Edinburgh Research Explorer

Disentangled Representations for Domain-generalized Cardiac Segmentation

Disentangled Representations for Domain-generalized Cardiac Segmentation

Xiao Liu[1], Spyridon Thermos[1], Agisilaos Chartsias[1], Alison O'Neil[1,3] and
Sotirios A. Tsaftaris[1,2]

[1] School of Engineering, University of Edinburgh, Edinburgh EH9 3FB, UK
[2] The Alan Turing Institute, London, UK
[3] Canon Medical Research Europe Ltd., Edinburgh, UK
Xiao.Liu@ed.ac.uk, SThermos@ed.ac.uk, Agis.Chartsias@ed.ac.uk,
Alison.ONeil@eu.medical.canon, S.Tsaftaris@ed.ac.uk

Abstract. Robust cardiac image segmentation is still an open challenge
due to the inability of the existing methods to achieve satisfactory per-
formance on unseen data of different domains. Since the acquisition and
annotation of medical data are costly and time-consuming, recent work
focuses on domain adaptation and generalization to bridge the gap be-
tween data from different populations and scanners. In this paper, we
propose two data augmentation methods that focus on improving the
domain adaptation and generalization abilities of state-to-the-art car-
diac segmentation models. In particular, our "Resolution Augmenta-
tion" method generates more diverse data by rescaling images to dif-
ferent resolutions within a range spanning different scanner protocols.
Subsequently, our "Factor-based Augmentation" method generates more
diverse data by projecting the original samples onto disentangled latent
spaces, and combining the learned anatomy and modality factors from
different domains. Our extensive experiments demonstrate the impor-
tance of efficient adaptation between seen and unseen domains, as well
as model generalization ability, to robust cardiac image segmentation.[4]

Keywords: Cardiac image segmentation · Data augmentation · Disen-
tangled factors mixing · Domain adaptation · Domain generalization

1 Introduction

Recent advances in machine and deep learning, as well as in computational hard-
ware, have significantly impacted the field of medical image analysis [14], and
more specifically the task of cardiac image segmentation [5]. Since cardiovascular
disease continues to be a major public health concern, accurate automatic seg-
mentation of cardiac structures is an important step towards improved analysis
for prevention and diagnosis. Most existing deep learning methods rely on fully or
semi-supervised setups, using pixel-level annotations provided by clinical experts
for model training. These methods achieve satisfactory performance in predicting

---

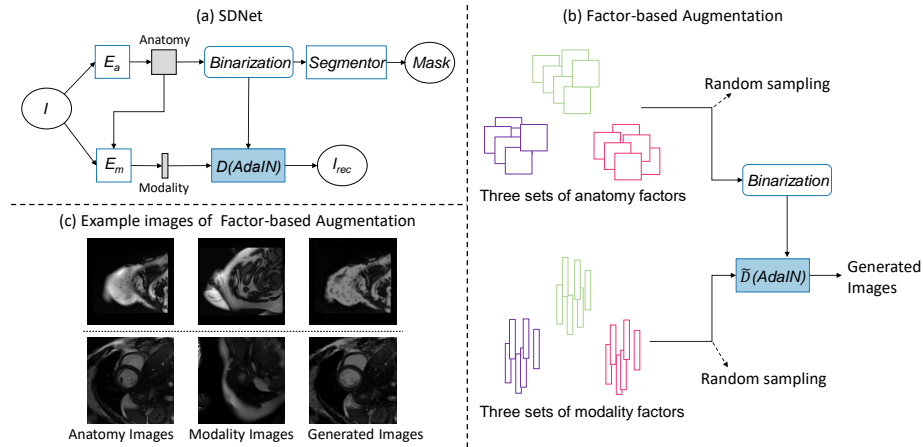[4] The code will be made publicly available.

Fig. 1. (a) SDNet: E_a: anatomy encoder, E_m: modality encoder, $D(AdaIN)$: AdaIN decoder. I is the input image to the model and I_{rec} is the output image of the AdaIN decoder *i.e.* the reconstructed image. $Mask$ is the predicted segmentation mask for the input image. (b) Illustration of Factor-based Augmentation: $\tilde{D}(AdaIN)$ is a pretrained AdaIN decoder. (c) Example images produced by Factor-based Augmentation. Anatomy Images provide anatomy factors, Modality Images provide modality factors, and Generated Images are the combination of the anatomy and modality factors.

left ventricle (LV), right ventricle (RV), and left ventricular myocardium (MYO) binary masks from unseen data within the same domain. However, such models struggle to achieve similar performance on unseen data from other domains, *e.g.* captured by another scanner or from a different population [19].

