Multimodal Cardiac Segmentation Using Disentangled Representation Learning

Agisilaos Chartsias[1(✉)], Giorgos Papanastasiou[2,3], Chengjia Wang[2,3], Colin Stirrat[2,3], Scott Semple[2,3], David Newby[2,3], Rohan Dharmakumar[5], and Sotirios A. Tsaftaris[1,4]

[1] School of Engineering, University of Edinburgh, Edinburgh EH9 3FB, UK
agis.chartsias@ed.ac.uk
[2] Edinburgh Imaging Facility QMRI, Edinburgh EH16 4TJ, UK
[3] Centre for Cardiovascular Science, Edinburgh EH16 4TJ, UK
[4] The Alan Turing Institute, London, UK
[5] Cedars Sinai Medical Center, Los Angeles, CA, USA

Abstract. Magnetic Resonance (MR) protocols use several sequences to evaluate pathology and organ status. Yet, despite recent advances, the analysis of each sequence's images (modality hereafter) is treated in isolation. We propose a method suitable for multimodal and multi-input learning and analysis, that disentangles anatomical and imaging factors, and *combines* anatomical content across the modalities to extract more accurate segmentation masks. Mis-registrations between the inputs are handled with a Spatial Transformer Network, which non-linearly aligns the (now intensity-invariant) anatomical factors. We demonstrate applications in Late Gadolinium Enhanced (LGE) and cine MRI segmentation. We show that multi-input outperforms single-input models, and that we can train a (semi-supervised) model with few (or no) annotations for one of the modalities. Code is available at https://github.com/agis85/multimodal_segmentation.

Keywords: Multimodal segmentation · Disentanglement · Representation learning · Cardiac MR

1 Introduction

MR is non-invasive and offers high soft-tissue contrast suitable for numerous applications. Multiple sequences are used in a single MR session, producing images of different contrast (modalities), that are characterised by disparities in overall image quality and signal-to-noise ratio, but also provide complementary information of anatomy and function. Developing methods to automatically segment tissue from such multimodal data remains important: for example in cardiac MR, cine and LGE needs to be jointly assessed to characterise myocardial infarction [11], since cine contains high anatomical information, whereas LGE focuses on nulling myocardial signal to detect hyper-intense infarct zones.

To this date, processing of such multimodal data treats each modality in isolation. Yet, jointly considering different modalities should be beneficial to obtain information from another modality that better captures anatomy (see Fig. 1 for a motivating example). Herein, we offer a step change: we propose a model designed to overcome challenges presented by multimodal analysis in cardiac MR solving the core problems of representation learning, cross-modal registration, information fusion and segmentation all in a joint end-to-end fashion in a semi-supervised setting, without requiring exhaustive annotations.

Deep learning has been successfully used for automating segmentation, however, most methods in the heart focus on single modalities. This is mainly because of the high variability observed in signal intensity patterns across different MR modality data and organ characteristics. While, in the brain, multimodal images are commonly used together [6], in the heart, multi-input processing and multimodal learning are substantially *challenging* due to inherent spatiotemporal and signal intensity differences (between modalities). These compromise learning direct pixel-to-pixel correspondences.

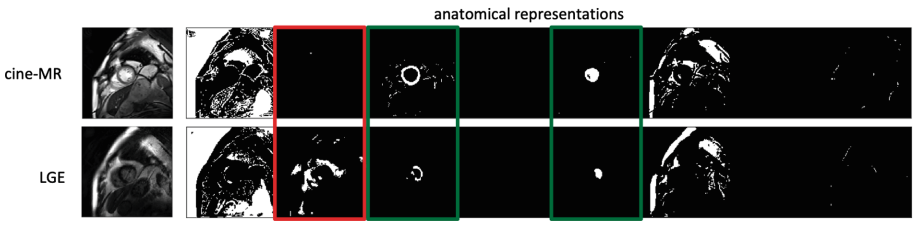

Fig. 1. Cine-MR and LGE images with corresponding anatomical factors. Common and unique information is marked with green and red boxes. Low tissue contrast (myocardial nulling) in LGE leads to poor separation in distinct channels between myocardium and surrounding tissues (e.g. ventricle). This can be corrected using the cine anatomy. (Color figure online)

We address the above difficulties, for the first time, with disentangled representations, i.e. mappings from multimodal images to corresponding anatomical and imaging factors. Anatomical factors contain structure (multi-channel binary maps); imaging factors contain input signal intensity characteristics. A Spatial Transformer Network (STN) [9] co-registers the corresponding (intensity-invariant) anatomical factors, avoiding the co-registration in image space (difficult in cardiac and other soft-tissue organs). We then combine (fuse) the aligned anatomical factors to find complementary features useful for segmentation.

