Adaptive Preprocessing for Generalization in Cardiac MR Image Segmentation

Firas Khader[1(✉)], Justus Schock[2], Daniel Truhn[1,3], Fabian Morsbach[1], and Christoph Haarburger[1]

[1] ARISTRA GmbH, Dusseldorf, Germany
{firas.khader,christoph.haarburger}@aristra.com
[2] Department of Diagnostic and Interventional Radiology, University Hospital Dusseldorf, Dusseldorf, Germany
[3] Department of Diagnostic and Interventional Radiology, University Hospital Aachen, Aachen, Germany

Abstract. Recent advances in deep learning have shown the capability to accurately segment cardiac structures in magnetic resonance images. However, while these models provide a good segmentation performance for the specified datasets, their generalization with respect to unseen data across different MRI scanners, vendors or clinics is still under investigation. Previous work that aims to increase the generalization performance provides proof that emphasizing the model design on a uniform preprocessing step may be more beneficial than searching for a better neural architecture. In this paper we build upon this idea and show that a carefully designed preprocessing pipeline plays an important role in enabling the neural network to generalize to the large variety in MRI images. We evaluate our model in the context of the Multi-Centre, Multi-Vendor & Multi-Disease Cardiac Image (M&Ms) Segmentation Challenge.

Keywords: Cardiac MRI · Cardiac segmentation

1 Introduction

In recent years, cardiac magnetic resonance imaging (MRI) has increasingly been used to assess cardiac function. To quantify parameters such as the ejection fraction and stroke volume, an accurate segmentation of the left ventricle, right ventricle and myocardium in both diastolic and systolic phase is necessary. When manually performed by a medical expert, the segmentation is time-consuming and subject to intra- and inter-rater variability. Therefore, automated segmentation approaches would help to improve reproducibility of segmentations and derived parameters, as well as save valuable time. A number automated segmentation approaches have been proposed [2], many of which have shown impressive performance in terms of typical segmentation performance measures [1,3]. It has even been discussed whether the problem is solved technically [2].

E. Puyol Anton et al. (Eds.): STACOM 2020, LNCS 12592, pp. 269–276, 2021.
https://doi.org/10.1007/978-3-030-68107-4_27

Clinically, the diversity in image acquisition parameters, sequences and reconstruction parameters is much higher than captured by most (public) research datasets. Therefore, the goal of the Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Image Segmentation Challenge (M&M) was to develop generalizable models that show consistent performance across institutions and scanner vendors. As the U-Net [5] still represents the current state-of-the-art method for medical image segmentation [4] our work is based on that approach. Previous work has shown that applying such preprocessing techniques can largely contribute to an improvement of the models' capability to generalize. The nnUNet proposed by Isensee et al. [4] makes use of such a preprocessing pipeline by introducing a resampling step together with a normalization routine in order to mitigate the large variability in medical datasets. Additionally, by performing a quantitative analysis of the training data, important architectural parameters such as the depth and kernel sizes as well as the input patch size can be extracted automatically. Along with the benefit of providing a way to automatically adapt to the variability of the datasets, this approach also reduces effort put into manually finding an appropriate set of hyperparameters for each dataset. In our challenge approach, we particularly focused on appropriate preprocessing to tackle differences between scanner vendors, institutions and acquisition parameters.

2 Materials

The provided dataset used to train our models consists of 150 annotated images that originate from two MRI vendors and were scanned in clinical centers across three different countries (Canada, Germany, Spain). Additionally, the challenge organizers provided 25 unannotated images from a third vendor. The images are four dimensional short-axis cardiac MRI images, given in the form (x, y, z, t), where t denotes the time. The resolution of the in-plane axis across all annotated training samples range from 0.97×0.97 mm to 1.625×1.625 mm, while the resolution for the through-plane axis ranges from 5 mm to 10 mm. Besides many healthy subjects, the dataset includes patients with hypertrophic and dilated cardiomyopathies. Creation of the respective ground-truth has been performed by experienced clinicians that were instructed to label the left- (LV) and right ventricle (RV), as well as the left ventricular myocardium (MYO) for the end-diastolic (ED) and end-systolic (ES) cardiac phases. The time points between these two phases were left unlabeled. Along with the three known vendors provided in the training set, one unseen vendor is included in the test set, resulting in 50 new studies from each of the vendors. To allow for appropriate model selection, 20% of this unseen dataset was used for validation purposes, while the rest is used to rank the challenge participants.

3 Methods

Building up on the premise that data-preprocessing and hyperparameter search for the training routine and architecture can be automated to a great extent by a quantitative analysis of the underlying datasets, we were inspired by the steps taken by Isensee et al. [4] to build an appropriate preprocessing pipeline that ensures better generalization performance. In the following, we will briefly outline the most important steps of the implemented pipeline.

3.1 Preprocessing

Preprocessing is used in our approach to facilitate the learning task for our neural network. This is achieved by performing basic image transformations that aim at harmonizing the variability of voxel geometries and intensity distributions in the training set. The steps listed below are thus incorporated into our preprocessing pipeline and were carried out in the same order as presented here.

Resampling. When dealing with medical images and especially MRI images, different scanners and protocols typically result in an anisotropic voxel spacing across the dataset. This can negatively affect the learning process of CNNs due to inconsistent voxel geometries across the dataset. Thus, in order to allow our architecture to learn about the spatial dimensions of anatomical structures, we resample our images and ground-truth segmentation masks to the median voxel spacing of the used training dataset. Resampling is performed as a nearest-neighbor interpolation for the one-hot encoded segmentation mask, while third-order spline interpolation is carried out for the images. Additionally, when resampling the plane constructed by the high resolution axis separately, bicubic interpolation is used.

