Style Data Augmentation for Robust Segmentation of Multi-modality Cardiac MRI

Buntheng Ly[1(✉)], Hubert Cochet[2], and Maxime Sermesant[1]

[1] Inria, Université Côte d'Azur, Sophia Antipolis, France
buntheng.ly@inria.fr
[2] IHU Liryc, University of Bordeaux, Pessac, France

Abstract. We propose a data augmentation method to improve the segmentation accuracy of the convolutional neural network on multi-modality cardiac magnetic resonance (CMR) dataset. The strategy aims to reduce over-fitting of the network toward any specific intensity or contrast of the training images by introducing diversity in these two aspects. The style data augmentation (SDA) strategy increases the size of the training dataset by using multiple image processing functions including adaptive histogram equalisation, Laplacian transformation, Sobel edge detection, intensity inversion and histogram matching. For the segmentation task, we developed the thresholded connection layer network (TCL-Net), a minimalist rendition of the U-Net architecture, which is designed to reduce convergence and computation time. We integrate the dual U-Net strategy to increase the resolution of the 3D segmentation target. Utilising these approaches on a multi-modality dataset, with SSFP and T2 weighted images as training and LGE as validation, we achieve 90% and 96% validation Dice coefficient for endocardium and epicardium segmentations. This result can be interpreted as a proof of concept for a generalised segmentation network that is robust to the quality or modality of the input images. When testing with our mono-centric LGE image dataset, the SDA method also improves the performance of the epicardium segmentation, with an increase from 87% to 90% for the single network segmentation.

Keywords: Image segmentation · Multi-modality · Cardiac magnetic resonance imaging · Late Gadolinium enhanced · Deep learning

1 Introduction

The combination of different MRI sequences, signal weighting techniques and contrast agents that are currently used for MRI gives rise to diverse modalities and qualities of the output image. Although each technique yields exploitable results, the variation in the image contrast can be detrimental for the development of automatic analysis tools in medical imaging.

M. Pop et al. (Eds.): STACOM 2019, LNCS 12009, pp. 197–208, 2020.
https://doi.org/10.1007/978-3-030-39074-7_21

To answer to the input diversity problem in machine learning segmentation, Seeböck et al. used an unpaired modality transfer generator network to reduce the variability between multi-centric datasets [12]. On the other hand, Isensee et al. proposed the nnU-Net (no-new U-Net), which automatically generates a CNN pipeline that is optimised for each specific dataset [5]. However, these methods require a sufficient mono-modality dataset, as they were built to be used for mono-modality segmentation.

In this study, we propose an alternative approach to this problem. We design a data augmentation method to train a single Deep Learning model to be robust to multi-modality input, including the modality that was not used for optimisation, thus the trained model can be used as a generalised segmentation tool. The style data augmentation introduces diversity of image contrast into the training dataset, with the goal to prevent the model from over-fitting toward the training image modality and to focus the network attention to the fundamental geometry features of the target. We base this method on the idea that despite having different contrasts, the organ geometry features are consistent between MRI modalities.

In this study, we use a 3D convolutional neural network for the segmentation [1]. Nonetheless, this method can be costly in term of memory usage and computation time. We have devised two strategies to combat these issues. Firstly, we proposed a minimalist U-Net inspired network, tailored to accelerate the convergence speed and to decrease memory usage. Secondly, we adopt the dual network strategy [6], which allows for the segmentation of high resolution targets.

2 Method

2.1 Thresholded Connection Layer Network

We propose a segmentation convolutional neural network called thresholded connection layer or TCL-Net. The network architecture is shown in Fig. 1. This architecture is an iteration of the U-Net architecture, originally proposed by Ronneberger et al. [11]. As such, the network follows the same U-shape design and is made up of an encoder and a decoder.

The architecture of TCL-Net exploits the segmentation network ultimate objective, which is to eliminate non-target pixels and to highlight the target pixels of the input image. TCL-Net uses, as the building unit, two consecutive, padded, $3 \times 3 \times 3$ convolutional layers, each followed by a normalisation and a non-linear activation layer. At the end of each encoder unit, a $2 \times 2 \times 2$ max pooling is applied. Correspondingly, a $2 \times 2 \times 2$ upsampling is applied to the output of the decoder unit.

