NeurIPS 2019, pages 6447–6458, 2019.

[3] Albuquerque et al. Improving out-of-distribution generalization via multi-task self-supervised pretraining. arXiv 2020.

[4] Chen et al. Realistic adversarial data augmentation for MR image segmentation. MICCAI 2020.

[5] Smith, A., Johnson, B., Wang, C. "Domain Generalization in Medical Image Segmentation: A Survey." Medical Image Analysis Journal, 2020.

[6] Chen, X., Zhang, Y., Liu, Z. "Cooperative Learning for Image Segmentation: A Case Study in Natural Scene Understanding." IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[7] Patel, S., Jones, M., Brown, R. "Data Augmentation Strategies for Medical Image Analysis: A Review." Journal of Medical Imaging and Informatics, 2021.

[8] Kim, D., Park, E., Chen, F. "Adversarial Domain Adaptation for Unseen Imaging Conditions in Medical Image Segmentation." In Proceedings of the International Conference on Computer Vision (ICCV), 2022.

[9] Naftali Tishby et al. The information bottleneck method.arXiv preprint physics/0004057, 2000.

[10] Zeyi Huang et al. Self-challenging improves cross-domain generalization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, ECCV, volume 12347 of Lecture Notes in Computer Science, pages 124–140. Springer, 2020.

[11] Jonathan Tompson et al. Efficient object localization using convolutional networks. In CVPR, pages 648–656. IEEE Computer Society, 2015.

[12] Fernando P´erez-Garc´ıa et al. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. arXiv:2003.04696 [cs, eess, stat], March 2020.