Better generalization of models is an open problem in computer vision [6, 7]. Learning a model with good generalization to the target domain is termed *domain adaptation* when labeled samples from the source domain and unlabeled samples from the target domain are given, or *domain generalization* when only labeled samples from the source domain are given.

Annotating more medical images is difficult due to time, cost, and privacy constraints. Therefore, researchers focus on bridging the gap between different domains by learning more general representations from the available data in a semi-supervised manner, *i.e.* domain adaptation [11, 20]. Existing methods in cardiac image segmentation achieve domain generalization either by disentangling spatial (anatomy) from imaging (modality) factors [3], or by augmenting the available data [2,4,21]. However, there is limited work [8] focusing on mixing disentangled representations, *i.e.* factors, to augment the original data.

In this paper, we propose two data augmentation methods, termed Resolution Augmentation (RA) and Factor-based Augmentation (FA), which are combined to improve domain adaption and generalization, thus boosting the performance of state-of-the-art models in Cardiac Magnetic Resonance (CMR) image segmentation. In particular, we use RA to remove the resolution bias by randomly
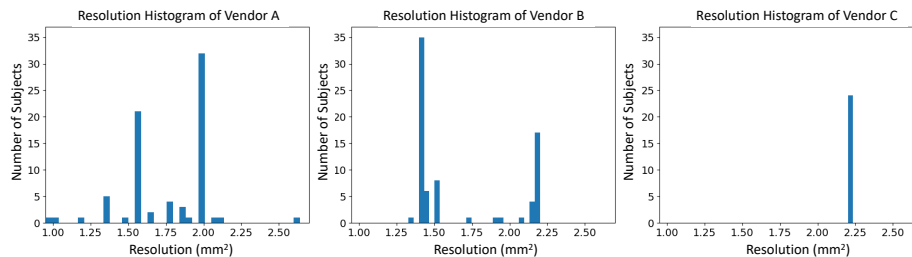
Fig. 2. Resolution histograms of the M&Ms challenge training data, broken down by vendor (from left to right: Vendors A, B and C).

rescaling the training images within a predetermined resolution range, while FA is used to increase diversity in the labeled data through mixing spatial and imaging factors, which we denote as the *anatomy* and *modality* factors, respectively. To extract these factors for FA, we pre-train the SDNet model introduced in [3] using the original (prior to augmentation) data. Experiments on the diverse dataset from the STACOM 2020 Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image Segmentation Challenge (M&Ms challenge) show the superiority of the proposed augmentation methods when combined with the state-of-the-art U-Net [17] and SDNet models.

2   Method

We train the SDNet model and employ RA and FA to generate a more diverse dataset. The model and FA setup is illustrated in Fig. 1.

2.1   Proposed augmentation methods

Resolution Augmentation (RA): It is common in MRI data that the imaging resolution is different for each study due to variation of the scanner parameters. Variation in the imaging resolution can cause the cardiac anatomy to vary significantly in size (*i.e.* area in pixels), beyond normal anatomical variability.

The training dataset contains subjects scanned by scanners from three vendors *i.e.* Vendors A, B and C. In Fig. 2, we show histograms of the training dataset image resolutions (Sec. 3 has more details on the dataset). We observed that the histograms of subjects imaged by scanners from different vendors are distinct from one other *i.e.* this is a bias with respect to the dataset. To reduce this bias, we propose to augment the training dataset such that the resolutions of subjects are equally distributed from 0.954 mm^2 to 2.692 mm^2 per pixel (the minimum and maximum values observed in the data from the 3 vendors), by rescaling the original image to a random resolution in this range. Finally we center-crop the rescaled image to uniform dimensions of 224×224 pixels.

Factor-based Augmentation (FA): As shown in Fig. 1(a), SDNet decomposes the input image into two latent representations, *i.e.* anatomy and modality factors, respectively. In particular, the anatomy factor contains spatial information about the input, and the modality factor contains non-spatial information only, namely imaging specific characteristics. Ideally, the anatomy factors would not encode the variation caused by different scanners, rather this information would be encoded in the modality factors. Motivated by this, we propose to augment the training dataset by combining different anatomy and modality factors to generate new data.

