Weakly-Supervised Deep Learning for Left Ventricle Fibrosis Segmentation in Cardiac MRI Using Image-Level Labels

Roel C Klein[1], Florence E van Lieshout[1], Maarten Z Kolk[1], Kylian van Geijtenbeek[1], Romy Vos[1], Samuel Ruiperez-Campillo[2], Ruibin Feng[2], Brototo Deb[2], Prasanth Ganesan[2], Reinoud Knops[1], Ivana Isgum[3,4], Sanjiv Narayan[2], Erik Bekkers[4], Bob de Vos[3], Fleur V Tjong[1]
on behalf of the DEEP RISK ICD study consortium

[1] Department of Cardiology, Amsterdam University Medical Center location University of Amsterdam, Amsterdam, Netherlands
[2] Department of Medicine, Stanford University, Stanford, United States
[3] Department of Biomedical Engineering, Amsterdam University Medical Center location University of Amsterdam, Amsterdam, Netherlands
[4] Faculty of Science, University of Amsterdam, Amsterdam, Netherlands

Abstract

Automated segmentation of myocardial fibrosis in late gadolinium enhancement (LGE) cardiac MRI (CMR) has the potential to improve efficiency and precision of diagnosis and treatment of cardiomyopathies. However, state-of-the-art Deep Learning approaches require manual pixel-level annotations. Using weaker labels can greatly reduce manual annotation time and expedite dataset curation, which is why we propose fibrosis segmentation methods using either slice-level or stack-level fibrosis labels.

5759 short-axis LGE CMR image slices were retrospectively obtained from 482 patients. U-Nets with slice-level and stack-level supervision are trained with 446 weakly-labeled patients by making use of a myocardium segmentation U-Net and fibrosis classification Dilated Residual Networks (DRN). For comparison, a U-Net is trained with pixel-level supervision using a training set of 81 patients.

On the proprietary test set of 24 patients, pixel-level, slice-level and stack-level supervision reach Dice scores of 0.74, 0.70 and 0.70, while on the external Emidec dataset of 100 patients Dice scores of 0.55, 0.61 and 0.52 were obtained. Results indicate that using larger weakly-annotated datasets can approach the performance of methods using pixel-level annotated datasets and potentially improve generalization to external datasets.

1. Introduction

Evaluation of myocardial fibrosis in the left ventricle (LV) using cardiac magnetic imaging (CMR) with late gadolinium enhancement (LGE) can help diagnose cardiovascular diseases and risk of heart failure [1]. However, the manual segmentation of fibrosis by clinicians is time-consuming, hindering its application in practice. To expedite quantitative evaluation, different methods for automatic and semi-automatic fibrosis segmentation have been created. For fully-automatic segmentation, most recent research has focused on fully-supervised deep learning segmentation networks [2]. A major factor holding back the usage of these models is the scarcity of annotated data. In order to train these networks a handcrafted dataset of pixel-level fibrosis annotations is needed, which is time-consuming to create, requires a high level of expertise and shows high interobserver variability. One approach to reduce the workload of creating training data and reduce training label interobserver variability is to switch from fully-supervised segmentation to weakly-supervised segmentation, which has not been attempted yet for myocardial fibrosis segmentation.

The LGE CMR images for one patient consist of a stack of 2D image slices, so the weakest label is a stack-level binary label. The advantage is that this allows for rapid expansion of training data, albeit this approach gives little information compared to pixel-level labels. A compromise is to use slice-level fibrosis labels, which gives more information than stack-level labels, but is still considerably faster to manually label than pixel-level labels. To aid our weakly-supervised fibrosis segmentation methods we also make use of a myocardium segmentation model, which is trained on a smaller subset of the weakly-labeled training data. This work proposes approaches for fibrosis segmentation using either slice-level or stack-level supervision and compares it to standard pixel-level supervision.

2. Materials

2.1. Deep Risk dataset

The Deep Risk dataset is a private dataset acquired at the Amsterdam University Medical Center, used for training and evaluating the deep learning models. It consists of 5759 short-axis (SAX) LGE CMR image slices, with an average of 11.9 image slices per patient. The patients were at risk for ventricular arrhythmia, for which they have since received an implantable cardioverter defibrillator (ICD). A subset of 117 patients was randomly selected for manual pixel-level myocardium and fibrosis annotations. This fully-labeled dataset was subsequently split into training, validation and test sets of 81, 12 and 24 patients respectively. A weakly-annotated training set, provided with slice-level and stack-level binary fibrosis labels, extends the number of training patients used for weakly-supervised fibrosis segmentation to 446.