Contributions: **(1)** Multimodal learning based on disentangled representations, that combines information present in different modalities without the explicit requirement for registered image pairs. **(2)** An application in cardiac segmentation, in which we improve on the segmentation accuracy of single-input (unimodal) models. **(3)** Semi-supervised learning: when few (or no) labels are

available, we transfer information from the other modality and use reconstruction costs.

2 Related Work

Disentangled Representations: Decomposing the feature space into spatial and style-like factors has shown success in computer vision [7,13], and recently in semi-supervised cardiac segmentation [2] and multimodal registration [17]. In medical imaging, disentangled representations have more stringent requirements, since the anatomical factors must have semantic and quantifiable meaning (e.g. be useful for segmentation). Our proposed method thus differs significantly from related multimodal methods; it strives for anatomical factors to be semantic and geometrically consistent across modalities, as well as maintain the image dimensions to allow a direct mapping to segmentations. These properties are essential for anatomical registration and fusion, as well as semi-supervised learning.

Multiple Inputs in Cardiac: Level sets have been applied for cine-MR and LGE segmentation given shape constraints, generated by convolutional networks [14]. In [8], unannotated data were translated into a modality with annotations using "style transfer". However, this relies on learning good pixel-wise transformations, which is not always possible [23]. Also the lack of an explicit fusion mechanism may be problematic when images exhibit low contrast-to-noise between different organs. Non-deep learning approaches include multimodal atlases [25], whereas simultaneous segmentation and registration of multimodal cardiac MR images has been proposed with Multivariate Mixture Models [24].

Multimodal Learning: In medical imaging, e.g. brain MRI, most multimodal approaches assume perfect alignment between the inputs. Many methods have been proposed for synthesis [10], and segmentation, for example with concatenated multi-channel inputs [5,6]. To aid the learning process, in [20] they use cross-modal convolutions and convolutional LSTMs, whereas in [4] they propose densely connected streams (one per modality) to fuse high and low level features.

One approach to handle unregistered multimodal data is to treat them separately and share parts of single-input models. An empirical study of different sharing options [22] concluded that a common feature space connected with individual encoders and decoders has the best performance. Small mis-registrations have been previously handled with an affine STN [10]. Our method is able to fuse multimodal information, and differently from [10], uses a non-linear STN.

3 Proposed Approach

Multimodal Spatial Decomposition Network (MMSDNet) consists of multiple components (see Fig. 2), described in Sects. 3.1 and 3.2. At inference time, MMSDNet can take as input a 2D image (of either modality) or two images (of different modalities) simultaneously. One encoder per modality extracts anatomical factors, which are used for segmentation or input reconstruction. If multimodal

image pairs are available, anatomical factors are aligned by a STN, and combined to produce a fused anatomy, which is used for the final segmentation mask.

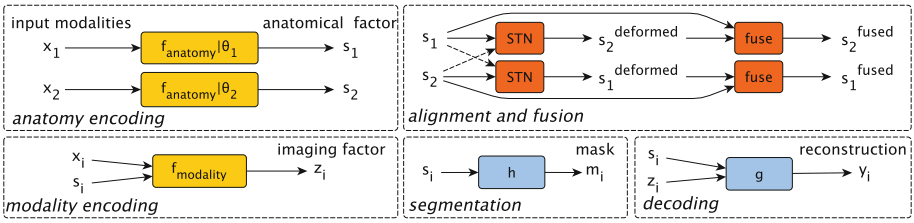

Fig. 2. MMSDNet components. Top left: anatomy encoders (one per modality) extract anatomical factors from images. Top right: misalignments are corrected with a STN; aligned factors are then fused to produce one factor. Bottom left: imaging factors are extracted by a modality encoder. Bottom right: the anatomical factor produces a segmentation; anatomical and imaging factors together reconstruct an image.