Considering the three sets of data from Vendors A, B and C, we first pre-train a SDNet model in a semi-supervised manner using the original data. Using this model at inference, we decompose the three sets of data into three sets of anatomy and modality factors. As shown in Fig. 1(b), we randomly sample an anatomy factor from the three anatomy factor sets and a modality factor from the three modality factor sets. A new image can be generated by processing the two factors with the decoder of the pre-trained SDNet model. By repeating this augmentation process, we generate a larger and more diverse dataset that contains segmentation annotations. The segmentation mask of the generated data is the mask of the image providing the anatomy factor, *if* the image is labeled, otherwise the generated data is unlabeled. Some indicative examples of FA are visualized in Fig. 1(c).

2.2   Model Architecture

As depicted in Fig. 1(a), the SDNet model consists of 4 basic modules, namely the anatomy encoder E_a, the modality encoder E_m, the segmentor, and the AdaIN-based decoder $D(AdaIN)$.

The **anatomy encoder** is realized as a U-Net network that consists of 4 downsampling and upsampling convolutional layers coupled with batch normalization [10] layers and ReLU [15] non-linearities. The output feature of the anatomy encoder has 8 channels, while the feature values are thresholded to 0 and 1 by a differentiable rounding operator. By adopting the thresholding constraint and supervision provided by the segmentation masks, the encoded anatomy factor is forced to contain more spatial information about the image.

The **modality encoder** consists of 2 downsampling convolutional layers (4×4 kernel size) that are followed by a global averaging pooling layer, which is used to eliminate the spatial (anatomical) information. The output of the pooling layer is then projected to an 8-dimensional vector (modality factor) using a Multi-Layer Perceptron (MLP) network.

The **segmentor** has 2 convolutional layers using 3×3 kernels, coupled with batch normalization and ReLU activation layers, as well as a 1×1 convolution followed by a channel-wise softmax function. The input to the segmentor is the thresholded anatomy factor. The output target of this segmentor is the ground-truth segmentation masks when the masks are available. In the learning process, this segmentor encourages the anatomy encoder to encode more spatial

information about the image such that the segmentor can learn to predict the corresponding masks more efficiently.

Finally, for the **AdaIN decoder**, we use the AdaIN module as in [9], in order to combine the anatomy and modality representations to reconstruct the image. In particular, the decoder consists of 3 convolutional layers (3×3 kernel size) coupled with adaptive instance normalization and ReLU activation layers. A final convolutional layer with 7×7 kernels is used for the reconstruction, followed by a hyperbolic tangent activation function that normalizes the values of the generated image into the [0,1] range. As discussed in [9], the AdaIN decoder normalizes the anatomy factor by firstly applying instance normalization to significantly remove the non-spatial information, then allowing the modality factor to define the new mean and standard derivation of the normalized anatomy factor. In this way, the decoder encourages the anatomy factor to contain spatial information and also force the modality factor to contain non-spatial information. On the other hand, by using AdaIN, the decoder does not simply ignore the modality factor that has a much smaller dimensionality than that of the anatomy factor.

2.3   Objective Function and Model Training

Apart from the original SDNet objective functions, we additionally use focal loss as presented in [13]. Focal loss is widely used in segmentation tasks and helps the model to achieve better accuracy by addressing the class imbalance problem. The augmented overall objective function is defined as:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{z_{rec}} + \lambda_3 \mathcal{L}_{dice} + \lambda_4 \mathcal{L}_{focal}, \tag{1}$$

where $\mathcal{L}_{rec}$ is the $L1$ distance between the input and the reconstructed image, $\mathcal{L}_{z_{rec}}$ denotes the $L1$ distance between the sampled and the re-encoded modality vector (latent regression), $\mathcal{L}_{dice}$ is the segmentation dice loss [18], and $\mathcal{L}_{focal}$ is the segmentation focal loss. Since we train SDNet in both fully supervised and semi-supervised setups, we set the hyperparameters $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$ when training using labeled data, while when training using unlabeled data we set $\lambda_1 = \lambda_2 = 1$ and $\lambda_3 = \lambda_4 = 0$.

3   Experiment

3.1   Dataset description and preprocessing

We train and validate our proposed method on the M&Ms challenge dataset of 350 subjects. Some subjects have hypertrophic and dilated cardiomyopathies (and some are healthy) but disease labels are not provided. Subjects were scanned in clinical centres in 3 different countries using 4 different magnetic resonance scanner vendors *i.e.* Vendor A, B, C and D in this paper. The M&Ms

Challenge training dataset contains 75 **labeled** subjects scanned using technology from Vendor A, 75 **labeled** subjects scanned by Vendor B, and 25 **unlabeled** subjects scanned by Vendor C.[5] The M&Ms challenge test dataset contains 200 subjects (50 from each vendor, including the seen 25 unlabeled subjects from Vendor C). From these, 80 subjects (20 from each vendor) are used to validate the model (results reported in Table 1) and the rest will be used for final challenge rankings. Subjects scanned by Vendor D were unseen during model training.