For the normalisation layer, we use the instance normalisation function [13], since the training is done with a single-input batch. We used the LeakyReLU [9] as the activation function of both convolutional layers of the encoder unit. For the decoder units, the LeakyReLU layer is used after the first convolutional and the ThresholdedReLU [7] is used after the second convolutional layer, before the

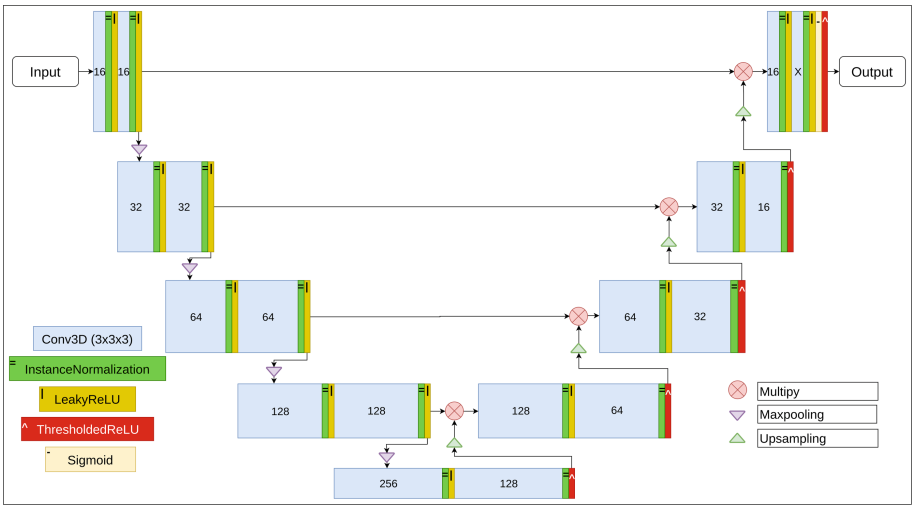


Fig. 1. Thresholded connection layer network. Noted that the ThresholdedReLU layers (red boxes) are only used at the end of each decoder unit. *The number indicates the filter of each convolutional layer. X is set according to the segmentation label (1 for single label segmentation).* (Color figure online)

upsampling layer. We used 0.3 as the coefficient of the LeakyReLU layer and 0.5 as the threshold value of the ThresholdedReLU layer.

The LeakyReLU layer would allow the negative pixels of the feature matrices to pass through, while the ThresholdedReLU would reduce pixels smaller than the threshold to zero. The goal is to let the features to be liberally processed through each level of the encoder and to only apply thresholding at the very end of each resolution level. In order for the network to preserve the elimination progress from the thresholding, we used multiplication to connect the output of the encoder with the decoder, instead of concatenation. Additionally, the multiplication operation could greatly amplify or reduce the value of the output features, which influences the elimination likelihood of each pixel in the next thresholded layer.

Toward the end of TCL-Net, the sigmoid activation layer is added to scale each pixel's value down to between 0 and 1. The output of the sigmoid function can be used to gauge the certainty of the segmentation at the pixel level. The final ThresholdedReLU layer is used as the final processing function to eliminate the pixels less than 0.5 from the output segmentation. The last two activation layers would facilitate the integration of the Dice loss function detailed in Subsect. 2.4.

2.2 Dual U-Net Strategy

In this study, we implement the dual U-Net strategy [6], where two networks are trained independently but can be used consecutively in the segmentation

pipeline. The first network is trained to segment the target from the low resolution inputs, as the original image has to be shrunk down to reduce memory consumption. The segmented results of the first network are used to crop the original images, which will be used as input for the second network.

To crop the output of the first network, we round up all the nonzero pixel values to 1, then only one biggest region of connected positive pixels is kept. Note that we integrate this strategy with TCL-Net, where thresholding is applied at the end of the network. Additional thresholding might be necessary with other architecture. To take into account the segmentation error, we apply the binary dilation transformation on the cluster using a $5\times5\times5$ spherical structure element. Finally, the original image is cropped using the bonding box of the dilated region.