2.2. Emidec dataset

The Emidec dataset is an open-source dataset [3], which in our case is only used for external evaluation of the deep learning models. It also consists of SAX LGE CMR images, acquired at the University Hospital of Dijon for patients admitted in the cardiac emergency department with symptoms of a heart attack. The Emidec training set is used for evaluation, which has openly available manual pixel-level annotations of myocardium and fibrosis for 100 patients.

3. Methodology

This work will compare a standard fully-supervised myocardial fibrosis segmentation approach to our weakly-supervised approach. Two types of weak fibrosis labels are considered: a per-slice and a per-stack binary label. The weakly-supervised approach is summarised as follows (see figure 1):

1. **Training a myocardium segmentation model**, using ground truth pixel-level myocardium segmentations as labels.
2. **Training a fibrosis classification model** for either the slice-level or stack-level classification task, using corresponding weak fibrosis labels and the trained myocardium segmentation model.
3. **Creating pixel-level fibrosis pseudo-labels**, using the trained fibrosis classification model and weak fibrosis labels.
4. **Training a fibrosis segmentation model**, using the pixel-level fibrosis pseudo-labels.

The fully-supervised fibrosis segmentation approach is analogous, skipping steps 2 and 3, and training (4) on ground truth pixel-level fibrosis labels.

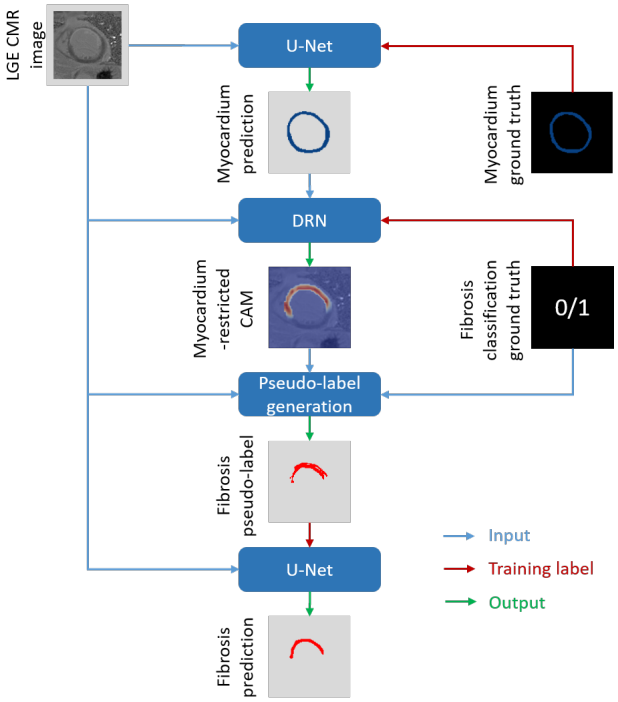


Figure 1: Pipeline for weakly-supervised fibrosis segmentation. Slice-level supervision uses a 2D DRN and slice-level fibrosis labels, while stack-level supervision uses a 3D DRN and stack-level fibrosis labels.

3.1. Myocardium segmentation model

The stack of image slices makes a 3D image, but given the coarse resolution on the z-axis a 2D model is chosen instead. The model architecture is a 2D U-Net, a popular neural network designed for medical image segmentation with an encoder-decoder structure and skip connections [4]. The model is trained using the Dice loss and ground truth pixel-level myocardium labels.

3.2. Fibrosis classification models

Fibrosis classification models provide Class Activation Maps (CAMs), used to initialize weakly supervised segmentation. A convolutional backbone is followed by a 1x1 convolution with one output channel, which outputs the CAM. After this, spatial Global Average Pooling (GAP) aggregates the CAM into an image-level prediction, thus enabling training with image-level labels. Early experiments showed that normal GAP leads to poor fibrosis localization, where high CAM values are mainly located in the center of the image, i.e. the blood pool. To force CAMs into relevant regions, a myocardium restricted pooling is

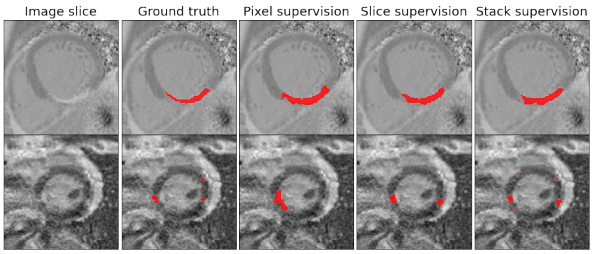
applied. Predictions from the myocardium segmentation model are used to mask out CAM pixels outside of the myocardium prior to GAP. Both LGE CMR images and predictions of the myocardium segmentation model are given at the input layer.

