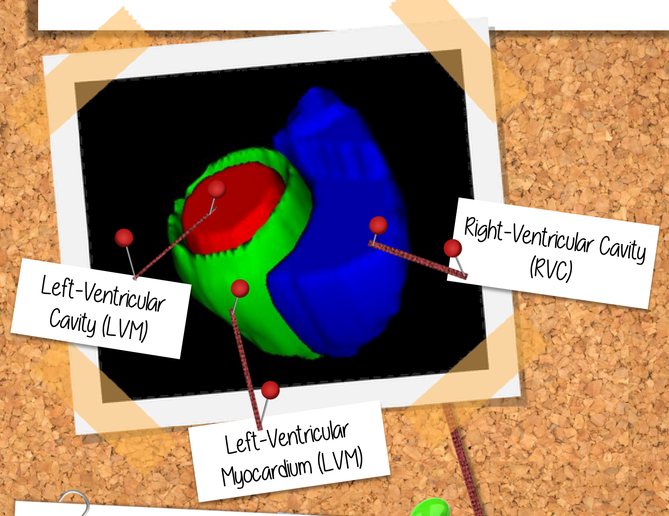
Real-time Prediction of Segmentation Quality
Swapping Expert Annotations for Generated Labels

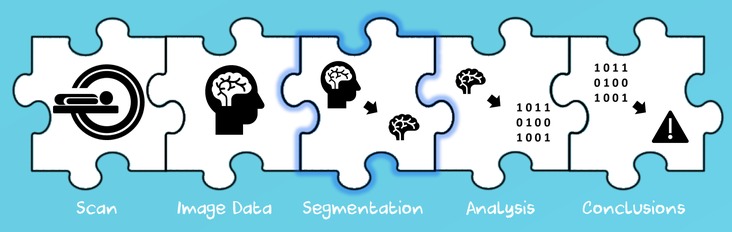
Rob Robinson[1] ✉, O. Oktay[1], W. Bai[1], V.V. Valindria[1], M.M. Sanghvi[3,4], N. Aung[3,4], J.M. Paiva[3], F. Zemrak[3,4], K. Fung[3,4], E. Lukaschuk[5], A.M. Lee[3,4], V. Carapella[5], Y.J. Kim[5,6], B. Kainz[1], S.K. Piechnik[5], S. Neubauer[5], S.E. Petersen[3,4], C. Page[2], D. Rueckert[1], B. Glocker[1]

[1] BioMedIA Group, Imperial College London, UK
[2] Research & Development, GlaxoSmithKline, UK
[3] NIHR Barts, Queen Mary's University UK
[4] Barts Heart Centre, Barts NHS Trust, London, UK
[5] Radcliffe Dept. Medicine, University of Oxford, UK
[6] Yonsei University College of Medicine, South Korea
r.robinson16@imperial.ac.uk



Can we do deep learning without the large manually-annotated dataset?

1



Desirable to perform quality control on-the-fly in real time. Deep Learning is ideal, but requires a large, labelled dataset. These don't exist in practice. So what can we do?

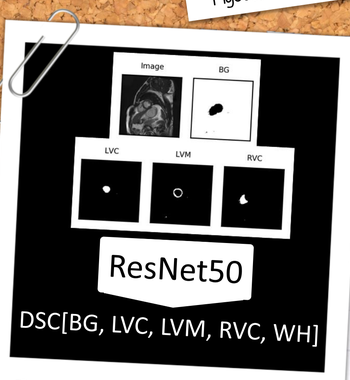


Fig. 2 Network Architecture

In both experiments we use the same architecture. Input: [224,224,8,5] array where the 5 channels are the 3D image and 4 one-hot-encoded segmentations, one for each of the classes: background, left-ventricular cavity, left-ventricular myocardium and right ventricular cavity. Network: A normal ResNet50 Residual Network Output: [1, 5] array of predicted Dice Similarity Coefficients (DSC) for each class and one for the binary 'whole heart' case.

