Cardiac Segmentation of LGE MRI with Noisy Labels

Holger Roth(✉), Wentao Zhu, Dong Yang, Ziyue Xu, and Daguang Xu

NVIDIA, Bethesda, USA
{hroth,wentaoz,dongy,ziyuex,daguangx}@nvidia.com

Abstract. In this work, we attempt the segmentation of cardiac structures in late gadolinium-enhanced (LGE) magnetic resonance images (MRI) using only minimal supervision in a two-step approach. In the first step, we register a small set of five LGE cardiac magnetic resonance (CMR) images with ground truth labels to a set of 40 target LGE CMR images without annotation. Each manually annotated ground truth provides labels of the myocardium and the left ventricle (LV) and right ventricle (RV) cavities, which are used as atlases. After multi-atlas label fusion by majority voting, we possess noisy labels for each of the targeted LGE images. A second set of manual labels exists for 30 patients of the target LGE CMR images, but are annotated on different MRI sequences (bSSFP and T2-weighted). Again, we use multi-atlas label fusion with a consistency constraint to further refine our noisy labels if additional annotations in other modalities are available for a given patient. In the second step, we train a deep convolutional network for semantic segmentation on the target data while using data augmentation techniques to avoid over-fitting to the noisy labels. After inference and simple post-processing, we achieve our final segmentation for the targeted LGE CMR images, resulting in an average Dice of 0.890, 0.780, and 0.844 for LV cavity, LV myocardium, and RV cavity, respectively.

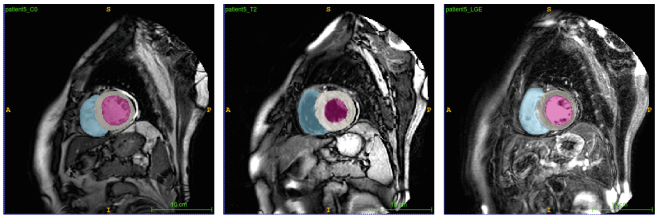
Keywords: LGE MRI · CMR · Cardiac segmentation · Deep learning · Multi-atlas label fusion · Noisy labels

1 Introduction

Segmentation of cardiac structures in magnetic resonance images (MRI) has potential uses for many clinical applications. In particular for cardiac magnetic resonance (CMR) images, late gadolinium-enhanced (LGE) imaging is useful to visualize and detect myocardial infarction (MI). Another common CMR sequence is T2-weighted imaging which highlights acute injury and ischemic regions. Additionally, balanced-steady state free precession (bSSFP) cine sequences can be utilized to analyze the cardiac motion of the heart [1,2]. Each CMR sequence is typically acquired independently, and they can exhibit significant spatial deformations among each other even when stemming from the

M. Pop et al. (Eds.): STACOM 2019, LNCS 12009, pp. 228–236, 2020.
https://doi.org/10.1007/978-3-030-39074-7_24

same patient. Nevertheless, segmentation of different anatomies from LGE could still benefit from the combination with the other two sequences (T2 and bSSFP) and their annotations. An example of different CMR sequences utilized in this work can be seen in Fig. 1. LGE enhances infarcted tissues in the myocardium and therefore is an important sequence to focus on for the detection and quantification of myocardial infarction. The infarcted myocardium tissue appears with a distinctively brighter intensity than the surrounding healthy regions. In particular, LGE images are important to estimate the extent of the infarct in comparison to the myocardium [1]. However, manual delineation of the myocardium is time-consuming and error-prone. Therefore, automated and robust methods for providing a segmentation of the cardiac anatomy around the left ventricle (LV) are needed to support the analysis of myocardial infarction. Modern semantic segmentation methods utilizing deep learning have significantly improved the performance in various medical imaging applications [3–6]. At the same time, deep learning methods typically require large amounts of annotated data in order to train sufficiently robust and accurate models depending on the difficulty of the task. However, in many use cases, the availability of such annotated cases may be limited for a specific targeted image modality or sequence. For CMR applications containing multiple sequences, annotations for the same anatomy of interest might be available for sequences other than the target one of the same patient. In this work, we attempt the segmentation of cardiac structures in LGE cardiac magnetic resonance (CMR) images utilizing classical methods from multi-atlas label fusion in order to provide "noisy" pseudo labels to be used for training deep convolutional neural network segmentation models.



(a) bSSFP (b) T2 (c) LGE

Fig. 1. Sagittal view of different cardiac magnetic resonance (CMR) image sequences of the same patient's heart. Images (a–c) show balanced-steady state free precession (bSSFP), T2-weighted, and late gadolinium-enhanced (LGE) images with overlays of the corresponding manual ground truth (g.t.) annotations [patient 2 of the challenge dataset].

