Realistic Adversarial Data Augmentation for MR Image Segmentation

Chen Chen[1](✉), Chen Qin[1], Huaqi Qiu[1], Cheng Ouyang[1], Shuo Wang[3], Liang Chen[1,4], Giacomo Tarroni[1,2], Wenjia Bai[3,4], Daniel Rueckert[1]

[1] BioMedIA Group, Department of Computing, Imperial College London, UK
[2] CitAI Research Centre, Department of Computer Science, City, University of London, UK
[3] Data Science Institute, Imperial College London, UK
[4] Department of Brain Sciences, Imperial College London, UK
chen.chen15@imperial.ac.uk

Abstract. Neural network-based approaches can achieve high accuracy in various medical image segmentation tasks. However, they generally require large labelled datasets for supervised learning. Acquiring and manually labelling a large medical dataset is expensive and sometimes impractical due to data sharing and privacy issues. In this work, we propose an adversarial data augmentation method for training neural networks for medical image segmentation. Instead of generating pixel-wise adversarial attacks, our model generates plausible and realistic signal corruptions, which models the intensity inhomogeneities caused by a common type of artefacts in MR imaging: bias field. The proposed method does not rely on generative networks, and can be used as a plug-in module for general segmentation networks in both supervised and semi-supervised learning. Using cardiac MR imaging we show that such an approach can improve the generalization ability and robustness of models as well as provide significant improvements in low-data scenarios.

Keywords: Image segmentation, Adversarial data augmentation, MR

1 Introduction

Segmentation of medical images is an important task for diagnosis, treatment planning and clinical research [1]. Recent years have witnessed the fast development of deep learning for medical imaging with neural networks being applied to a variety of medical image segmentation tasks [2, 3]. Deep learning-based approaches in general require a large-scale labelled dataset for training, in order to achieve good model generalization ability and robustness on unseen test cases. However, acquiring and manually labelling such large medical datasets is extremely challenging, due to the difficulties that lie in data collection and sharing, as well as to the high labelling costs [4].

To address the aforementioned problems, one of the commonly adopted strategies is data augmentation, which aims to increase the diversity of the

available training data without collecting and manually labelling new data. Conventional data augmentation methods mainly focus on applying simple *random* transformations to labelled images. These random transformations include intensity transformations (e.g. pixel-wise noise, image brightness and contrast adjustment) and geometric transformations (e.g. affine, elastic transformations). Recently, there is a growing interest in developing generative network-based methods for data augmentation [5, 6, 7, 8], which have been found effective for one-shot brain segmentation [5] and low-shot cardiac segmentation [7]. Unlike conventional data augmentation, which generates new examples in an uninformative fashion and does not account for complex variations in data, this generative network-based method is data-driven, learning optimal image transformations from the underlying data distribution in the real world [7]. However, in practice, training generative networks is not trivial due to their sensitivity to hyper-parameters tuning [9] and it can suffer from the mode collapse problem.

In this work, we introduce an effective adversarial data augmentation method for medical imaging without resorting to generative networks. Specifically, we introduce a realistic intensity transformation function to amplify intensity non-uniformity in images, simulating potential image artefacts that may occur in clinical MR imaging (i.e. bias field). Our work is motivated by the observations that MR images often suffer from low-frequency intensity corruptions caused by inhomogeneities in the magnetic field. This artefact cannot be easily eliminated [10, 11] and can be regarded as a physical attack to neural networks, which have been reported to be sensitive to intensity perturbations [12, 13]. To efficiently improve the model generalizability and robustness, we apply adversarial training to directly search for optimal intensity transformations that benefit model training. By continuously generating these realistic, 'hard' examples, we prevent the network from over-fitting and, more importantly, encourage the network to defend itself from intensity perturbations by learning robust semantic features for the segmentation task.