Data

biobank uk — Improving the health of future generations

16,100 2D Stack CMR + Random Forest segmentations. Manual annotations available

2

Use a 'reverse testing' strategy to generate a large set of predictions

↓

Train a network to predict the predictions

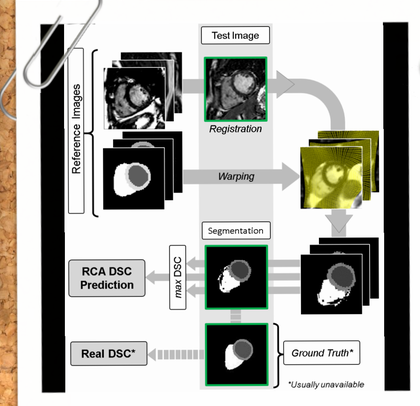


Fig. 1 Reverse Classification Accuracy (RCA)

RCA could be used to generate predictions of quality for a large database of images. Only requirement is a small set of ground truth labelled reference images.

Advantages:
1. No need for expensive, expert manual labelling
2. Extend the training dataset by segmenting at different qualities (e.g. random forest depths)

3

Compare network prediction of RCA predictions (Experiment 2) with network predictions of real DSC from ground truth, manual labelmaps (Experiment 1)

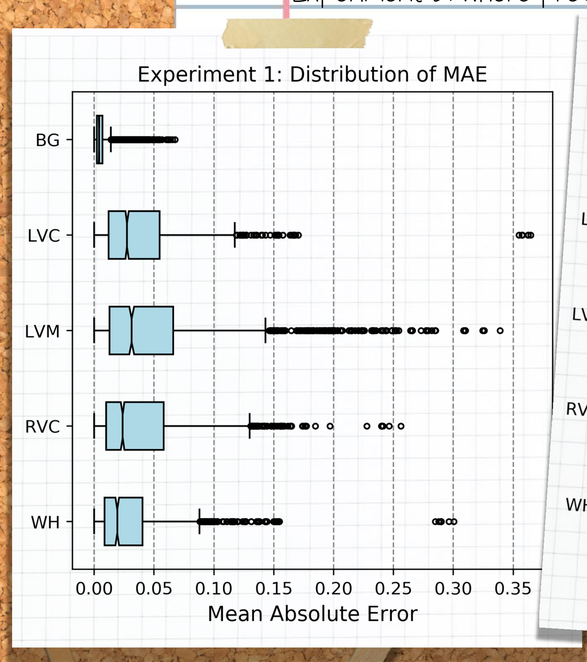
Fig. 3: Results from Both Experiments

Table (top) showing mean absolute error for the different classes from Experiments 1 (left) and 2 (right). MAE for good (DSC $\geq$ 0.5) and poor (DSC < 0.5) ranges shown to be of similar order. True (TPR) and false (FPR) positive rates are shown below. Standard deviation in brackets.

Mean Absolute Error (MAE)

	Experiment 1			Experiment 2		
Class	$0 \leq DSC \leq 1$ $n = 1,610$	$DSC < 0.5$ $n = 817$	$DSC \geq 0.5$ $n = 793$	$0 \leq DSC \leq 1$ $n = 288$	$DSC < 0.5$ $n = 160$	$DSC \geq 0.5$ $n = 128$
BG	0.008 (0.011)	0.012 (0.014)	0.004 (0.002)	0.034 (0.042)	0.048 (0.046)	0.074 (0.002)
LV	0.038 (0.040)	0.025 (0.024)	0.053 (0.047)	0.120 (0.128)	0.069 (0.125)	0.213 (0.065)
LVM	0.055 (0.064)	0.027 (0.027)	0.083 (0.078)	0.191 (0.218)	0.042 (0.041)	0.473 (0.111)
RVC	0.039 (0.041)	0.021 (0.020)	0.058 (0.047)	0.127 (0.126)	0.076 (0.109)	0.223 (0.098)
WH	**0.031 (0.035)**	**0.018 (0.018)**	**0.043 (0.043)**	**0.139 (0.091)**	**0.112 (0.093)**	**0.188 (0.060)**
	TPR 0.975	FPR 0.060	**Acc. 0.965**	TPR 0.879	FPR 0.000	**Acc. 0.906**

Graphs (bottom) of MAE over full DSC range. Low MAE for WH case in both experiments. Expected larger errors for Experiment 2 where predictions are used as GT labels.