For this study, we are interested in left ventricular segmentation from CMR images, specifically from late Gadolinium enhanced (LGE) images, in which the myocardial scar is visible. The first network is used to locate the epicardium, and then second network can either be used to refine the segmentation of the same target or smaller targets such as endocardium and myocardial scar.

2.3 Style Data Augmentation

The style data augmentation strategy focuses on introducing contrast diversity in the training dataset, via different image processing algorithms. The aim is to prevent the model from over-fitting to any specific contrast and to focus the optimisation toward the fundamental geometry features of target.

The image transformation algorithms were selected arbitrarily, as the goal is to simply increase the variety of the training images. For this study, we selected 5 transformation functions, including adaptive histogram equalisation [3], Laplacian transformation, Sobel edge detection, intensity inversion and histogram matching [10], as shown in Fig. 2.

The histogram matching method can be used to convert the histogram of the original training images (C0 and T2) toward the histogram of the validation images (LGE) without the ground truth mask. More details on the dataset used for this study is described in Subsect. 3.1. These functions were applied to the normalized original image using the functions provided by SimpleITK's python package [8,14].

Our goal is to pre-train a segmentation model that is robust to any unknown-to-the-model modality. As the method focuses on the geometry features, it is only suited to be used for the image modalities where the target shape is consistent.

2.4 Experimental Setting

To validate the effectiveness of TCL-Net, we compare the validation result of the new architecture with a baseline network proposed by Isensee et al. [4]. Both networks were trained using a single 3D input per batch. The 3D images are interpolated to equalise the spacing of each dimension, thus the extracted data would closely correspond to the physical size. We use linear and nearest neighbour interpolation methods on the greyscale and mask images, respectively. The interpolated images are then resized to $128 \times 128 \times 128$. Finally, the

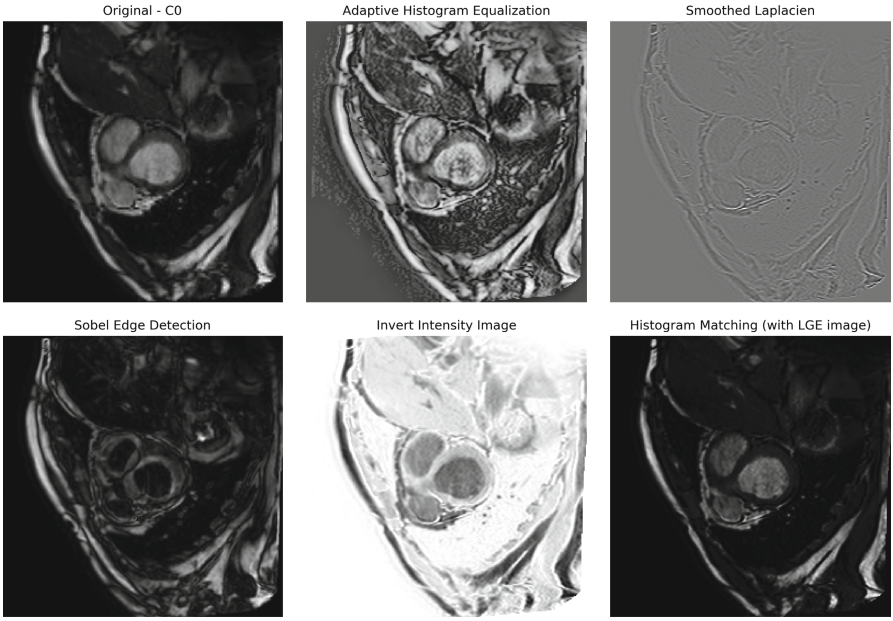


Fig. 2. Different variations of input training images and the image processing methods. *C0 denotes the steady-state free precession CMR modality image.*

images are normalised using linear normalisation function to bring the greyscale value between [0–255]. To test the validity of our method, we do not apply any shape transformation for additional augmentation or any complex pre-processing method on the validation or training images.