Our main contributions can be summarised as follows: (1) We introduce a realistic adversarial intensity transformation model for data augmentation in MRI, which simulates intensity inhomogeneities which are common artefacts in MR imaging. The proposed data augmentation is complementary to conventional data augmentation methods. (2) We present a simple yet effective framework based on adversarial training to learn adversarial transformations and to regularize the network for segmentation robustness, which can be used as a plug-in module in general segmentation networks. More importantly, unlike conventional adversarial example construction [14, 15, 16], generating adversarial bias fields does not require manual labels, which makes it applicable for both supervised and semi-supervised learning, see Sec. 2.2. (3) We demonstrate the efficacy of the proposed method on a public cardiac MR segmentation dataset in challenging low-data settings. In this scenario, the proposed method greatly outperforms competitive baseline methods, see Sec. 3.2.

Related work. Recent studies have shown that adversarial data augmentation, which generates adversarial data samples during training, is effective to improve

model generalization and robustness [15, 17]. Most existing works are based on designing attacks with pixel-wise noise, i.e. by adding gradient-based adversarial noise [14, 18, 19, 20, 21]. More recently, there have been studies showing that neural networks can also be fragile to other, more natural form of transformations that can occur in images, such as affine transformations [22, 23, 24], illumination changes [24], and small deformations [13, 25]. In medical imaging, designing and constructing realistic adversarial perturbations, which can be used for improving medical image segmentation networks, has not been explored in depth.

2 Adversarial Data Augmentation with Robust Optimization

In this work, we aim at generating realistic adversarial examples to improve model generalization ability and robustness, given a limited number of training examples. To achieve the goal, we first introduce a physics-based intensity transformation model that can simulate intensity inhomogeneities in MR images. We then propose an adversarial training method, which finds effective adversarial transformation parameters to augment training data, and then regularizes the network with a distance loss function which penalizes network's sensitivity to such adversarial perturbations. Since our method is based on virtual adversarial training (VAT) [20], we will first briefly review VAT before introducing our method.

2.1 Virtual Adversarial Training

VAT is a regularization method based on adversarial data augmentation, which can prevent the model from over-fitting and improve the generalization performance and robustness [20]. Given an input image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ (H,W,C denote image height, width, and number of channels, respectively) and a classification network $f_{cls}(\cdot;\theta)$, VAT first finds a small adversarial noise $\mathbf{r}^{\text{adv}} \in \mathbb{R}^{H \times W \times C}$ to construct its adversarial example $\mathbf{I}^{adv} = \mathbf{I} + \mathbf{r}^{\text{adv}}$ (as shown in Fig.1A), with the goal of maximising the Kullback−Leibler (KL) divergence $\mathcal{D}_{\text{KL}}$ between an original probabilistic prediction $f_{cls}(\mathbf{I};\theta)$ and its perturbed prediction $f_{cls}(\mathbf{I}+\mathbf{r}^{\text{adv}};\theta)$. The adversarial example is then used to regularize the network for robust feature learning.

The adversarial noise can be generated by taking the gradient of $\mathcal{D}_{\text{KL}}$ with respect to a random noise vector: $\mathbf{r}^{\text{adv}} = \epsilon \cdot \frac{r'}{\|r'\|_2}$,$r' = \nabla_{\mathbf{r}} \mathcal{D}_{\text{KL}}[f(\mathbf{I};\theta) \parallel f(\mathbf{I}+\mathbf{r};\theta)]$. Here ϵ is a hyper-parameter that controls the strength of perturbation. After finding adversarial examples, one can utilize them for robust learning, which penalizes the network's sensitivity to local perturbations. This is achieved by adding $\mathcal{D}_{\text{KL}}$ to its main objective function.

