Deep Generative Model-based Quality Control for Cardiac MRI Segmentation

[illegible] [1](✉), [illegible] [2], [illegible] [2], [illegible] [1], [illegible] [2], [illegible] [2], [illegible] [1], [illegible] [2], [illegible] [1,3]

[1] Data Science Institute, Imperial College London, London, UK
[2] BioMedIA Group, Department of Computing, Imperial College London, UK
[3] Department of Brain Sciences, Imperial College London, UK
shuo.wang@imperial.ac.uk

Abstract. In recent years, convolutional neural networks have demonstrated promising performance in a variety of medical image segmentation tasks. However, when a trained segmentation model is deployed into the real clinical world, the model may not perform optimally. A major challenge is the potential poor-quality segmentations generated due to degraded image quality or domain shift issues. There is a timely need to develop an automated quality control method that can detect poor segmentations and feedback to clinicians. Here we propose a novel deep generative model-based framework for quality control of cardiac MRI segmentation. It first learns a manifold of good-quality image-segmentation pairs using a generative model. The quality of a given test segmentation is then assessed by evaluating the difference from its projection onto the good-quality manifold. In particular, the projection is refined through iterative search in the latent space. The proposed method achieves high prediction accuracy on two publicly available cardiac MRI datasets. Moreover, it shows better generalisation ability than traditional regression-based methods. Our approach provides a real-time and model-agnostic quality control for cardiac MRI segmentation, which has the potential to be integrated into clinical image analysis workflows.

Keywords: Cardiac segmentation · Quality control · Generative model.

1 Introduction

[illegible]

Related work:
[6,7,8] [9,10]

Learning-based quality control:
[11] 42
85%
[12]
[13]
[14]

Registration-based quality control:
[15]
[9]
[16]

Contributions:

2 Methodology

2.1 Problem Formulation

Let F denote a segmentation function that maps an image I to a segmentation $\hat{S} = F(I)$. The segmentation $\hat{S}$ is evaluated by a quality function $q(S_{gt}, \hat{S})$ where S_{gt} is the ground truth segmentation and q is a scalar (e.g. Dice). Since S_{gt} is unknown at test time, we aim to learn a function Q such that $Q(\hat{S}; I) \approx q(S_{gt}, \hat{S})$.

2.2 Deep Generative Model-based Quality Control

Instead of learning Q directly from S_{gt}, we propose to learn it from a surrogate ground truth S_{sur} such that $q(S_{sur}, \hat{S}) \approx q(S_{gt}, \hat{S})$. This is achieved by projecting $(I, \hat{S})$ onto a good-quality manifold (Fig. 1).

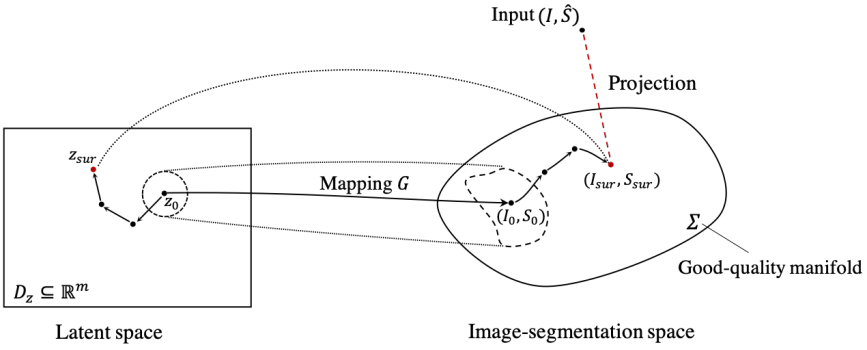

Fig. 1. Overview of the deep generative model-based quality control framework. The generative model G is trained to learn a mapping $G(z)$ from the low-dimensional latent space D_z to the good-quality manifold Σ. The input image-segmentation pair $(I, \hat{S})$ is projected to (S_{sur}, I_{sur}) on the manifold through iterative search, which is in turn used as surrogate ground truth for quality prediction. z_0 is the initial guess in the latent space and it converges to z_{sur}.