We use an initial learning rate of $1e-4$, which decays by half each 5 epochs with no validation improvement. An early stop is also programmed after 20 epochs of no increase in validation performance. At each epoch, 100 images will be chosen randomly from the training dataset to be used to train the network. The network is updated using Adam optimisation and Dice loss, calculated using Eq. 1, where $\hat{Y}$ is the prediction mask, and Y is the manual labelled mask. During the training, we also measure the original Dice coefficient [2] between the prediction and the manual mask, by applying the "half to even" round function to binarise the output segmentation, Eq. 2. The round function breaks the gradient chain, which prevents Dice coefficient from being used for backpropagation.

$$Dice_{Loss} = 1 - 2 * \frac{\sum(\hat{Y} * Y)}{\sum \hat{Y} + \sum Y} \tag{1}$$

$$Dice_{Coeff} = 2 * \frac{\sum(round(\hat{Y}) * Y)}{\sum round(\hat{Y}) + \sum Y} \tag{2}$$

3 Evaluation on Clinical Data

3.1 Materials

For the multi-modal dataset, we use the dataset provided in MS-CMRSeg 2019 segmentation challenge [15,16]. The challenge dataset consists of 135 3D CMR images of 45 patients taken under three modalities: T2-Weighted (T2), balanced-Steady State Free Precession (C0) and late Gadolinium enhanced (LGE), Fig. 3. The manually labelled masks were provided for the first 35 images of T2 and C0 but only for the first 5 images of LGE modality. The provided labels include epicardium and endocardium of the left ventricle and endocardium of the right ventricle. By applying our data augmentation method, we trained the network with 420 images of different variations of the original C0 and T2 images and validated the network with the 5 LGE images.

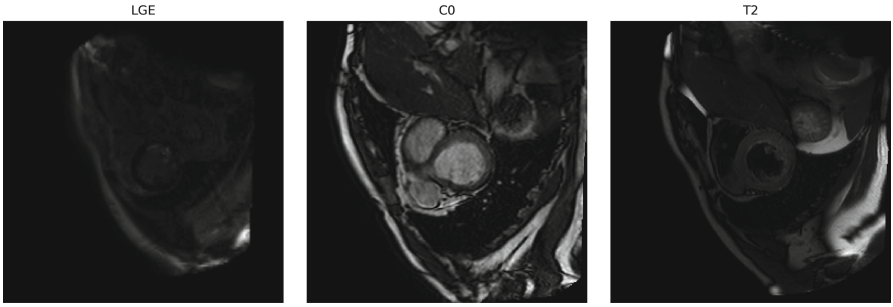


Fig. 3. Short-axis view of original LGE, C0 and T2 CMR images. Note that despite having different contrasts, all these images show the same anatomical structure.

While the SDA method was not designed to be used with mono-modal datasets, we still wish to study the effectiveness of the training input's contrast diversity under this context. We used our local dataset, which consists of 119 mono-centric LGE-CMR images provided by IHU Lyric. Since the dataset is mono-modal, we remove histogram matching from the SDA algorithms. The original dataset was first split in 9:1 ratio for training and validation, before the augmentation method was applied to the training images. We then compare the mean of best validation scores from 5 different sets of validation images of the model trained with and without SDA method.

3.2 Results

Multi-modality. The Table 1 shows the validation results of TCL-Net and Isensee on the multi-modal dataset. Using the TCL-Net with dual network and SDA method, the validation scores reach 0.967 and 0.904 for the epicardium and endocardium segmentations. This is a considerable improvement from 0.833 and

Table 1. The validation Dice coefficient on multi-modality dataset. **: the best results; w/o HM: without Histogram Matching.*

	Single Network		Dual Network	
	TCL-Net	Isensee's	TCL-Net	Isensee's
Epi				
Original(C0&T2)	0.833 ± 0.020	0.791 ± 0.015	0.878 ± 0.006	0.787 ± 0.007
+SDA	*0.915 ± 0.006	0.845 ± 0.013	*0.9677 ± 0.012	0.866 ± 0.011
+SDA(w/o HM)	0.908 ± 0.007	0.854 ± 0.129	0.9671 ± 0.004	0.874 ± 0.007
Endo				
Original(C0&T2)	0.692 ± 0.041	0.651 ± 0.008	0.767 ± 0.008	0.787 ± 0.005
+SDA	*0.865 ± 0.021	0.805 ± 0.038	*0.904 ± 0.003	0.836 ± 0.006
+SDA(w/o HM)	0.857 ± 0.011	0.780 ± 0.019	0.900 ± 0.007	0.839 ± 0.010