2.2 Adversarial Training by Modelling Intensity Inhomogeneities

In this work, we extend the VAT approach by introducing a new type of adversarial attack, namely intensity inhomogeneities (bias field) that often occur in

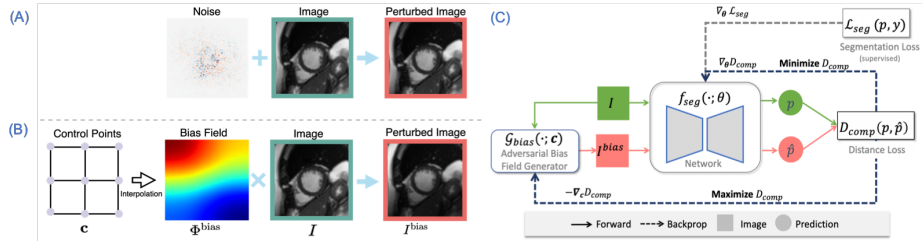


Fig. 1: (A) Adversarial example construction with additive gradient-based noise in VAT [20]; (B) Adversarial example construction with a multiplicative control point-based bias field (proposed); (C) Adversarial training with bias field perturbation.

MR imaging. In MR imaging, a bias field is a low frequency field that smoothly varies across images, introducing intensity non-uniformity across the anatomy being imaged. The model for the intensity non-uniformity can be defined as follows [10, 26]: $\mathbf{I}^{\text{bias}} = \mathcal{G}_{\text{bias}}(\mathbf{I}; \mathbf{c}) = \mathbf{I} \times \mathit{\Phi}^{\text{bias}}(\mathbf{c})$. Here, the intensity of the image $\mathbf{I}$ is perturbed by a multiplication with the bias field $\mathit{\Phi}^{\text{bias}} \in \mathbb{R}^{H \times W}$. As the bias field is typically composed of low frequencies and thus slowly varying across the image, it can be modelled using a set of uniformly distributed k by k points $\mathbf{c} = \{\mathbf{c}_{(i)}\}_{1 \ldots k \times k}$ [10], see Fig. 1B. A smooth bias field at the finest resolution is obtained by interpolating scattered control points with a third-order B-spline smoothing [27].

While one can repeatedly sample random bias fields for data augmentation, this might be computationally inefficient as it may generate images which are of no added value for model optimization. We therefore would like to construct adversarial examples (perturbed by bias field as described above) targeting the weakness of the network in an intelligent way. This allows the use of the generated adversarial examples to improve the model performance and robustness, which can be achieved via the following min-max game:

$$\begin{aligned} \min_{\theta} \max_{\mathbf{c}} \quad & \mathcal{D}_{\text{comp}}[f_{seg}(\mathbf{I}; \theta), f_{seg}(\mathcal{G}_{\text{bias}}(\mathbf{I}; \mathbf{c}); \theta)] \\ \text{subject to} \quad & \forall (x, y) \in \mathbb{R}^2,\ \mathit{\Phi}^{\text{bias}}_{(x,y)} > 0;\ |\mathit{\Phi}^{\text{bias}} - \mathbf{1}|_{\infty} \leq \alpha, 0 < \alpha < 1. \end{aligned} \tag{1}$$

As shown in Fig. 1C, given a segmentation network $f_{seg}(\cdot; \theta)$ and an input image $\mathbf{I}$, we first find optimal values for control points $\mathbf{c}$ in the search space to construct an adversarial bias field, so that it **maximizes** the distance measured by $\mathcal{D}_{\text{comp}}$ between the original prediction and the prediction after perturbation: $\mathbf{p} = f_{seg}(\mathbf{I}; \theta), \hat{\mathbf{p}} = f_{seg}(\mathcal{G}_{\text{bias}}(\mathbf{I}; \mathbf{c}); \theta)$, with θ fixed. We then optimize the parameters θ in the network to **minimize** the distance between the original prediction and the prediction after the generated adversarial bias attack $f_{seg}(\mathcal{G}_{\text{bias}}(\mathbf{I}; \mathbf{c}^{adv}); \theta)$.