Models for both slice-level and stack-level fibrosis classification are trained. For slice-level classification, the backbone is a 2D Dilated Residual Network (DRN) [5], where the stride in layer 4 is reduced from 2 to 1, resulting in a CAM resolution of 56x56. Stack-level classification requires a 3D model, which is created by replacing the 3x3 convolutions in the slice-level classification model with 3x3x1 convolutions. This means that inter-slice features are not considered, but the coarse resolution along the z-axis is an indicator that the benefit of inter-slice features to classification and CAM quality might be limited. Furthermore, adding inter-slice features could lead to model overfitting. The slice-level fibrosis classification model is trained using the standard binary cross-entropy loss. The stack-level fibrosis classification is instead trained using a weighted binary cross-entropy loss to compensate for the class imbalance on this label.

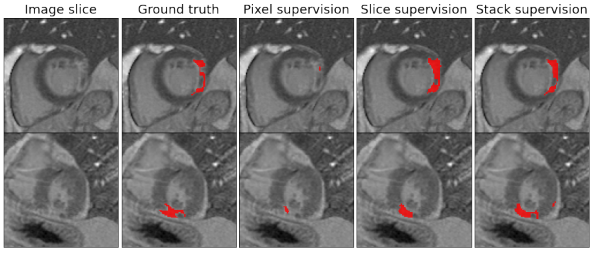
3.3. Fibrosis pseudo-label generation

A series of post-processing steps are performed to transform the restricted CAMs of fibrosis classification models into pseudo-labels, which are subsequently used to train fibrosis segmentation networks. Pseudo-label generation is largely the same for slice-level and stack-level supervision, except for two points. The first difference is that CAMs from either a slice-level or a stack-level fibrosis classification model are used as a starting point. Secondly, pseudo-labels for images with a negative weak label are automatically assumed to contain no fibrosis.

The following steps are therefore only performed for images with a positive weak label. Firstly, CAMs are resized to the original image size using bilinear interpolation. Secondly, the unbound CAM values are converted to a probability map by applying a ReLU and scaling the remaining positive values between 0.5 and 1.0. Next, a 3 class Multi-Otsu thresholding [6] is performed over image pixels corresponding to the remaining positive probabilities. All pixels values below the lowest threshold are then set to probability zero, using the prior knowledge that healthy myocardium is low intensity to improve pseudo-label precision. As a last step, the probability maps are given as input to a Dense Conditional Random Field (DenseCRF) [7], with the purpose of removing small, unconnected pixels with low probability from the pseudo-labels. This results in binary pixel-level fibrosis pseudo-labels. DenseCRF hyperparameters are set to $w^{(1)} = 0$, $w^{(2)} = 1$, $\theta_\gamma = 3$ and $k = 5$.



(a) Examples on the internal Deep Risk dataset. Here no clear preference can be found between pixel-level, slice-level and stack-level supervision.



(b) Examples on the external Emidec dataset. Pixel-level supervision underpredicts, likely due to the smaller fully-labeled training set. The bottom row image shows that stack-level supervision tends to predict false positives in areas disconnected from the actual fibrosis.

Figure 2: Examples of fibrosis segmentation using pixel-level, slice-level and stack-level supervision.

3.4. Fibrosis segmentation model

The fibrosis segmentation model is a 2D U-Net identical to the myocardium segmentation model, which is trained using the Dice loss. For weakly-supervised fibrosis segmentation, only the LGE CMR image slices are given as input. For fully-supervised fibrosis segmentation, predictions from the myocardium segmentation model are given as additional input, since this was found to improve results.

4. Results

For the myocardium segmentation average Dice scores of 0.84 and 0.76 were reached on the internal Deep Risk dataset and external Emidec dataset respectively (Table 2). Slice-level fibrosis classification reached AUROC scores of 0.92 and 0.86 for the Deep Risk and Emidec dataset, while scores of 0.88 and 0.81 are reached for stack-level fibrosis classification (Table 1). Table 2 shows the average fibrosis Dice score for patients with fibrosis. Here we can see that fully-supervised, pixel-level supervision performs best on the Deep Risk dataset (Dice 0.74), followed by slice-level supervision (Dice 0.70) and stack-level supervision (Dice 0.70). Qualitative examples in Figure 2a show no clear preference between the different supervision levels. However, on the external Emidec dataset we see that slice-level supervision performs best (Dice 0.61),

Table 1: Results for the slice-level and stack-level fibrosis classification tasks. Area under the ROC-curve (AUROC).