3.1 Model

Encoding: Assuming two input modalities, and image samples x_i (of height H and width W), where $i \in \{1, 2\}$, the anatomy factor is derived from an encoder $f_{anatomy}$ with parameters θ_i: $s_i = f_{anatomy}(x_i|\theta_i)$. Anatomy encoders are fully convolutional networks (architecture is shown in Fig. 3), which output $s_i \in \{0, 1\}^{H \times W \times 8}$, a one-hot encoding (in the channel dimension), 8-channel binary feature map of the same spatial dimension as the input (each channel represents a different anatomical area). These two restrictions encourage a semantic representation, since each tissue will be present only in one channel. They also disentangle anatomy from imaging, since a binary image does not encode any modality information in gray levels.

Alignment: The two anatomical factors are aligned using a Spatial Transformer Network (STN) [9] (architecture is shown in Fig. 3), which, through non-rigid registration, generates two deformed anatomies $s_1^{deformed} = stn(s_1, s_2)$ and $s_2^{deformed} = stn(s_2, s_1)$. The STN learns a matrix of 5×5 control points that define the displacement field, which registers the second to the first anatomical factor. Thin plate spline [1] is applied to interpolate the surface that passes through each control point.

Fusion: The deformed anatomy $s_1^{deformed}$ is an approximation of the anatomy s_2 corresponding to image x_2. Thus, it can be fused with s_2 to produce a single representation of x_2 that preserves the encoded multimodal anatomical features. We require the union of the aligned features, and thus use the pixel-wise max: $s_1^{fused} = \max(s_1^{deformed}, s_2)$. Accordingly, s_2^{fused} is also generated.

Segmentation: The previous steps produce six anatomical factors, namely s_1, s_2, $s_1^{deformed}$, $s_2^{deformed}$, s_1^{fused} and s_2^{fused}, which are used as input (one at a time) to a convolutional network $h(.)$ (architecture is shown in Fig. 3) to obtain the final segmentation masks. Depending on the inference task, we can get a

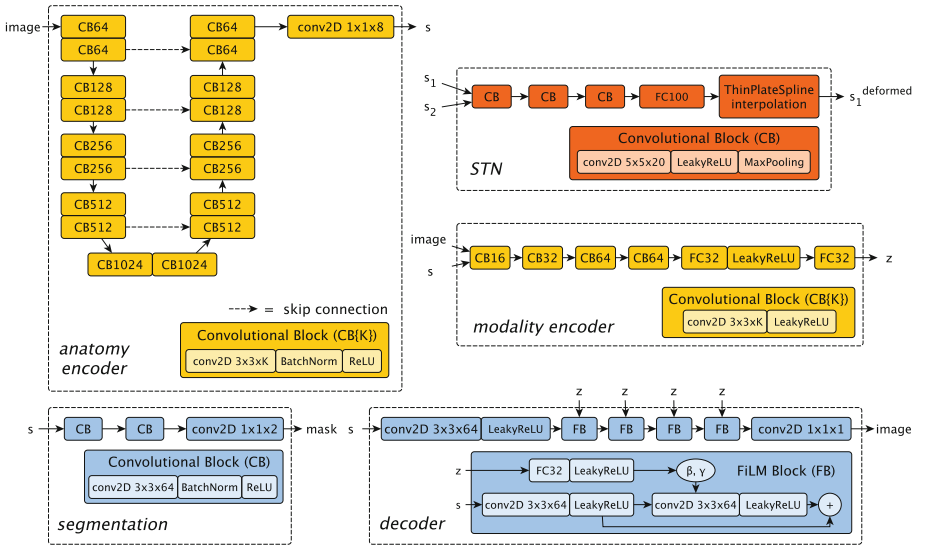

Fig. 3. Architectures of the MMSDNet components. Top left: the anatomy encoder follows a U-Net [18] architecture and maps an image to an anatomical factor s. Downsampling and upsampling are performed with max pooling and nearest neighbour interpolation respectively. Bottom left: the segmentation network is a small fully convolutional network that given s, produces a segmentation mask. Top right: the spatial transformer network consists of three convolutional and one fully connected layers and predicts the interpolation parameters used to register s_1 to s_2. Middle right: the modality encoder is a convolutional network that predicts the modality factor z. Bottom right: the decoder is a convolutional network that modulates an anatomy factor s with a modality factor z to generate an image.

segmentation using the appropriate anatomy, as also demonstrated in Sect. 4. If only x_i is available, the segmentation is obtained from s_i, whereas if both x_1, x_2 are available the fused anatomy s_i^{fused} produces the most accurate result.