2 Method

Our method can be described in two steps. In the first step, we register a small set, e.g. 5, LGE CMR with ground truth labels ("atlases") to a set of target LGE

CMR images without annotation. Each ground truth atlas provides manually annotated labels of the myocardium, and the left and right ventricle cavities. After multi-atlas label fusion by majority voting, we possess noisy labels for each of the targeted LGE images. A second set of manual labels exists for some of the patients of the targeted LGE CMR images, but are annotated on different MRI sequences (bSSFP and T2-weighted). Again, we use multi-atlas label fusion with a consistency constraint to further refine our noisy labels if additional annotations in other sequences are available for that patient. In the second step, we train a deep convolutional network for semantic segmentation on the target data while using data augmentation techniques to avoid over-fitting to the noisy labels. After inference and simple post-processing, we arrive at our final label for the targeted LGE CMR images.

2.1 Multi-atlas Label Fusion of CMR

Many methods of multi-atlas label fusion exist [7]. In this work, we use a well-established non-rigid registration framework based on a B-spline deformation model [8] using the implementation provided by [9]. The registration is driven by a similarity measurement $\mathcal{S}$ based on intensities from LGE, T2, and bSSFP images. We perform two sets of registrations

1. Inter-patient and intra-modality registration, i.e. the registration of LGE with annotations to the targeted LGE images of different patients.
2. Intra-patient and inter-modality registration, i.e. the registration of bSSFP/T2 with annotations to the targeted LGE images of the same patient.

In both cases, an initial affine registration is performed followed by non-rigid registration between the source image F (providing annotation, i.e. the "atlas") and the targeted reference image R. A coarse-to-fine registration scheme is used in order to first capture large deformations between the images, followed by more detailed refinements. The deformation is modeled with a 3D cubic B-spline model using a lattice of control points $\{\boldsymbol{\phi}\}$ and spacings between the control points of δ_x, δ_y, and δ_z along the x-, y-, and z-axis of the image, respectively. Hence, the deformation $\mathbf{T}(\boldsymbol{x})$ of a voxel $\boldsymbol{x} = (x, y, z)$ to the domain Ω of the target image can be formulated as

$$\mathbf{T}(\boldsymbol{x}) = \sum_{i,j,k} \beta^3\left(\frac{x}{\delta_x} - i\right) \times \beta^3\left(\frac{y}{\delta_y} - j\right) \times \beta^3\left(\frac{y}{\delta_z} - k\right) \times \boldsymbol{\phi}_{ijk}. \quad (1)$$

Here, β^3 represents the cubic B-Spline function. By maximizing an overall objective function

$$\mathcal{O}\left(I_{\mathrm{p}}, I_{\mathrm{s}}\left(\mathbf{T}\right); \{\boldsymbol{\phi}\}\right) = (1 - \alpha - \beta) \times \mathcal{S} - \alpha \times \mathcal{C}_{\mathrm{smooth}}(\mathbf{T}) - \beta \times \mathcal{C}_{\mathrm{inconsistency}}(\mathbf{T}), \quad (2)$$

we can find the optimal deformation field between source and targeted images. Here, the similarity measure $\mathcal{S}$ is constrained by two penalties $\mathcal{C}_{\mathrm{smooth}}$ and $\mathcal{C}_{\mathrm{inconsistency}}$ which aim to enforce physically plausible deformations. The contribution of each penalty term can be controlled with the weights α and β,

respectively. We use normalized mutual information (NMI) [10] which is commonly used in inter-modality registrations [7] as our driving similarity measure

$$\mathcal{S} = \frac{H(R) + H(F(\mathbf{T}))}{H(R, F(\mathbf{T}))}. \tag{3}$$

Here, $H(R)$ and $H(F(\mathbf{T}))$ are the two marginal entropies, and $H(R, F(\mathbf{T}))$ is the joint entropy. In [9], a Parzen Window (PW) approach [11] is utilized to fill the joint histogram necessary in order to compute the NMI between the images efficiently. To encourage realistic deformations, we utilize bending energy which controls the "smoothness" of the deformation field across the image domain Ω:

$$\begin{aligned}\mathcal{C}_{\text{smooth}} = \frac{1}{N} \sum_{\boldsymbol{x} \in \Omega} \Bigg(& \left| \frac{\partial^2 \mathbf{T}(\boldsymbol{x})}{\partial x^2} \right|^2 + \left| \frac{\partial^2 \mathbf{T}(\boldsymbol{x})}{\partial y^2} \right|^2 + \left| \frac{\partial^2 \mathbf{T}(\boldsymbol{x})}{\partial z^2} \right|^2 \\ & + 2 \times \left[\left| \frac{\partial^2 \mathbf{T}(\boldsymbol{x})}{\partial xy} \right|^2 + \left| \frac{\partial^2 \mathbf{T}(\boldsymbol{x})}{\partial yz} \right|^2 + \left| \frac{\partial^2 \mathbf{T}(\boldsymbol{x})}{\partial xz} \right|^2 \right] \Bigg).\end{aligned} \tag{4}$$