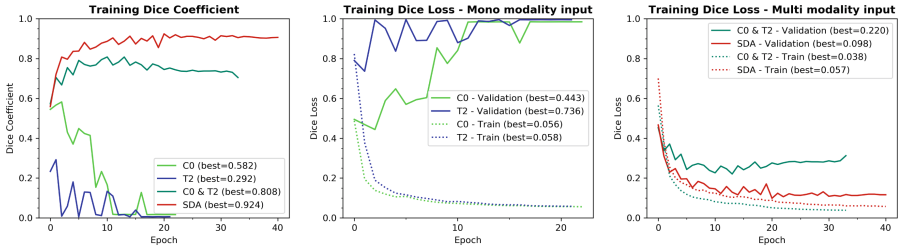
0.692 without these two improvements. There is a slight decrease in performance when histogram matching is removed from the SDA algorithms. Nonetheless, the network still performs better compared to training with only original images.

As shown in Fig. 4, TCL-Net performance is enhanced considerably with multi-modality training set (SDA and C0&T2) compared with mono-modality. We can observe in Fig. 4b that the model would quickly overfit to the training modality, as the gap between training and validation scores get higher each epoch. On the contrary, the over-fitting becomes less severe when there is diversity in the training input, as shown in Fig. 4c. The validation also appears more stable at the end of the training with SDA compared with the training with only original C0&T2.

The Fig. 5 shows the validation output of epicardium and endocardium segmentations using the dual TCL-Net models trained with SDA method. Both models perform well and produce accurate segmentation in the region where there is no myocardial scar. Yet, the models struggle at the scar regions, as pointed by the arrows in the Fig. 5b and c.

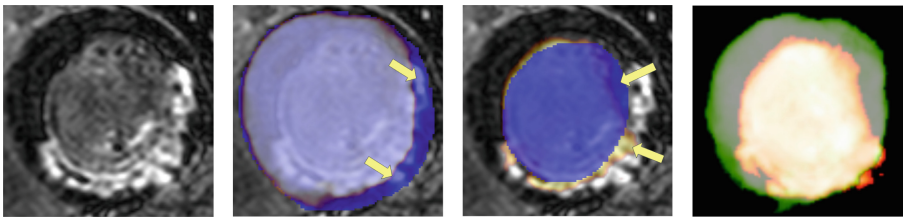
Mono-modality. When testing on the mono-centric and mono-modal dataset the SDA method does show improvement in validation Dice coefficients from 0.874 to 0.905 for the epicardium segmentation in the first TCL-Net, Table 2. However, the method has an adverse effect on the myocardial scar segmentation in the second TCL-Net. Figure 6 shows the validation segmentation output of the myocardial scar using the TCL-Net models trained with and without SDA method. Despite the poor Dice scores, both models can adequately detect the scar regions, Fig. 6c and b.

TCL-Net. As shown in Tables 1 and 2, TCL-Net achieves better final validation score than the baseline model, both in multi- and mono-modal datasets with or without SDA. Figure 7 shows the validation Dice coefficient of both networks during the first network training. Figure 7a shows that TCL-Net required less



(a) (b) (c)

Fig. 4. Validation results of epicardium segmentation of the first TCL-Net on multi-modal dataset.