Good-quality manifold: We define a good-quality manifold as the set of all good-quality image-segmentation pairs. Let G be a generative model that learns a mapping from a latent space D_z to the joint image-segmentation space $X = (I, S) \in D_I \times D_S$. Here D_I and D_S are the image and segmentation spaces. Given a set of good-quality image-segmentation pairs (I, S_{gt}), we assume that G learns a manifold $\Sigma \subset D_I \times D_S$, called the *good-quality manifold*. The

G $D_z \subseteq \mathbb{R}^m$

$$G(z) : D_z \ni z \mapsto X = G(z) \in D_I \times D_S \tag{1}$$

z m. $G(z)$

Iterative search in the latent space:

$$z_{sur} = \arg\min_{z \in D_z} \mathcal{L}(G(z), (I, \hat{S})) \tag{2}$$

$\mathcal{L}$ $G(z)$ $I, \hat{S})$.

Algorithm 1

Require: A trained generator $G : D_z \ni z \mapsto G(z) = (I, S) \in \Sigma$
Input: Image-segmentation pair $(I, \hat{S})$
Output: Quality prediction $Q(\hat{S}; I)$
Initialization $z = z_0 \in D_z$
while L not converge **do**
  $L = \mathcal{L}(G(z), (I, \hat{S}))$
  $grad = \nabla_z L$ # calculate gradient through back-propagation
  $z = z - \alpha \cdot grad$ # gradient descent with learning rate α
end while
$S_{sur} = G(z_{sur})$
$Q(\hat{S}; I) = q(S_{sur}, \hat{S})$ # perform quality control q (e.g. Dice)

2.3 Generative Model using VAE

E_φ D_ϕ, φ ϕ [17] $I, S)$ E_φ $\mathcal{N}(\mu_z, \sigma_z^2)$ μ_z σ_z^2

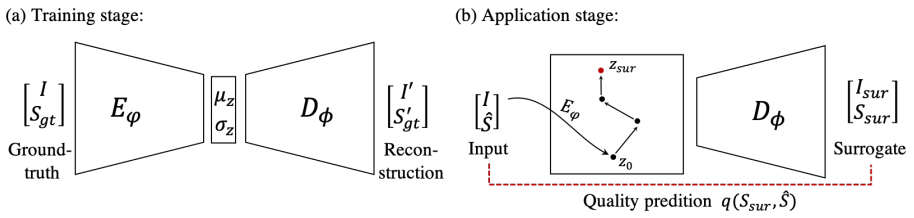

Fig. 2. Framework implementation using the variational autoencoder (VAE). In the training stage, the ground-truth image-segmentation pairs are used. In the application stage, the VAE decoder is used as the generator for iterative search of the surrogate segmentation on the good-quality manifold. Initial guess z_0 is from the encoder.

D_ϕ.

[18]

$$\mathcal{L}_{VAE} = \mathcal{L}_{recon} + \beta \cdot D_{KL}(\mathcal{N}(\mu_z, \sigma_z^2)||\mathcal{N}(0, \mathbf{I})) \quad (3)$$

$$\mathcal{L}_{recon} = BCE(S_{GT}, S'_{GT}) + MSE(I, I') \quad (4)$$

β m

D_ϕ

E_φ.

S_{sur}

$q(S_{sur}, \hat{S})$,

3 Experiments

3.1 Datasets

UK Biobank dataset:

1,500 (800 , 200 , 500).

1.8 1.8 68

2 10 12

ACDC dataset: 100 [illegible] [4] [illegible]

3.2 Experimental Design

[illegible]

VAE implementation and training: [illegible] *ReLU* [illegible] *Sigmoid* [illegible] m [illegible] 8 [illegible] β [illegible] 0.01 [illegible] 0.0001 [illegible] 16. [illegible] 100 [illegible] 96 [illegible] [0 [illegible] , 1] [illegible]

Baseline methods: [illegible] 42 [illegible] [11] [illegible] 8 [illegible] [12] [illegible]

Experiment 1: UK Biobank [illegible]

Experiment 2: ACDC [illegible]

Ablation study: [illegible] m [illegible] β [illegible]