In an ideal registration, the optimized transformations from F to R (forward) and R to F (backward) are the inverse of each other. i.e. $\mathbf{T}_{\text{forward}} = \mathbf{T}^{-1}_{\text{backward}}$ and $\mathbf{T}_{\text{backward}} = \mathbf{T}^{-1}_{\text{forward}}$ [12]. The used implementation by [13] follows the approach by [12] using compositions of $\mathbf{T}_{\text{forward}}$ and $\mathbf{T}_{\text{backward}}$ in order to include a penalty term that encourages inverse consistency of both transformations:

$$\mathcal{C}_{\text{inconsistency}} = \sum_{\boldsymbol{x} \in \Omega} \| \mathbf{T}_{\text{forward}} (\mathbf{T}_{\text{backward}} (\boldsymbol{x})) \|^2 + \sum_{\boldsymbol{x} \in \Omega} \| \mathbf{T}_{\text{backward}} (\mathbf{T}_{\text{forward}} (\boldsymbol{x})) \|^2 \tag{5}$$

At each level of the registration, both the image and control point grid resolutions are doubled compared to the previous level. We find suitable registration parameters for both type (1) and type (2) registrations using visual inspection of the transformed image and ground truth atlases. For type (1) registrations, multiple atlases are available to be registered with each target image. We perform a simple majority voting in order to generate our "noisy" segmentation label $\hat{Y}$ for each target image X.

2.2 Label Consistency with Same Patient Atlases

Because of anatomical consistency between different sequences of the same patient, we employ inter-modality registration to obtain noisy labels for LGE images in type (2) registrations. Two sets of segmentations, denoted by $\hat{Y}^{LGE}_{bSSFP}$ and $\hat{Y}^{LGE}_{T2}$, can be obtained from the registrations: bSSFP to LGE, and T2 to LGE. In order to make sure our noisy labels are accurate enough, we only employ the consistent region $\hat{Y}^{LGE}_{bSSFP} \bigcap \hat{Y}^{LGE}_{T2}$ where both segmentations agree. In the non-consistent regions, we still use the noisy label from type (1) registrations. In type (1) registrations, we use symmetric registration with bending energy factor $\alpha = 0.001$ and inconsistency factor $\beta = 0.001$. We use five resolution levels and the maximal number of iteration per level is 300. The final grid spacing along x,

y and z are the same with five voxels. In type (2) registrations, we use six levels and the maximal number of iteration per level is 4000. The final grid spacing along x, y and z are the same with one voxel.

2.3 Deep Learning Based Segmentation with Noisy Labels

In the second step, we train different deep convolutional networks for semantic segmentation on the target data while using data augmentation techniques (rotation, scaling, adding noise, etc.) to avoid over-fitting to the noisy labels.

Given all pairs of images X and pseudo labels $\hat{Y}$, we re-sample them to 1 mm^3 isotropic resolution and train an ensemble $\mathcal{E}$ of n fully convolutional neural networks to segment the given foreground classes, with $P(X) = \mathcal{E}(X)$ standing for the *softmax* output probability maps for the different classes in the image. Our network architectures follow the encoder-decoder network proposed in [14], named *AH-Net*, and [5] based on the popular 3D U-Net architecture [3] with residual connections [15], named *SegResNet*. For training and implementing these neural networks, we used the *NVIDIA Clara Train SDK*[1] and NVIDIA Tesla V100 GPU with 16 GB memory. As in [14], we initialize *AH-Net* from *ImageNet* pretrained weights using a ResNet-18 encoder branch, utilizing anisotropic ($3 \times 3 \times 1$) kernels in the encoder path in order to make use of pretrained weights from 2D computer vision tasks. While the initial weights are learned from 2D, all convolutions are still applied in a full 3D fashion throughout the network, allowing it to efficiently learn 3D features from the image. In order to encourage view differences in our ensemble models, we initialize the weights in all three major 3D image planes, i.e. $3 \times 3 \times 1$, $3 \times 1 \times 3$, and $1 \times 3 \times 3$, corresponding to axial, sagittal, and coronal planes of the images. This approach results in three distinct *AH-Net* models to be used in our ensemble $\mathcal{E}$. The Dice loss [4] has been established as the objective function of choice for medical image segmentation tasks. Its properties make it suitable for the unbalanced class labels common in 3D medical images:

$$\mathcal{L}_{Dice} = 1 - \frac{2\sum_{i=1}^{N} y_i \hat{y}_i}{\sum_{i=1}^{N} y_i^2 + \sum_{i=1}^{N} \hat{y}_i^2} \tag{6}$$