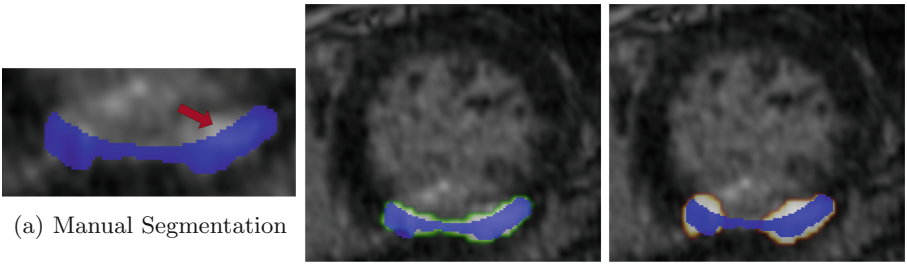
(a) Original image (b) Epicardium (c) Endocardium (d) Epi&Endo

Fig. 5. Epicardium and endocardium segmentation from LGE image using models trained with original and augmented C0 & T2 images (multi-modality dataset, dual TCL-Net with SDA). *Blue: Ground Truth; Orange, Green: Predicted Segmentations.* (Color figure online)

Table 2. The validation Dice coefficient on mono-modality dataset. **: best results.*

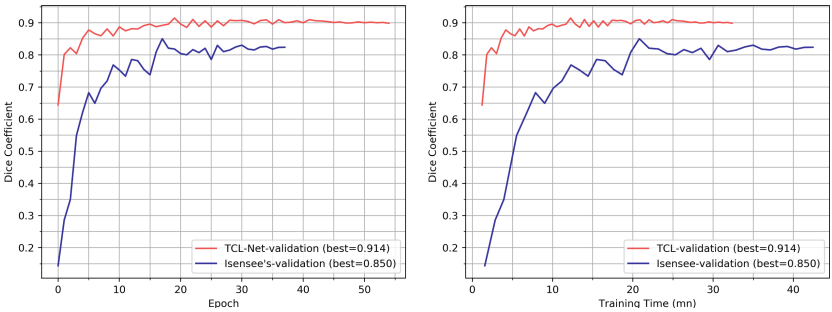
	Single Network	**Dual Network**
	Epicardium	Scar
TCL-Net		
Original(LGE)	0.874 ± 0.002	$*0.462 \pm 0.070$
+SDA	$*0.905 \pm 0.011$	0.444 ± 0.061
Isensee's		
Original(LGE)	0.853 ± 0.012	0.439 ± 0.042
+SDA	0.851 ± 0.012	0.348 ± 0.040

epochs for the optimisation. When factoring the training time in Fig. 7b, TCL-Net has faster training speed than the baseline network, with the validation Dice coefficient reaching 85% in less than 5 min.



(a) Manual Segmentation (b) Without SDA (c) With SDA

Fig. 6. Myocardial scar segmentation using dual U-Net strategy. *Blue: manual segmentation* (Color figure online)



(a) Corrected by training epoch. (b) Corrected by training time.

Fig. 7. Training and validation results of TCL-Net vs. Isensee's (multi-modality dataset, first network with SDA).

Dual Network Segmentation. The results from Tables 1 and 2 show that the dual network strategy increases significantly the segmentation accuracy. On top of that, compared with the single network, the dual network also produces higher resolution segmentation output, Fig. 8.

4 Discussion

SDA-Epicardium and Endocardium. The image processing functions implemented in SDA create images of different contrasts with defined border and geometric features, thus making the method applicable to the target regular structure such as the epicardium and endocardium. The results in Sect. 3.2 show the increase in performance in both mono- and multi-modal datasets for the segmentation of the epicardium. As shown in multi-modal experiments, the SDA improves segmentation validation score of the LGE images, without any optimisation with the actual LGE data.

The slight decrease in performance when histogram matching was not included in SDA further proves that the strategy does not overly depend on this

(a) Single Network (b) Dual Network

Fig. 8. Segmentation output of single vs. dual network.

particular transformation. It also validates that the increase in contrast diversity in augmentation algorithms leads to the increase in performance, rather than over-fitting. Nevertheless, the trained model still reaches a limit, as observed in dual network segmentation in Table 1.

SDA-Myocardial Scar. Because the model can no longer depend on image contrast for the segmentation when training with SDA method, it has to rely on the patterns of the target, such as the traces of the myocardium wall and the homogeneity of the intensity of each structure. Therefore, the method might not be suitable for the targets without uniform structure, such as the myocardial scar. For instance, when training on C0 and T2 images of MS-CMRSeg challenge, the model is only familiar with homogeneous myocardium. Thus, it does not perform well when scar is present on the myocardium of LGE images, Fig. 5.