3.2 Additional Networks and Losses

Our end-to-end strategy enables the model to learn from multimodal data to separate anatomy from imaging characteristics, whilst doing good segmentation, registration and reconstruction. Critically, reconstruction enables semi-supervised learning, aided via adversarial objectives on segmentation masks. Below we detail the breakdown of the overall training loss, $L = \lambda_1 L_{KL} + \lambda_2 L_{seg} + \lambda_3 L_{adv} + \lambda_4 L_{rec} + \lambda_5 L_{z_{rec}}$. (The λ's are set to 0.1, 10, 1, 10, 1 respectively.)

$\mathbf{L_{KL}}$ and $\mathbf{L_{z_{rec}}}$: Given an image x_i, from modality i, then a corresponding anatomy s can either be the encoded $s_i = f_{anatomy}(x_i|\theta_i)$, or the deformed $s_j^{deformed}$ and fused anatomies s_j^{fused} if x_i has a paired slice x_j in modality j. Key is the disentanglement of the latent space into anatomical s_i and imaging factors z_i (8-dimensional vector), which requires a **modality encoder**,

$z_i = f_{modality}(x_i, s_i)$, and a **decoder**. The decoder reconstructs the input, $\hat{x_i} = g(s, z_i)$, using FiLM [16], by modulating s with scaling and offset parameters β and γ, that are learned from z_i. The network architectures of both the modality encoder and the decoder are shown in Fig. 3. The posterior distribution given the inputs $q(z|x, s)$ is modelled after the Variational Autoencoder [12] to follow a Gaussian prior $p(z) = \mathcal{N}(0, 1)$, by minimising the KL-divergence between q and p: $L_{KL} = D_{KL}(q(z|x, s)\|p(z))$. The representation disentanglement is further encouraged by a $z-$reconstruction cost using the ℓ_1 loss: $L_{z_{rec}} = \|z - f_{modality}(\hat{x_i}, f_{anatomy}(\hat{x_i}|\theta_i))\|_1$, where $\hat{x_i}$ is produced by a z that is sampled from the Gaussian prior.

$\mathbf{L_{rec}}$: Image reconstruction between the input and synthetic image is trained with $L_{rec} = \sum_{s\in\{s_i, s_j^{deformed}, s_j^{fused}\}} \|x_i - g(s, z_i)\|_1$. Essential for disentanglement is the cross-reconstruction between modalities by properly mixing the anatomical and modality factors. In addition, the reconstruction error is back-propagated to the STN and provides the learning signal for aligning anatomical factors.

$\mathbf{L_{seg}}$: When segmentation masks m_i, corresponding to the input x_i, are available, then a supervised cost is defined using differentiable Dice between real and predicted masks: $L_{seg} = \sum_{s\in\{s_i, s_j^{deformed}, s_j^{fused}\}} Dice(m_i, h(s))$.

$\mathbf{L_{adv}}$: Finally, an unsupervised cost with least-squares adversarial loss [15] is defined, $L_{adv} = \sum_{s\in\{s_i, s_j^{deformed}, s_j^{fused}\}} [D_M(h(s))^2 + (D_M(m) - 1)^2]$, using a discriminator over masks D_M. Here, the encoder $f_{anatomy}$ and segmentor h are trained to minimise L_{adv} adversarially against D_M which maximises it.

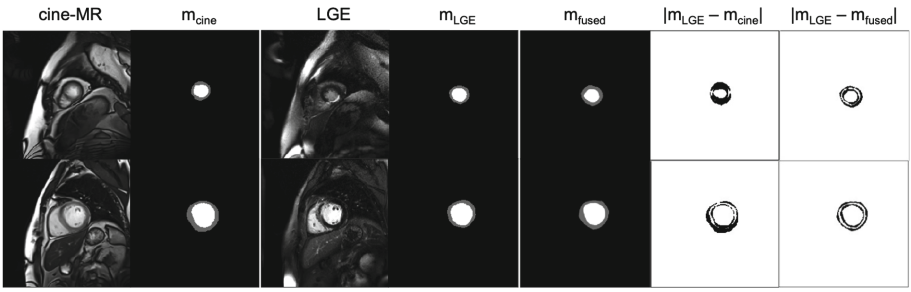

Fig. 4. Two segmentation examples from LGE+cine dataset. Each row shows a paired cine-MR and LGE with their respective ground truth masks (m_{cine} and m_{LGE}); the MMSDNet predicted mask (m_{fused}); and finally, the absolute difference of m_{LGE} with m_{cine} and m_{fused} respectively. Row-wise: $Dice(m_{cine}, m_{LGE})$=0.51, $Dice(m_{fused}, m_{LGE})$=0.81, $Dice(m_{cine}, m_{LGE})$=0.77, $Dice(m_{fused}, m_{LGE})$=0.89.