Finding adversarial bias fields. To find the optimal values for the control points $\mathbf{c}$ for adversarial example construction, we use the gradient descent algorithm and search the values of control points in its log space for numerical

stability [10,26], which allows to produce positive bias fields. Specifically, similar to the projected gradient decent (PGD) attack construction in [15], we first randomly initialize the values of control points and then apply a projected gradient ascent algorithm to iteratively update $\mathbf{c}$ with n steps: $\mathbf{c} \leftarrow \Pi(\mathbf{c} + \xi \cdot \mathbf{c}'/\|\mathbf{c}'\|_2)$ where $\mathbf{c}' = \nabla_{\mathbf{c}} \mathcal{D}_{\text{comp}}[f_{seg}(\mathbf{I};\theta), f_{seg}(\mathcal{G}_{\text{bias}}(\mathbf{I};\mathbf{c});\theta)]$. Π denotes the projection function which projects $\mathbf{c}$ onto the feasible set, and ξ is the step size. For neural networks, gradients $\mathbf{c}'$ can be efficiently computed with back-propagation. Φ^{bias} is updated by first interpolating the coarse-grid control points (log values at the current iteration) to its finest grid using B-spline convolution, and then taking the exponential function for value recovering. Finally, the generated bias field is rescaled to meet the magnitude constraint in Eq. 1.

Composite distance function $\mathcal{D}_{\text{comp}}$. Here, we propose a composite distance function $\mathcal{D}_{\text{comp}}$ to enhance its discrimination ability between the original prediction $\mathbf{p}$ (short for $f_{seg}(\mathbf{I};\theta)$) and the prediction after perturbation $\hat{\mathbf{p}}$, for *semantic segmentation* tasks. This composite loss consists of (1) the original $\mathcal{D}_{\text{KL}}$ used in VAT, which measures the difference between distributions and (2) a contour-based loss function $\mathcal{D}_{\text{contour}}$ [28] which is specifically designed to capture mismatch between object boundaries: $\mathcal{D}_{\text{comp}}(\mathbf{p}, \hat{\mathbf{p}}) = \mathcal{D}_{\text{KL}}[\mathbf{p} \,||\, \hat{\mathbf{p}}] + w\mathcal{D}_{\text{contour}}(\mathbf{p}, \hat{\mathbf{p}})$; $\mathcal{D}_{\text{contour}}(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{m \in M} \sum_{S_{x,y}} \|S(\mathbf{p}^m) - S(\hat{\mathbf{p}}^m)\|_2$. M denotes foreground channels, $S_{x,y}$ denote two Sobel filters in x- and y-direction for edge extraction and w controls the relative importance of both terms.

Optimizing segmentation network. After constructing the adversarial examples, one can compute $\mathcal{D}_{\text{comp}}$ and apply it to regularizing the network, encouraging the network to be less sensitive to adversarial perturbations, and thus produce consistent predictions. Since this algorithm uses probabilistic predictions (produced by the network) rather than manual labels for adversary construction, it can be applied to both labelled (l) and unlabelled data (u) for supervised and semi-supervised learning [20]. The loss functions for the two scenarios are defined as: $\mathcal{L}_{\text{SU}} = \mathcal{L}_{\text{seg}}(\mathbf{p}^{(l)}, \mathbf{y}_{gt}^{(l)}) + \lambda_l \mathcal{D}_{\text{comp}}(\mathbf{p}^{(l)}, \hat{\mathbf{p}}^{(l)})$; $\mathcal{L}_{\text{SE}} = \mathcal{L}_{\text{SU}} + \lambda_u \mathcal{D}_{\text{comp}}(\mathbf{p}^{(u)}, \hat{\mathbf{p}}^{(u)})$. $\mathcal{L}_{\text{seg}}$ denotes a general task-related segmentation loss function for supervised learning (e.g. cross-entropy loss) and $\mathbf{y}_{gt}^{(l)}$ denotes ground truth.

3 Experiments

To test the efficacy of the proposed method, we applied it to training a segmentation network for the left ventricular myocardium from MR images in low-data settings. We compared the results with several competitive baseline methods.

3.1 Dataset and Experiment Settings

ACDC dataset. Experiments were performed on a public benchmark dataset for cardiac MR image segmentation: The Automated Cardiac Diagnosis Challenge (ACDC) dataset [29] [5]. This dataset was collected from 100 subjects which

[5] https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html

were evenly classified into 5 groups: 1 normal group (NOR) and 4 pathological groups with cardiac abnormalities: dilated cardiomyopathy(DCM); hypertrophic cardiomyopathy (HCM); myocardial infarction with altered left ventricular ejection fraction (MINF); abnormal right ventricle (ARV). The left ventricular myocardium in end-diastolic and end-systolic frames were manually labelled.