Classification task	AUROC Deep Risk	AUROC Emidec
Slice-level	0.92	0.86
Stack-level	0.88	0.81

Table 2: Dice scores for myocardium segmentation and fibrosis segmentation with different supervision levels. Average and standard deviation.

Tissue	Supervision	Dice Deep Risk	Dice Emidec
Myocardium	Pixel-level	0.84±0.05	0.76±0.10
Fibrosis	Pixel-level	0.74±0.36	0.55±0.40
Fibrosis	Slice-level	0.70±0.37	0.61±0.37
Fibrosis	Stack-level	0.70±0.37	0.52±0.38

followed by pixel-level supervision (Dice 0.55) and lastly stack-level supervision (Dice 0.52). Qualitative examples are seen in Figure 2b, where pixel-level supervision heavily underpredicts compared to slice-level and stack-level supervision. This poor generalization is likely due to the smaller fully-labeled training set and the distribution shift with respect to imaging and patients. It can also be seen that stack-level supervision makes small predictions in healthy areas, something which heavily drops Dice scores for healthy slices and is a consequence of the limited information in stack-level labels.

5. Discussion and Conclusion

This work has introduced two methods for weakly-supervised myocardial fibrosis segmentation that drastically reduce manual labelling time, using slice-level and stack-level supervision. Results indicate that the usage of larger weakly-labeled datasets can approach the performance reached using painstakingly created pixel-level datasets and potentially improve generalization to new data sources.

One limitation of the weakly-supervised methods is that they still require pixel-level myocardium annotations. If this reliance can be removed or reduced, manual annotation time could be further shortened. Another point for improvement is the incorporation of inter-slice features, something that despite the distance between slices should be able to improve results. More research is necessary to answer how pixel-level, slice-level and stack-level supervision scale with increasing amounts of training data, which in combination with the respective labelling time can help determine which type of supervision is to be preferred in practice. Another line of research could could investigate how to best combine smaller fully-labelled datasets with larger weakly-labeled datasets in order to optimally use all available data.

Acknowledgments

We acknowledge the DEEP RISK ICD study investigators: C.P. Allaart, M.J.W Götte, J.L. Selder, A.C.L. van der Lingen.

This publication is part of the project DEEP RISK ICD (with project number 452019308) of the Rubicon research programme (personal grant F.V.Y.T) which is (partly) financed by the Dutch Research Council (NWO). This research is partly funded by the Amsterdam Cardiovascular Sciences (personal grant F.V.Y.T).

References

[1] Kuruvilla S, Adenaw N, Katwal AB, Lipinski MJ, Kramer CM, Salerno M. Late gadolinium enhancement on cardiac magnetic resonance predicts adverse cardiovascular outcomes in nonischemic cardiomyopathy. Circulation Cardiovascular Imaging 2014;7(2):250–258.

[2] Wu Y, Tang Z, Li B, Firmin D, Yang G. Recent advances in fibrosis and scar segmentation from cardiac mri: A state-of-the-art review and future perspectives. Frontiers in Physiology 2021;12. ISSN 1664-042X.

[3] Lalande A, Chen Z, Pommier T, Decourselle T, Qayyum A, Salomon M, Ginhac D, Skandarani Y, Boucher A, Brahim K, de Bruijne M, Camarasa R, Correia TM, Feng X, Girum KB, Hennemuth A, Huellebrand M, Hussain R, Ivantsits M, Ma J, Meyer C, Sharma R, Shi J, Tsekos NV, Varela M, Wang X, Yang S, Zhang H, Zhang Y, Zhou Y, Zhuang X, Couturier R, Meriaudeau F. Deep learning methods for automatic evaluation of delayed enhancement-mri. the results of the emidec challenge. Medical Image Analysis 2022;79:102428. ISSN 1361-8415.

[4] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In Navab N, Hornegger J, Wells WM, Frangi AF (eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing. ISBN 978-3-319-24574-4, 2015; 234–241.

[5] Yu F, Koltun V, Funkhouser TA. Dilated residual networks. CoRR 2017;abs/1705.09914.

[6] Liao PS, Chen TS, Chung PC. A fast algorithm for multilevel thresholding. J Inf Sci Eng 09 2001;17:713–727.

[7] Krähenbühl P, Koltun V. Efficient inference in fully connected crfs with gaussian edge potentials. In NIPS. 2011; .

Address for correspondence:

Fleur Tjong
Department of Cardiology, Amsterdam UMC
Meibergdreef 9, 1105 AZ Amsterdam
f.v.tjong@amsterdamumc.nl