4 Experiments and Discussion

Data: We evaluate MMSDNet in LGE segmentation using a private dataset acquired at Edinburgh Imaging Facility QMRI with image pairs of 28 patients from cine-MR and LGE [19]. Myocardial contours are provided for the end

diastolic frame of the cine-MR and the LGE data. The spatial resolution is $1.562\,\text{mm}^2$ per pixel, and the slice thickness is 9 mm. The dataset contains 358 expertly paired cine-MR and LGE images and their corresponding segmentation masks. The image resolution is 208×208 pixels.

Baselines: A lower-bound is obtained from the Dice between the real masks of both modalities, referred as "copy masks". This is repeated after affine image registration using mutual information, followed by symmetric diffeomorphic using cross-correlation [21]. We also consider uni- and multi-modal single-input U-Nets by mixing training data. The uni-modal UNet is trained only with the LGE images (UNet-single), whereas the multi-modal UNet is trained with both LGE and cine images (UNet-both). Finally, we compare with DualStream [22] setup of two encoders and decoders, the most recent deep learning method for unpaired multimodal segmentation.

Training and Evaluation: We train, using data augmentations of rotation, translation and scaling in Keras [3], with the Adam optimiser and a learning rate of 0.0001. Results are produced by held-out test sets on 3-fold cross-validation, where the training, validation and test sets are split using 70%, 15% and 15% of the dataset subjects, respectively.

4.1 Multi-input vs. Single-input Segmentation

Initially, we test whether multiple inputs benefit LGE segmentation, compared to single-input models. Two experimental scenarios are considered: LGE masks are available during training or not. Table 1 compares the performance of MMSDNet with the baselines and presents the mean test Dice score of Left Ventricle (LV) and myocardium (MYO) segmentation, as well as their average.

Given fully annotated LGE data (100% column of Table 1), the highest Dice is achieved when using multiple inputs at inference time (MMSDNet-multi), confirming knowledge transfer from source to target modality. The effect of multimodal registration is qualitatively demonstrated in Fig. 4, which shows the improvement achieved by MMSDNet compared with the cine segmentation. MMSDNet, which is trained with multiple inputs, outperforms a single-input U-Net, even when at inference time the paired cine-MR image is not available (referred to as MMSDNet-single in Table 1). Most importantly, when LGE masks are not available during training, but only images (0% column of Table 1), the U-Net and DualStream baselines fail to achieve accurate LGE segmentation since they are only trained on cine-MR data. MMSDNet, with the use of its unsupervised objectives, can still learn multimodal features and outperforms the registration baseline. The achieved Dice scores are comparable with the ones reported in related works [14,24].

4.2 Segmentation with a Varying Number of Annotations

Here we vary the amount of LGE annotations during training to demonstrate the unique capabilities of semi-supervised learning in our approach. In this experiment a fixed number of annotated cine-MR images is used, that is equal

Table 1. Average myocardium and left ventricle test Dice results when training with a varying amount of masks. Best results are underlined; * denotes statistical significance at 0.05 compared to the best baseline. Number of cine-MR masks is always at 100%.