Image pre-processing. We used the same image preprocessing as in [7], where all images were bias corrected using N4 algorithm [10]. In addition, all images were centrally cropped into 128×128, given that the heart is generally located in the center of the image. This saves computational costs.

Random data augmentation (Rand Aug). We applied a strong random data augmentation method to our training data as a basic setting. Random affine transformation (i.e. scaling, rotation, translation), random horizontal and vertical flipping, random global intensity transformation (brightness and contrast) [7] and elastic transformation were applied.

Training details. For ease of comparison, same as [7], we adopted the commonly-used 2D U-net as our segmentation network, which takes 2D image slices as input. The Adam optimizer with a batch size of 20 was used to update network parameters. For the proposed method, we first trained the network with the default data augmentation (Rand Aug) for 10,000 iterations (learning rate=$1e^{-3}$), and then finetuned the network by adding the proposed adversarial training using a smaller learning rate ($1e^{-5}$) for 2,000 iterations. The common standard cross-entropy loss function was used as $\mathcal{L}_{\text{seg}}$. For bias field construction, we adopted the B-spline convolution kernel (order=3) with 4×4 control points. The kernel was provided by AirLab library [30]. We empirically set: $\alpha = 0.3$, $w = 0.5$, $\lambda_l = 1$ and $\lambda_u = 0.1$. Besides, we found that in our experiments, one step searching in the inner loop produced sufficient improvement. Thus, we set $n = 1, \xi = 1$ to save computational cost. All the experiments were performed on an Nvidia® GeForce® 2080 Ti with Pytorch.

3.2 Experiments and Results

Experiment 1: Low-shot learning. In this experiment, the proposed method was evaluated in both *supervised* learning and *semi-supervised* learning scenarios, where only 1 or 3 labelled subjects are available. Specifically, we used the same data splitting setting as in [7]. The ACDC dataset was split into 4 subsets: a labelled set (where N_l images were sampled from for training), unlabelled training set (N=25), validation set (N=2), test set (N=20). N denotes the number of subjects. Details of the low-data setting can be found in [7]. For one-shot learning (N_l=1) and three-shot learning (N_l=3) in both supervised and semi-supervised settings, we trained the network for five times, each with a different labelled set.

We compared the proposed method (**Adv Bias**) with several competitive data augmentation methods including **VAT** [20], an effective data mixing-based method (**Mixup**) [31] for supervised learning and the state-of-the-art semi-supervised generative model-based method(**cGANs**) [7]. For VAT and Mixup, we used the set of hyperparameters that achieved the best performance on the

Table 1: Comparison of the proposed method (Adv Bias) to other data augmentation methods.

Setting	Method	# labelled subjects 1	3
Supervised	No Aug	0.293	0.544
	Rand Aug	0.560	0.796
	+Mixup [31]	0.575	0.801
	+VAT [20]	0.570	0.811
	+Adv Bias	**0.650**	**0.826**
Semi-supervised	+VAT [20]	0.625	0.826
	+Adv Bias	0.692	**0.830**
	cGANs [7]	**0.710**	0.823

Table 2: Segmentation performance of the proposed method and baseline methods across five populations. All were trained with NOR cases only.