For each subject, we have 2D cardiac image slice acquisitions captured at multiple phases in the cardiac cycle (including end-systole and end-diastole). We train on the 2D images independently because we adopt 2D convolution neural networks in our model. Following [1], we normalize the training data intensity distribution to a Gaussian distribution with a mean of 0 and a standard deviation of 1. Overall, there are 1,738 pairs of images and masks from A and 1,546 pairs of images and masks from B. Apart from these labeled images, there are 47,346 unlabeled images from A, B and C.

3.2   Model Training

Models are trained using the Adam [12] optimizer with an initial learning rate of 0.001. To stabilize the training, we set the new learning rate to 10% of the previous rate when the validation Dice similarity coefficient between the predicted and the ground-truth masks does not improve for 2 consecutive epochs. Following the original training setting of SDNet, we set the batch size to 4 and train the model for 50 epochs. All models are implemented in PyTorch [16] and trained using an NVidia 1080 Ti GPU.

3.3   Results and Discussions

To verify the effectiveness of the proposed augmentation methods, we train 5 models for the purpose of ablation: **a)** the U-Net model using samples augmented with RA (U-Net+RA), **b)** the original SDNet model trained in a fully supervised fashion (FS SDNet), **c)** the fully supervised SDNet model using samples augmented with RA (FS SDNet+RA), **d)** the SDNet model trained in a semi-supervised fashion, using samples augmented with RA (SS SDNet+RA), and **e)** the semi-supervised SDNet using samples augmented with both FA and RA (SS SDNet+RA+FA). We train the fully supervised models with labeled samples from vendors A and B, while in the semi-supervised scenario we use all available data (labeled and unlabeled) for training. Table 1 reports the per vendor average Dice scores.

Since we are only allowed to validate our model a limited number of times, the results are not comprehensive, therefore we will do pairwise analysis below.

---

[5] We denote subjects scanned by scanners of Vendor A and B as labeled. However, it is notable that only the end-diastole and end-systole phases are labeled.

Table 1. Evaluation of the 5 models. Average Dice similarity coefficients are reported. Bold numbers denote the best performances across the 5 models. LV: left ventricle, MYO: left ventricular myocardium and RV: right ventricle.

Model	Vendor A			Vendor B			Vendor C			Vendor D		
	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV	LV	MYO	RV
U-Net+RA	0.900	0.829	0.811	0.937	0.877	0.907	0.856	0.837	0.852	0.762	0.651	0.503
FS SDNet	0.901	0.837	0.822	0.942	0.877	0.920	0.851	0.826	**0.853**	0.734	0.618	0.474
FS SDNet+RA	0.905	0.846	0.828	**0.945**	0.886	**0.921**	0.855	0.843	0.843	**0.819**	0.749	**0.624**
SS SDNet+RA	**0.909**	**0.854**	**0.841**	**0.945**	**0.887**	0.916	0.843	0.837	0.813	0.811	**0.752**	0.553
SS SDNet+RA+FA	**0.909**	0.846	0.778	0.939	0.882	0.909	**0.863**	**0.847**	0.843	0.812	0.712	0.498

Does RA help? By inspection of the **FS SDNet** and **FS SDNet+RA** results, we observe that for Vendor A and B, RA helps to achieve better overall performance compared to the respective baseline and RA substantially boosts performance on the unseen Vendor D. Subsequently, for Vendor C, the models have similar performance in the LV class, while FS SDNet+RA achieves the best performance in the MYO class (0.843 Dice score). However, the models using RA do not perform well in the RV class. In the case of Vendor C (no labelled examples), we know that the sample resolution of 24 out of 25 subjects made available at training time from Vendor C is 2.203 mm^2, and the resolution of 32 out of 75 Vendor A training subjects is around 2.000 mm^2, thus we argue that the resolution bias for Vendor A and Vendor C are already similar in the original data, and becoming invariant to this bias by rescaling the Vendor A data does not further help the performance on Vendor C.

Does FA help? Inspection of **SS SDNet+RA** and **SS SDNet+RA+FA** results shows that FA has mixed performance. It performs well on Vendor C for the scenario of domain adaptation (target samples available but unlabeled), where Vendor C is one of the vendors providing modality factors. However, FA performs poorly on Vendor D for domain generalization (unseen target), where Vendor D is not involved in the augmentation process.