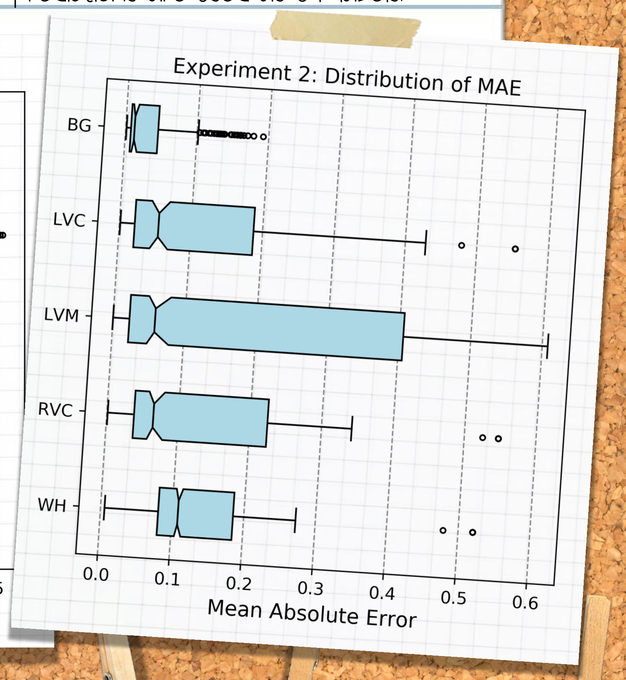
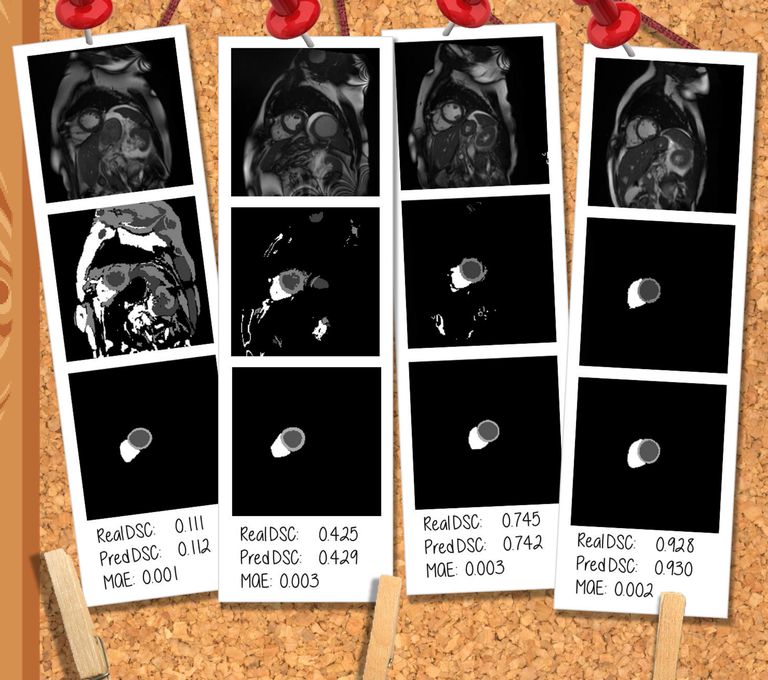


Fig. 4: Example Predictions

Predictions of network trained in Experiment 1. Showing (top) the midsagittal slice from 2D stack of CMR, (middle) random forest generated segmentation and (bottom) the manual ground truth labelmap. Examples segmentations are ordered by increasing quality from left to right. Real and predicted DSC shown with mean absolute error (MAE)



4

What's the point?? There are still some things that need a little more consideration...

- ✓ Real-time feedback to technicians? 40 ms!
- ✓ Is the image analyzable?
- ✓ Reacquisition with patient in-situ? Reduced stress
- ✓ Cost/efficiency benefits?

TO DO

- ✓ Network will learn how to assess Random Forest segmentations. ~~So train the network with CNN/ensemble of segmentations~~
- ✓ Train the network on the logit transform of the DSC to give more fine-grain prediction at higher DSC

References

[1] Valindria, V. V, Lavdas, I, Bai, W, Kamnitsas, K, Aboagye, E. O, Rockall, A. G, Rueckert, D, Glocker, B. (n.d.) "Reverse Classification Accuracy...", 7(4), pp. 1–10, Feb 2017

[2] Petersen, SE et al. "Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort." JCMR 19(1) (2017)

BioMedIA Imperial College London · Imperial College London · King's College London · Medical Imaging EPSRC Centre for Doctoral Training · EPSRC · gsk · MIDL Amsterdam 2018 · MISS 2018 Medical Imaging Summer School · BioMedEng18 · MICCAI2018 Granada Spain