Population	Rand Aug	+Mixup	+VAT	**+Adv Bias** (Proposed)
NOR	0.911	0.901	0.909	**0.912**
DCM	0.831	0.803	0.843	**0.871**
HCM	0.871	0.881	**0.891**	0.890
MINF	0.805	0.789	0.824	**0.847**
ARV	0.843	0.844	0.843	**0.853**
Average	0.841	0.833	0.853	**0.868**

validation set and applied the same training procedure. For cGANs, we report the results of one-shot and three-shot learning in their original paper for reference, which were tested on the same test set. Table 1 compares the segmentation accuracy obtained by different data augmentation methods. Each reported value is the average Dice score of 20 test cases. In the supervised learning setting (no access to unlabelled images), when only one or three labelled subject was available, the proposed method clearly outperformed all baseline methods. For semi-supervised learning, the proposed methods outperformed VAT, especially when only one labelled subject is available (0.686 vs 0.625). The proposed method achieves competitive results compared to the semi-supervised GAN-based method (cGANs) as well. Of note, cGANs adopts two additional GANs to sample geometric transformations and intensity transformations from unlabelled images. This is why it was only compared in the semi-supervised learning setting here. On the contrary, our approach is applicable to both low-shot supervised learning and semi-supervised learning. In addition, cGANs contains more parameters than our method and thus it might be less computationally efficient.

Experiment 2: Learning from limited population. In this experiment, we trained the network using only normal healthy subjects (NOR) and evaluated its performance on pathological cases. 20 healthy subjects were split into 14/2/4 subjects for training, validation and test. This setting simulates a practical data scarcity problem, where pathological cases are rarer, compared to healthy data. As shown in Table 2, while the conventional method (Rand Aug) achieved excellent performance on the test healthy subjects (NOR), its performance dropped on pathological cases. Interestingly, applying Mixup did not help to solve this population shift problem, but rather slightly reduced the average performance compared to the baseline, from 0.841 to 0.833. This might be due to the fact that Mixup generates unrealistic images through its linear combination of paired images, which may modify semantic features and affect representation learning for *precise* segmentation. By contrast, our method outperformed both Mixup and VAT, yielding substantial and consistent improvements across five different populations. Notably, we attained evident improvement on the most challenging

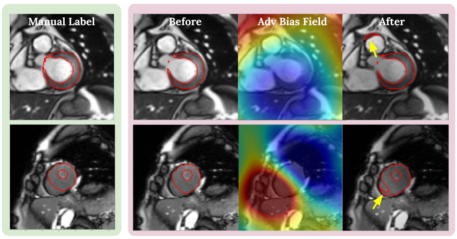


Fig. 2: Visualization of generated adversarial examples and failed network predictions. Before/After: network prediction before/after bias field attack (Adv Bias Field).

MINF images (0.805 vs 0.847), where the shape of the myocardium is clearly irregular. As shown in Fig. 2, the proposed method does not only generate adversarial examples during training, but also increases the variety of image styles while preserving the shape information. Augmenting images with various styles can encourage the network to learn high-level shape-based representation instead of texture-based representation, leading to improved network robustness on unseen classes, as discussed in [32]. By contrast, VAT only introduces imperceptible noise, failing to model realistic image appearance variations.

Ablation study. To get a better understanding of the effectiveness of adversarial bias field, we compared it to data augmentation using random bias field, using experiment setting 2. Results clearly showed that training with adversarial bias field improved the model generalization ability, increasing the Dice score from 0.852 to 0.868. On the other hand, applying $\mathcal{D}_{\text{comp}}$ to regularize the network improved the average Dice score from 0.859 to 0.868, compared to the one trained with only $\mathcal{D}_{\text{KL}}$. Unlike random-based approach, constructing adversarial attacks considers both the posterior probability information estimated by the model and semantic information from images. In experiments, we found these attacks focused on attacking challenging images on which the network was uncertain, e.g. object boundary is not clear or there is another similar structure presented, see Fig. 2. In the same spirit of online hard example mining, utilizing these borderline examples during training helps the network to improve its generalization and robustness ability. Please find more details in the supplementary material.

4 Discussion and Conclusion

In this work, we presented a realistic adversarial data augmentation method to improve the generalization and robustness for neural network-based medical image segmentation. We demonstrated that by modelling bias field and introducing adversarial learning, the proposed method is able to promote the learning of robust semantic features for cardiac image segmentation. It can also alleviate the data scarcity problem, as demonstrated in the low-data setting and

cross-population experiments. The proposed method does not rely on generative networks but instead employs a small set of explainable and controllable parameters to augment data with image appearance variations which are realistic for MR. It can be easily extended for multi-class segmentation and used in general segmentation networks for improving model generalization and robustness.