4 Results and Discussion

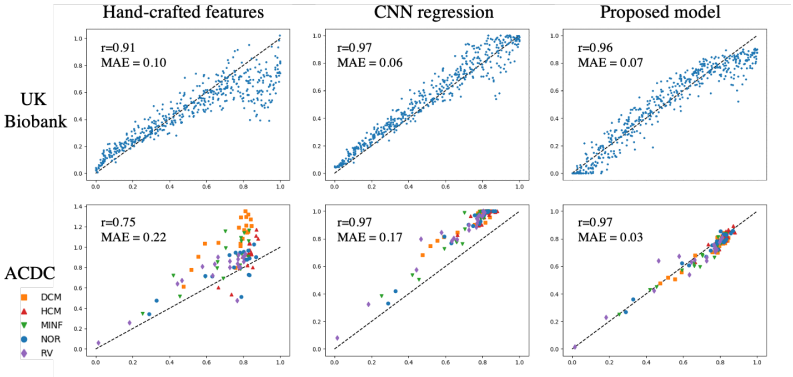

Fig. 3. Comparison of the performance of different quality control methods. The x-axis is the real Dice of each subject and the y-axis is the predicted Dice by each method. The dashed line is the $y = x$, plotted for reference. Top row: UK Biobank data ($n = 500$). Bottom row: ACDC data ($n = 100$), with five subgroups plotted in different colors.

Table 1. Quality control performance of three models on two cardiac datasets. The Pearson correlation coefficient r and the mean absolute error (MAE) between predicted and true Dice metrics are reported. MAE is reported as mean (standard deviation). For ACDC dataset, the performance on five subgroups [4] are aslo reported.

Dataset	Hand-crafted features [11]		CNN regression [12]		Proposed model	
	r	MAE	r	MAE	r	MAE
UK Biobank	0.909	0.100(0.100)	**0.973**	**0.061(0.049)**	0.958	0.067(0.052)
ACDC	0.728	0.182(0.130)	0.968	0.165(0.044)	**0.969**	**0.033(0.028)**
DCM	0.802	0.353(0.123)	0.956	0.186(0.034)	**0.964**	**0.036(0.020)**
HCM	0.838	0.131(0.075)	0.836	0.155(0.027)	**0.896**	**0.023(0.015)**
MINF	0.815	0.184(0.123)	0.969	0.156(0.051)	**0.976**	**0.033(0.029)**
NOR	0.775	0.114(0.065)	0.985	0.158(0.040)	**0.985**	**0.026(0.017)**
RV	0.877	0.129(0.073)	0.974	0.169(0.053)	**0.960**	**0.045(0.042)**

prior to

5 Conclusion

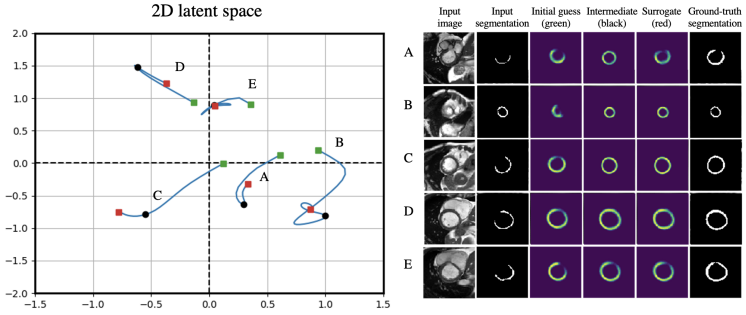

Fig. 4. Visualisation of searching path in a two-dimensional latent space. Left: searching paths for five exemplar samples (green point: initial guess from the VAE encoder; black point: intermediate state during iterative search; red point: convergence point for surrogate segmentation). Right: the input image and segmentation, reconstructed segmentations along the searching path and the ground-truth segmentation.

References

1. WHO: Scale up prevention of heart attack and stroke. `https://www.who.int/cardiovascular_diseases/world-heart-day/en/` accessed 16 March 2020.
2. Bai, W., Sinclair, M., Tarroni, G., Oktay, O., Rajchl, M., Vaillant, G., Lee, A.M., Aung, N., Lukaschuk, E., Sanghvi, M.M., et al.: Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance **20**(1) (2018) 65
3. Tao, Q., Yan, W., Wang, Y., Paiman, E.H., Shamonin, D.P., Garg, P., Plein, S., Huang, L., Xia, L., Sramko, M., 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) (2019) 81–88
4. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al.: Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE transactions on medical imaging **37**(11) (2018) 2514–2525
5. Zheng, Q., Delingette, H., Duchateau, N., Ayache, N.: 3-d consistent and robust segmentation of cardiac images by deep learning with spatial propagation. IEEE transactions on medical imaging **37**(9) (2018) 2137–2148
6. Tarroni, G., Oktay, O., Bai, W., Schuh, A., Suzuki, H., Passerat-Palmbach, J., de Marvao, A., ORegan, D.P., Cook, S., Glocker, B., et al.: Learning-based quality control for cardiac mr images. IEEE transactions on medical imaging **38**(5) (2018) 1127–1138
7. Carapella, V., Jiménez-Ruiz, E., Lukaschuk, E., Aung, N., Fung, K., Paiva, J., Sanghvi, M., Neubauer, S., Petersen, S., Horrocks, I., et al.: Towards the semantic enrichment of free-text annotation of image quality assessment for uk biobank cardiac cine mri scans. In: Deep Learning and Data Labeling for Medical Applications. Springer (2016) 238–248
8. Zhang, L., Gooya, A., Dong, B., Hua, R., Petersen, S.E., Medrano-Gracia, P., Frangi, A.F.: Automated quality assessment of cardiac mr images using convolu-