LGE masks:	100%			50%			25%			0%		
	MYO	LV	avg	MYO	LV	avg	MYO	LV	avg	MYO	LV	avg
Copy masks	50_{05}	81_{06}	67_{6}	50_{05}	81_{06}	67_{06}	50_{05}	81_{06}	67_{6}	50_{05}	81_{06}	67_{06}
Registration	51_{08}	80_{07}	68_{07}	51_{08}	80_{07}	68_{07}	51_{08}	80_{07}	68_{07}	51_{08}	80_{07}	68_{07}
UNet-single	66_{07}	87_{03}	78_{04}	64_{11}	83_{13}	76_{12}	51_{10}	75_{15}	66_{14}	-	-	-
UNet-both	$\underline{69}_{02}$	$\underline{89}_{02}$	$\underline{81}_{03}$	64_{10}	84_{08}	76_{08}	56_{09}	79_{12}	71_{10}	27_{17}	44_{27}	38_{23}
DualStream	65_{01}	86_{03}	80_{06}	64_{05}	84_{04}	76_{03}	48_{08}	69_{17}	61_{13}	27_{17}	44_{27}	38_{23}
MMSDNet-single	$\underline{69}_{02}$	86_{04}	80_{04}	64_{08}	81_{10}	75_{08}	61_{07}	84_{06}	75_{06}	56_{07}	83_{04}	72_{06}
MMSDNet-multi	$\underline{69}_{03}$	$\underline{89}_{02}$	$\underline{81}_{02}$	$\underline{65}_{04}$	$\underline{85}_{04}$	$\underline{77}_{04}$	$\underline{63}^{*}_{03}$	$\underline{87}^{*}_{04}$	$\underline{77}^{*}_{03}$	$\underline{59}^{*}_{05}$	$\underline{84}^{*}_{03}$	$\underline{74}^{*}_{04}$

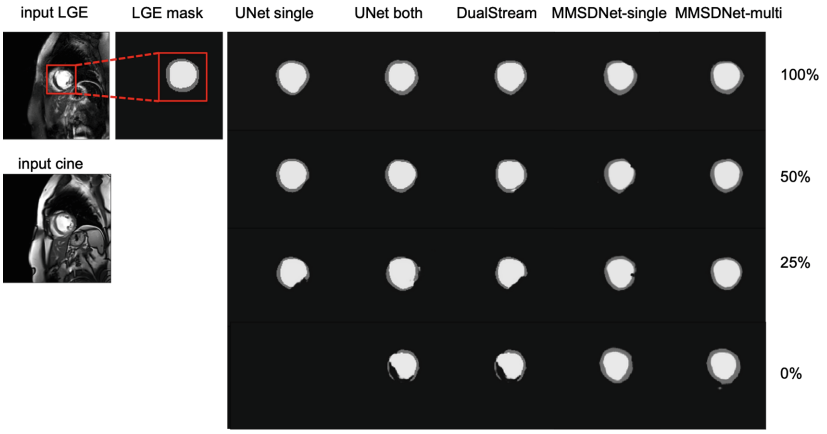

Fig. 5. LGE segmentations when training with varying amounts of LGE annotations.

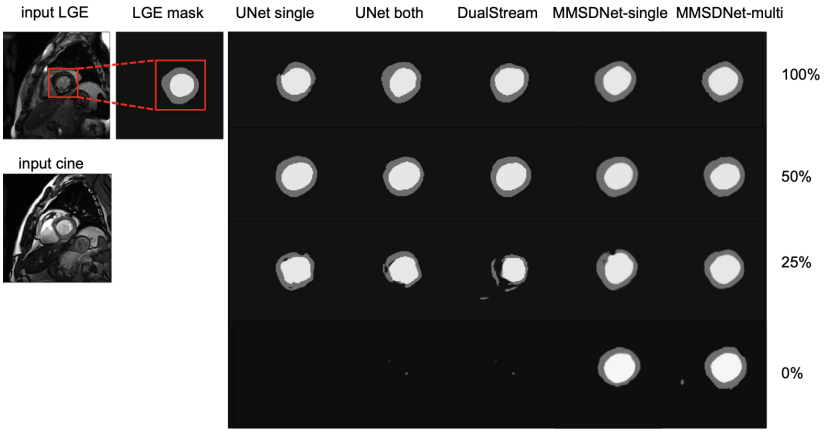

Fig. 6. LGE segmentations when training with varying amounts of LGE annotations. Observe that the baselines did not produce any segmentation mask when trained only with cine-MR data, i.e. for the 0% case.

to the number of LGE images at 100%. Qualitative testing set examples in Fig. 5 and Fig. 6, show the predictions of baseline and MMSDNet models with varying amount of training data. Observe how our approach offers more consistency.

Table 1 reinforces these observations quantitatively on segmentation accuracy for MMSDNet and various baselines. When the number of images is high (above 50%), all methods perform on par. However, as they decrease, the performance of the baselines also decreases. MMSDNet though is consistent and maintains a good performance even when training with no LGE masks. The performance of MMSDNet-multi is always higher than MMSDNet-single, suggesting that our method can leverage information from cine-MR to improve segmentation.