Here, y_i is the predicted probability from our network f and $\hat{y_i}$ is the label from our "noisy" label map $\hat{Y}$ at voxel i. For simplicity we show the Dice loss for one foreground class in Eq. 6. In practice, we minimize the average Dice loss across the different foreground classes. After inference and simple post-processing, we arrive at our final label set for the targeted LGE CMR images. We resize the ensemble models' prediction maps to the original image resolution using trilinear interpolation, fuse each probability map using an *median* operator in order to reduce outliers. Then, the label index is assigned using the *argmax* operator:

$$Y(X) = \text{argmax}\,(\ \text{median}\,(\ \{\mathcal{E}_0(X), \ldots, \mathcal{E}_n(X)\}\ )\ ) \tag{7}$$

[1] https://devblogs.nvidia.com/annotate-adapt-model-medical-imaging-clara-train-sdk.

Finally, we apply 3D largest connected component analysis on the foreground in order to remove isolated outliers.

3 Experiments and Results

3.1 Challenge Data

The challenge organizers provided the anonymized imaging data of 45 patients with cardiomyopathy who underwent CMR imaging at the Shanghai Renji hospital, China, with institutional ethics approval. For each patient, three CMR sequences (LGE, T2, and bSSF) are provided as multi-slice images in the ventricular short-axis views acquired at breath-hold. Slice-by-slice manual annotations of the right and left ventricular, and ventricular myocardium have been generated as gold-standard using ITK-SNAP[2] for training of the mdoels and for evaluation the segmentation results. The manual segmentation took about 20 min/case as stated by the challenge organizers. We also use ITK-SNAP for all the visualizations shown in this paper. For more details, see the challenge website[3]. The available training and test data have the following characteristics:

Training data:

- Patient 1-5:
  - LGE CMR (image + manual label) for validation
  - T2-weighted CMR (image + manual label)
  - bSSFP CMR (image + manual label)
- Patient 6-35:
  - T2-weighted CMR (image + manual label)
  - bSSFP CMR (image + manual label)
- Patient 36-45:
  - T2-weighted CMR (only image)
  - bSSFP CMR (only image)

Test data:

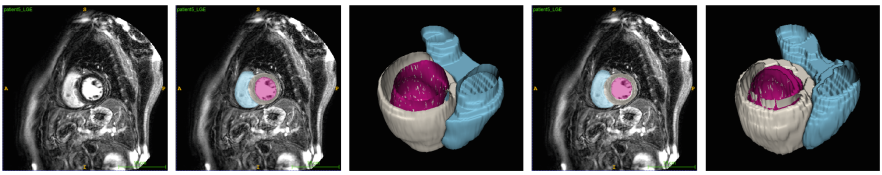
- Patient 6-45:
  - LGE CMR (only image)

As one can see, only five ground truth annotations are available in the targeted LGE images. However, 30 images have gold standard annotations available in different image modalities, i.e. bSSFP and T2. We use all available annotations for type (1) and type (2) multi-atlas label fusion approaches described in Sect. 2. After "noisy" label generation for all testing LGE images, we train our deep neural network ensemble to produce the final prediction labels for 40 LGE images in the test set. The five manually annotated LGE cases are used as the validation set during deep neural network training in order to find the best model

[2] http://www.itksnap.org.
[3] https://zmiclab.github.io/mscmrseg19/data.html.

Table 1. Evaluation scores on 40 LGE test images as provided by the challenge organizers. Both overlap and surface distance-based metrics are shown. LV and RV denote the left and right ventricle, respectively.