tional neural networks. In: International Workshop on Simulation and Synthesis in Medical Imaging, Springer (2016) 138–145
9. Robinson, R., Valindria, V.V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., et al.: Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study. Journal of Cardiovascular Magnetic Resonance **21**(1) (2019) 18
10. Albà, X., Lekadir, K., Pereanez, M., Medrano-Gracia, P., Young, A.A., Frangi, A.F.: Automatic initialization and quality control of large-scale cardiac mri segmentations. Medical image analysis **43** (2018) 129–141
11. Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., Grady, L.: Evaluating segmentation error without ground truth. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2012) 528–536
12. Robinson, R., Oktay, O., Bai, W., Valindria, V.V., Sanghvi, M.M., Aung, N., Paiva, J.M., Zemrak, F., Fung, K., Lukaschuk, E., et al.: Real-time prediction of segmentation quality. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2018) 578–585
13. Hann, E., Biasiolli, L., Zhang, Q., Popescu, I.A., Werys, K., Lukaschuk, E., Carapella, V., Paiva, J.M., Aung, N., Rayner, J.J., et al.: Quality control-driven image segmentation towards reliable automatic image analysis in large-scale cardiovascular magnetic resonance aortic cine imaging. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer (2019) 750–758
14. Liu, F., Xia, Y., Yang, D., Yuille, A.L., Xu, D.: An alarm system for segmentation algorithm based on shape model. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 10652–10661
15. Valindria, V.V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B.: Reverse classification accuracy: predicting segmentation performance in the absence of ground truth. IEEE transactions on medical imaging **36**(8) (2017) 1597–1606
16. Haskins, G., Kruger, U., Yan, P.: Deep learning in medical image registration: a survey. Machine Vision and Applications **31**(1) (2020) 1–18
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations. (2014)
18. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-VAE: Learning basic visual concepts with a constrained variational framework. In: International Conference on Learning Representations. (2017)
19. Bojanowski, P., Joulin, A., Lopez-Paz, D., Szlam, A.: Optimizing the latent space of generative networks. In: International Conference on Machine Learning. (2018)

Appendix

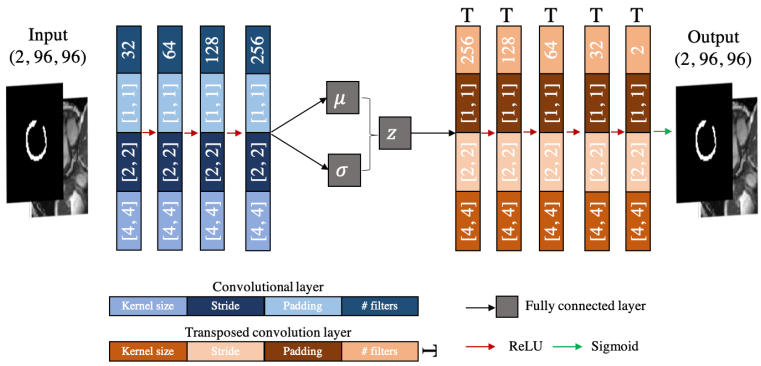

Fig. A1. Network architecture of the VAE.

Table A1. Ablation study of the latent space dimension m and the hyperparameter *beta*. The mean absolute error (MAE) between predicted and true Dice metrics on UKBB validation set are reported. m=8 and $\beta = 0.01$ were selected as the optimal parameters according to the results.

m	β			
	0	1E-3	1E-2	1E-1
2	0.095	0.094	0.105	0.502
4	0.086	0.092	0.114	0.416
8	0.088	0.072	**0.068**	0.229
16	0.147	0.141	0.095	0.091