5 Conclusion

We demonstrated multimodal segmentation using input images of different modalities. We devise representation disentanglement to extract the individual anatomical factors, and then use these factors to fuse common and unique information. Our results show that accurate segmentation can be achieved when combining multimodal images, even when no annotations of the target modality are available (during training). We used two MR modalities with expert pairing of the inputs. Our methodology can be extended for additional modalities, by adding new encoders and by accordingly learning a pairing mechanism. Both are under investigation, along with further applications in other organs.

Acknowledgements. This work was supported by UK EPSRC (EP/P022928/1) and US National Institutes of Health (1R01HL136578-01), and used resources from the Edinburgh Compute and Data Facility. S.A. Tsaftaris acknowledges the Royal Academy of Engineering and the Research Chairs and Senior Research Fellowships scheme.

References

1. Bookstein, F.L.: Principal warps: thin-plate splines and the decomposition of deformations. IEEE PAMI **11**(6), 567–585 (1989)
2. Chartsias, A., et al.: Disentangled representation learning in cardiac image analysis. Med. Image Anal. **58**, 101535 (2019)
3. Chollet, F.: Keras (2015). https://keras.io
4. Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C., Ayed, I.B.: HyperDense-Net: a hyper-densely connected CNN for multi-modal image segmentation. In: IEEE TMI (2018)
5. Fidon, L., et al.: Scalable multimodal convolutional networks for brain tumour segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 285–293. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_33
6. Havaei, M., et al.: Brain tumor segmentation with deep neural networks. Med. Image Anal. **35**, 18–31 (2017)
7. Huang, X., Liu, M.-Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 179–196. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_11

8. Huo, Y., et al.: SynSeg-Net: synthetic segmentation without target modality ground truth. In: IEEE TMI (2018)
9. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: NIPS, pp. 2017–2025 (2015)
10. Joyce, T., Chartsias, A., Tsaftaris, S.A.: Robust multi-modal MR image synthesis. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 347–355. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_40
11. Kim, H.W., Farzaneh-Far, A., Kim, R.J.: Cardiovascular magnetic resonance in patients with myocardial infarction: current and emerging applications. JACC **55**(1), 1–16 (2009)
12. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: ICLR (2014)
13. Lee, H.-Y., Tseng, H.-Y., Huang, J.-B., Singh, M., Yang, M.-H.: Diverse image-to-image translation via disentangled representations. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11205, pp. 36–52. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01246-5_3
14. Liu, J., Xie, H., Zhang, S., Gu, L.: Multi-sequence myocardium segmentation with cross-constrained shape and neural network-based initialization. CMIG **71**, 49–57 (2019)
15. Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P.: On the effectiveness of least squares generative adversarial networks. In: IEEE PAMI (2019)
16. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: visual reasoning with a general conditioning layer. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
17. Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., Kamen, A.: Unsupervised deformable registration for multi-modal images via disentangled representations. In: Chung, A.C.S., Gee, J.C., Yushkevich, P.A., Bao, S. (eds.) IPMI 2019. LNCS, vol. 11492, pp. 249–261. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20351-1_19
18. 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
19. Stirrat, C.G., et al.: Ferumoxytol-enhanced magnetic resonance imaging assessing inflammation after myocardial infarction. Heart **103**(19), 1528–1535 (2017)
20. Tseng, K.-L., Lin, Y.-L., Hsu, W., Huang, C.-Y.: Joint sequence learning and cross-modality convolution for 3D biomedical segmentation. In: CVPR, pp. 6393–6400 (2017)
21. Tustison, N.J., Yang, Y., Salerno, M.: Advanced normalization tools for cardiac motion correction. In: Camara, O., Mansi, T., Pop, M., Rhode, K., Sermesant, M., Young, A. (eds.) STACOM 2014. LNCS, vol. 8896, pp. 3–12. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-14678-2_1
22. Valindria, V., et al.: Multi-modal learning from unpaired images: application to multi-organ segmentation in CT and MRI. In: WACV (2018)
23. Zhang, Z., Yang, L., Zheng, Y.: Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network. In: CVPR, pp. 9242–9251 (2018)
24. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. In: IEEE PAMI (2019)
25. Zhuang, X., Shen, J.: Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. Med. Image Anal. **31**, 77–87 (2016)