The inconsistency in contrast of the myocardial scar may explain why the network achieves better result without SDA for the scar segmentation on the mono-modal dataset. As shown in Fig. 6a by the red arrow, the scar region does not include the entire area of the same intensity, since the upper area belongs to the cavity of the ventricle. Because the scar does not have specific geometric shape like the epicardium or endocardium, the model trained with only the original image would perform better, since it can depend more on the specific contrast of the LGE modality during optimisation than the model trained with SDA.

TCL vs. Isensee. In their original paper [4], Isensee et al. integrate a more elaborate preprocessing technique on the input image than what is done in this experiment. Therefore, our experiment might not present the optimal condition for the baseline network. Our goal is to simply compare the performance of the larger network with our new architecture on the minimal processing datasets. The TCL-Net architecture used in the experiment is considerably smaller with only 3,529,635 trainable parameters, than Isensee's model, which has 8,294,659.

The experiment shows that compared to Isensee's, our architecture achieves faster convergence and better validation performance.

5 Conclusion

We proposed a data augmentation strategy that increases the accuracy of the segmentation and is invariant to the modality of the validation image. The SDA strategy forces the network to be independent from the input image modality and prevents it from over-fitting to any specific contrast. This validates our theory that the diversity in training input increases the neural network performance.

The image transformation algorithms in SDA can also be seen as placeholders and be easily replaced by the real world MR modalities. Our current experiment uses the validation result of LGE images to terminate the training, thus making the trained coefficients bias toward the LGE modality. A more diverse real-world multi-modality dataset is needed to improve the universality of the trained network.

The efficiency of SDA method also challenges the traditional concept of complex normalisation or equalisation of the dataset in medical image segmentation. It pushes the boundary of the convolutional neural network in term of its flexibility and adaptability toward the input quality in semantic segmentation task.

Acknowledgement. This research is a collaboration between Inria Sophia Antipolis - Méditerrané and IHU Lyric. This work is possible due to the datasets provided by MICCAI's MS-CMRSeg 2019 challenge and IHU Lyric and the NEF computational cluster provided by Inria. The author would like to thank the work of relevant engineers and scholars.

References

1. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
2. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
3. Hummel, R.: Image enhancement by histogram transformation. Comput. Graph. Image Process. **6**(2), 184–195 (2008)
4. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 287–297. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_25
5. Isensee, F., Petersen, J., Kohl, S.A.A., Jäger, P.F., Maier-Hein, K.H.: nnU-Net: breaking the spell on successful medical image segmentation 1, 1–8 (2019)

6. Jia, S., et al.: Automatically segmenting the left atrium from cardiac images using successive 3D U-nets and a contour loss. In: Pop, M., et al. (eds.) STACOM 2018. LNCS, vol. 11395, pp. 221–229. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12029-0_24
7. Konda, K., Memisevic, R., Krueger, D.: Zero-bias autoencoders and the benefits of co-adapting features. ICLR **2015**, 1–11 (2014)
8. Lowekamp, B.C., Chen, D.T., Ibáñez, L., Blezek, D.: The design of SimpleITK. Front. Neuroinformatics **7**, 45 (2013)
9. Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. ICML **28**, 6 (2013)
10. Nyúl, L.G., Udupa, J.K., Zhang, X.: New variants of a method of MRI scale standardization. IEEE Trans. Med. Imaging **19**(2), 143–150 (2000)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. Seeböck, P., et al.: Using CycleGANs for effectively reducing image variability across OCT devices and improving retinal fluid segmentation (2019)
13. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance Normalization: The Missing Ingredient for Fast Stylization (2016)
14. Yaniv, Z., Lowekamp, B.C., Johnson, H.J., Beare, R.: Simpleitk image-analysis notebooks: a collaborative environment for education and reproducible research. J. Digit. Imaging **31**(3), 290–303 (2018)
15. Zhuang, X.: Multivariate mixture model for cardiac segmentation from multi-sequence MRI. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 581–588. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_67
16. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 2933–2946 (2018)