Acknowledgements. This work was supported by the SmartHeart EPSRC Programme Grant (EP/P001009/1). HQ was supported by the EPSRC Programme Grant (EP/R005982/1).

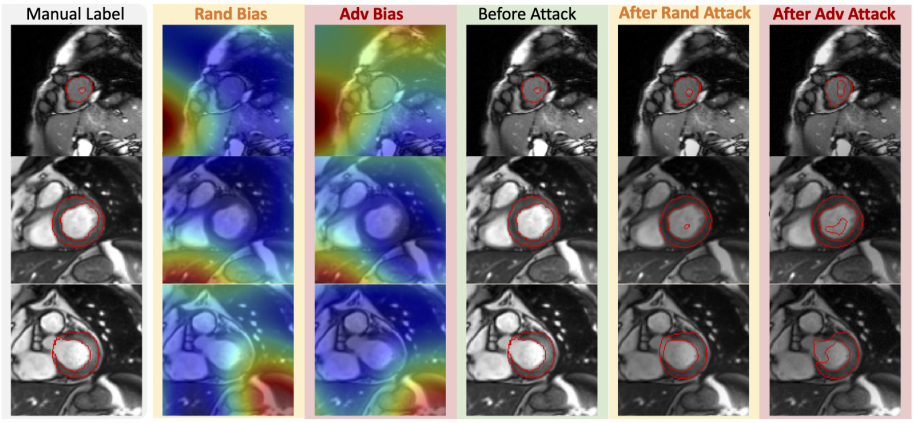
References

1. Erik Smistad, Thomas L Falch, Mohammadmehdi Bozorgi, Anne C Elster, and Frank Lindseth. Medical image segmentation on gpus–a comprehensive review. *Medical image analysis*, 20(1):1–18, 2015.
2. Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, June 2017.
3. Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.
4. Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey Chiang, Zhihao Wu, and Xiaowei Ding. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Arxiv*, August 2019.
5. Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *CVPR*, pages 8543–8553, 2019.
6. Jianfei Liu, Christine Shen, Tao Liu, Nancy Aguilera, and Johnny Tam. Active appearance model induced generative adversarial network for controlled data augmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 201–208. Springer International Publishing, 2019.
7. Krishna Chaitanya, Neerav Karani, Christian F Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu. Semi-Supervised and Task-Driven data augmentation. In *International Conference on Information Processing in Medical Imaging*, pages 29–41. Springer, 2019.
8. Yunyan Xing, Zongyuan Ge, Rui Zeng, Dwarikanath Mahapatra, Jarrel Seah, Meng Law, and Tom Drummond. Adversarial pulmonary pathology translation for pairwise chest x-ray data augmentation. In *MICCAI*, October 2019.
9. Na Lei, Dongsheng An, Yang Guo, Kehua Su, Shixia Liu, Zhongxuan Luo, Shing-Tung Yau, and Xianfeng Gu. A geometric understanding of deep learning. *Engineering*, 2020.
10. Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: improved N3 bias correction. *TMI*, 29(6):1310–1320, June 2010.
11. Nadieh Khalili, Nikolas Lessmann, Elise Turk, N Claessens, Roel de Heus, Tessel Kolk, Max A Viergever, Manon JNL Benders, and Ivana Išgum. Automatic brain tissue segmentation in fetal mri using convolutional neural networks. *Magnetic resonance imaging*, 64:77–89, 2019.