Best model: Out of our two augmentation methods, we observe that RA has the most reliable performance, and we choose to submit the **FS SDNet+RA** model for final evaluation in the M&Ms challenge. We can see by comparing **U-Net+RA** with **FS SDNet+RA** that our chosen architecture of SDNet has good generalization performance compared to a standard U-Net, and this is particularly evident for the unseen vendor D, supporting our choice of the disentangled SDNet architecture even when FA is not employed.

Improvements to FA: Satisfyingly, our FA method yields a benefit for Vendor C (whilst RA did not give a significant benefit to Vendor C). However, FA does not give considerable improvements on domain generalization. It may be that improvements to the model training would further enhance the benefit of FA.

On the other hand, we should also distinguish the effect of simply training on reconstructed images from the effect of FA. In future, we believe that this method can be extended to achieve better domain generalization once we can manipulate the factors realistically to generate out-of-distribution samples.

4   Conclusion

In this paper, two data augmentation methods were proposed to address the domain adaptation and generalization problems in the field of CMR image segmentation. In particular, a geometry-related augmentation method was introduced, which aims to remove the scale and resolution bias of the original data. Further, the second proposed augmentation method aims bridge the gap between populations and data captured by different scanners. To achieve this, the original data is projected onto a disentangled latent space and generates new samples by combining disentangled factors from different domains. The presented experimental results showcase the contribution of the geometry-based method to CMR image segmentation through boosting the domain generalization, while also demonstrating the contribution of the disentangled factors mixing method to the domain adaptation.

5   Acknowledgement

References

1. Baumgartner, C.F., Koch, L.M., Pollefeys, M., Konukoglu, E.: An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation. In: Proc. STACOM. pp. 111–119 (2017)
2. Chaitanya, K., Karani, N., Baumgartner, C., Donati, O., et. al: Semi-supervised and task-driven data augmentation. In: Proc. IPMI. pp. 29–41 (2019)
3. Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D.E., Dharmakumar, R., Tsaftaris, S.A.: Disentangled representation learning in cardiac image analysis. Medical Image Analysis **58**(101535) (2019)
4. Chen, C., Qin, C., Qiu, H., Ouyang, C., Wang, S., Chen, L., Tarroni, G., Bai, W., Rueckert, D.: Realistic adversarial data augmentation for MR image segmentation. In: Proc. MICCAI (2020)
5. Chen, C., Qin, C., Qiu, H., Tarroni, G., Duan, J., Bai, W., Rueckert, D.: Deep learning for cardiac image segmentation: A review. Frontiers in Cardiovascular Medicine **7**(25), 1–33 (2020)
6. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proc. ICML. pp. 1180–1189 (2015)
7. Ghifary, M., Bastiaan Kleijn, W., et. al: Domain generalization for object recognition with multi-task autoencoders. In: Proc. ICCV. pp. 2551–2559 (2015)
8. Hsu, W., Zhang, Y., Weiss, R.J., Chung, Y., Wang, Y., Wu, Y., Glass, J.: Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In: Proc. ICASSP. pp. 5901–5905 (2019)
9. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proc. ICCV. pp. 1501–1510 (2017)
10. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. Proc. ICML p. 448–456 (2015)
11. Jiang, J., Hu, Y.C., Tyagi, N., Zhang, P., Rimner, A., et. al: Tumor-aware, adversarial domain adaptation from ct to mri for lung cancer segmentation. In: Proc. MICCAI. pp. 777–785. Springer (2018)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR (2015)
13. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. ICCV. pp. 2980–2988 (2017)
14. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., et. al: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
15. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: Proc. ICML. p. 807–814 (2010)
16. Paszke, A., Gross, S., Massa, F., Lerer, A., et. al: PyTorch: An imperative style, high-performance deep learning library. In: Proc. NeurIPS. pp. 8026–8037 (2019)
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Proc. MICCAI. pp. 234–241 (2015)
18. Sudre, C.H., Li, W., et. al: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Proc. ML-CDS. pp. 240–248 (2017)
19. Tao, Q., Yan, W., Wang, Y., et. al: Deep learning-based method for fully automatic quantification of left ventricle function from cine mr images: a multivendor, multicenter study. Radiology **290**(1), 81–88 (2019)
20. Yang, J., Dvornek, N.C., Zhang, F., Zhuang, J., Chapiro, J., et. al: Domain-agnostic learning with anatomy-consistent embedding for cross-modality liver segmentation. In: Proc. ICCV Workshops. pp. 0–0 (2019)
21. Zhao, A., et. al: Data augmentation using learned transformations for one-shot medical image segmentation. In: Proc. CVPR. pp. 8543–8553 (2019)