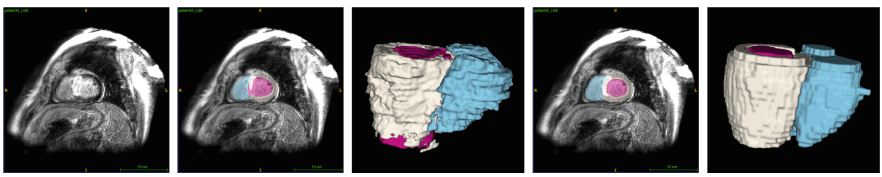
Metric	LV cavity	LV myocardium	RV cavity	Average
Dice	0.890	0.780	0.844	0.838
Jaccard	0.805	0.642	0.735	0.727
Surface distance [mm]	2.13	2.32	2.80	2.41
Hausdorff distance [mm]	11.6	16.3	18.1	15.3



(a) LGE (b) g.t. (c) g.t. 3D (d) pred. (e) pred. 3D

Fig. 2. Comparison of the available ground truth annotation (b) and (c) in a validation LGE dataset and our model's prediction (d) and (e) [patient 2 of the challenge dataset].

parameters and avoid overfitting completely to the noisy labels. Throughout the challenge, the authors are blinded to the ground truth of the test set during model development and evaluation. Our evaluation scores on the test set are summarized in Table 1. A comparison of the available ground truth annotation in a validation LGE dataset and our model's prediction is shown in Fig. 2.



(a) LGE (b) pseudo (c) pseudo (d) pred. (e) pred. 3D

Fig. 3. Comparison of the result after multi-atlas label fusion (b) and (c) in a testing LGE dataset (a) and our model's prediction (d) and (e) [patient 45 of the challenge dataset].

4 Discussion and Conclusion

In this work, we combined classical methods of multi-atlas label fusion with deep learning. We utilized the ability of multi-atlas label fusion to generate labels for

new images using only a small set of labeled images of the targeted image modality as atlases, although resulting in less accurate (or "noisy") labels when compared to manual segmentation. Furthermore, we enhanced the noisy labels by merging more atlas-based label fusion results if annotations of the same patient's anatomy are available in different image modalities. Here, they came from different MRI sequences, but they could potentially stem from even more different modalities like CT, using multi-modality similarity measures to drive the registrations. After training a round of deep convolutional neural networks on the "noisy" labels, we can see a clear visual improvement over multi-atlas label fusion result. This points to the fact that neural networks can still learn correlations of the data and the desired labels even when training labels are not as accurate as ground truth supervision labels [16]. The networks are able to compensate for some of the non-systematic errors in the "noisy" labels and hence improve the overall segmentation. We are blinded to the test set ground truth annotations and cannot quantify these improvements but visually, the improvements are noticeable as shown in Fig. 3. In conclusion, we achieved the automatic segmentation of cardiac structures in LGE magnetic resonance images by combing classical methods from multi-atlas label fusion and modern deep learning-based segmentation, resulting in visually compelling segmentation results.

References

1. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 2933–2946 (2019)
2. 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
3. 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
4. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 565–571. IEEE (2016)
5. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 311–320. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_28
6. Zhu, W., et al.: AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. Med. Phys. **46**(2), 576–589 (2019)
7. Iglesias, J.E., Sabuncu, M.R.: Multi-atlas segmentation of biomedical images: a survey. Med. Image Anal. **24**(1), 205–219 (2015)
8. Rueckert, D., Sonoda, L., Hayes, C., Hill, D., Leach, M., Hawkes, D.: Nonrigid registration using free-form deformations: application to breast MR images. IEEE Trans. Med. Imaging **18**(8), 712–721 (1999)

9. Modat, M., et al.: Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. **98**(3), 278–284 (2010)
10. Studholme, C., Hill, D.L., Hawkes, D.J.: An overlap invariant entropy measure of 3D medical image alignment. Pattern Recogn. **32**(1), 71–86 (1999)
11. Mattes, D., Haynor, D.R., Vesselle, H., Lewellen, T.K., Eubank, W.: PET-CT image registration in the chest using free-form deformations. IEEE Trans. Med. Imaging **22**(1), 120–128 (2003)
12. Feng, W., Reeves, S., Denney, T., Lloyd, S., Dell'Italia, L., Gupta, H.: A new consistent image registration formulation with a B-spline deformation model. In: ISBI, pp. 979–982 (2009)
13. Modat, M., Cardoso, M.J., Daga, P., Cash, D., Fox, N.C., Ourselin, S.: Inverse-consistent symmetric free form deformation. In: Dawant, B.M., Christensen, G.E., Fitzpatrick, J.M., Rueckert, D. (eds.) WBIR 2012. LNCS, vol. 7359, pp. 79–88. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31340-0_9
14. Liu, S., et al.: 3D anisotropic hybrid network: transferring convolutional features from 2D images to 3D anisotropic volumes. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11071, pp. 851–858. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_94
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)
16. Heller, N., Dean, J., Papanikolopoulos, N.: Imperfect segmentation labels: how much do they matter? In: Stoyanov, D. (ed.) LABELS/CVII/STENT -2018. LNCS, vol. 11043, pp. 112–120. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01364-6_13