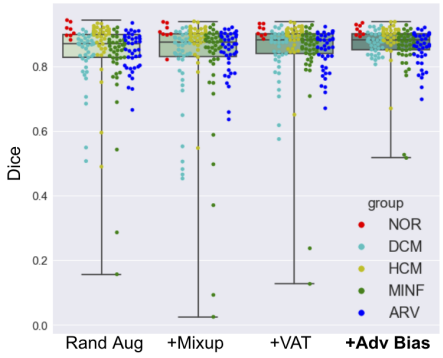
12. Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. robustness: adversarial examples for medical imaging. In *MICCAI*, 2018.
13. Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Intelligent image synthesis to attack a segmentation CNN using adversarial learning. In *Simulation and Synthesis in Medical Imaging - 4th International Workshop, SASHIMI 2019, Held in Conjunction with MICCAI 2019*, pages 90–99, 2019.
14. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.
15. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, June 2017.
16. Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018.
17. Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C. Duchi, Vittorio Murino, and Silvio Savarese. Generalizing to unseen domains via adversarial data augmentation. In *NeurIPS*, pages 5339–5349, 2018.
18. Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017.
19. Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. In *NIPS*, April 2019.
20. Takeru Miyato, Shin-Ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: A regularization method for supervised and Semi-Supervised learning. *TPAMI*, 2018.
21. Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. robustness: Adversarial examples for medical imaging. In *MICCAI*, March 2018.
22. Can Kanbak, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Geometric robustness of deep networks: Analysis and improvement. In *CVPR*, pages 4441–4449, 2018.
23. Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 1802–1811, Long Beach, California, USA, 2019. PMLR.
24. Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi-Keung Tang, and Alan L. Yuille. Adversarial attacks beyond the image space. In *CVPR*, pages 4302–4311, 2019.
25. Rima Alaifari, Giovanni S. Alberti, and Tandri Gauksson. Adef: an iterative algorithm to construct adversarial deformations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

26. John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.
27. Jean Gallier and Jean H Gallier. *Curves and surfaces in geometric modeling: theory and algorithms*. Morgan Kaufmann, 2000.
28. 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 *International Workshop on Statistical Atlases and Computational Models of the Heart*, pages 209–219. Springer, 2019.
29. Olivier Bernard, Alain Lalande, and Pierre-Marc others. Deep learning techniques for automatic MRI cardiac Multi-Structures segmentation and diagnosis: Is the problem solved? *TMI*, 0062(11):2514–2525, November 2018.
30. Robin Sandkühler, Christoph Jud, Simon Andermatt, and Philippe C Cattin. AirLab: Autograd image registration laboratory. *Arxiv*, June 2018.
31. Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018.
32. Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

Supplementary material: Realistic Adversarial Data Augmentation for MR Image Segmentation



Supple. Figure 1: **Performance of adversarial bias field attack (Adv Bias) vs random bias field attack (Rand Bias)**



Supple. Figure 2: **Segmentation accuracy of each method for cardiac myocardium segmentation across five different populations.** Each dot represents the Dice score for each test subject, and its color indicates its group. Our method (column 4) produces more accurate segmentation on **unseen** pathological cases than the baselines. This indicates that the proposed method can improve the model robustness for abnormal cases, even the network was only trained with normal cases (NOR).

Supple. Table 1: $\mathcal{D}_{\text{comp}}$ vs $\mathcal{D}_{\text{KL}}$

Method	Distance Loss	Dice	HD	VolumeSim
VAT	$\mathcal{D}_{\text{KL}}$	0.853	6.678	0.949
VAT	$\mathcal{D}_{\text{comp}}$	0.856	6.331	0.946
Adv Bias	$\mathcal{D}_{\text{KL}}$	0.859	6.330	0.949
Adv Bias	$\mathcal{D}_{\text{comp}}$	**0.868**	**5.912**	**0.957**

HD: Hausdorff distance; VolumeSim: Volume similarity index [**?**]. Reported values are average scores across all test subjects from five populations ($20 \times 4 + 4 = 84$ subjects). The same applies to Table 2.

Supple. Table 2: Random bias field vs Adversarial bias field

Method	Distance Loss	Dice	HD	VolumeSim
Rand Bias	$\mathcal{D}_{\text{comp}}$	0.852	6.25	0.941
Adv Bias	$\mathcal{D}_{\text{comp}}$	**0.868**	**5.91**	**0.957**