Resizing. In order to provide as much context as possible to the CNN, we did not process the images patch-based but as a whole. To allow batch-processing in the training routine, the images were resampled to a common shape. This is necessary because, similar to the voxel spacings, the original shape of the images varies across different MRI scanners. Furthermore, the difference in voxel spacings may also lead to a different shape after resampling, even when the original shape is equal. Therefore, we have decided to resize all the inputs to the median shape of the dataset that is acquired after resampling the images to the median voxel spacing. Resizing is performed by cropping and padding operations to ensure that the voxel spacing across all images remains consistent.

Intensity Normalization. The intensity ranges of the acquired images typically vary across MRI scans from different scanners and protocols. Therefore, we normalized our input images by performing a z-score normalization using the standard deviation and the mean of the respective images. However, as proposed in [4], normalization was performed only on the non-zero part if one quarter of the image is non-zero.

3.2 Augmentation

Augmentation techniques are applied to the input image and the corresponding one-hot encoded segmentation mask to increase the robustness of our model to unseen data. These include mirroring along the image plane and an elastic deformation to simulate the contraction and relaxation of the respective left- and right ventricle as well as the left ventricular myocardium. Following the proposed method for performing the elastic deformation by Simard et al. [6], we first create random displacement fields $\Delta x(x, y)$ and $\Delta y(x, y)$ in the range between -1 and +1. Subsequent convolution with a Gaussian filter then allows a conform displacement of adjacent pixels by choice of a sufficiently large standard deviation σ. In this case, σ has been chosen to equal 30. In order to control the intensity of the deformation, the resulting displacement field is additionally multiplied by a factor α, whose value has been chosen to equal 1550 in our experiments.

3.3 Architecture

Similarly to the network architecture chosen by Iseensee et al. [4] we have decided to base our architecture on the U-Net [5]. In order to allow our model to appropriately adjust to dataset-specific features, the number of pooling layers and the corresponding kernels are chosen automatically. This was done by defining a minimal edge length for the feature map in the bottleneck that should not be undershot when downsampling the image by means of a max-pooling operation. As a result, every axis may be subject to a different number of pooling layers. To ensure that axes with a relatively small dimensionality don't fall below the minimal edge length in the bottleneck, while still enforcing the previously computed number of pooling steps, no max-pooling was performed on these axes in the first layers of the network. This was achieved by assigning a value of 1 to the corresponding kernel dimension of the max-pooling layer.

On the other hand, the kernel dimension corresponding to the longest axis is set to 2, starting at the first layer of the encoder network. Throughout the course of our experiments, two different U-Net architectures have been implemented to deal with the three-dimensional input data: The first one consists of a simple 2D U-Net, where the input to the network is chosen to be the image plane, formed by the two axes with the highest resolution. The second architecture replaces every 2D layer in the 2D U-Net by a corresponding 3D layer, in order to construct a 3D network.

3.4 Hyperparameters

While most hyperparameters are set automatically by the dataset analysis explained above, we still had to tune few parameters manually. Given that we have trained our models on a GTX 1080 TI and have chosen the median size after resampling as the common size for all inputs, the batch size has been tuned to consume the remaining GPU memory. This adjustment results in a batch size of 2 in the case of our 3D U-Net and a batch size of 23 in the case of the 2D U-Net. Training has been performed using the Adam optimizer with a learning rate of $1 \cdot 10^{-3}$. The loss function used throughout our training routine is the soft dice loss.

3.5 TTA and Ensembling

In order to increase the accuracy of our model we incorporated test time augmentation (TTA) into our post-processing pipeline TTA is applied by performing horizontal flips and rotations in 90 degree steps from 0 to 270. The final model output is then a result of taking the mean over the predictions for each augmentation step. To further boost the performance, we also built an ensemble of models by means of a majority voting over the predicted labels. The models used in the ensembling are acquired by performing a 5-fold cross-validation for the 2D and 3D U-Net.

4 Results

In order to evaluate our approach and tune hyperparmeters such as early stopping, we performed a 5-fold cross validation over the 150 labelled training images. Splitting between training, validation and test set was performed at patient level to prevent information leakage. For each split, we assigned 64% of the data to the training set, 16% to the validation set and the remaining 20% to the test set for performance assessment. The results for 2D and 3D U-Nets with various combinations of post-processing are depicted in Table 1. With a 3D network

Table 1. Result metrics for all models. All values refer to medians and interquartile ranges based on the 5 folds on the annotated training data.

Predictor	Dice LV	Dice RV	Dice MYO	Hausdorff LV	Hausdorff RV	Hausdorff MYO
2D noTTA noEnsembling	0.69 [0.52, 0.77]	0.78 [0.66, 0.84]	0.43 [0.29, 0.52]	50.33 [27.57, 78.91]	29.50 [15.14, 44.84]	110.43 [92.25, 133.1]
2D TTA ensembling	0.72 [0.60, 0.80]	0.83 [0.78, 0.88]	0.60 [0.50, 0.69]	21.2 [16.37, 27.44]	13.94 [10.57, 15.42]	17.31 [14.50, 23.75]
3D+2D TTA ensembling	0.82 [0.73, 0.87]	0.90 [0.86, 0.93]	0.79 [0.73, 0.82]	15.32 [12.15, 20.04]	**9.96 [8.47, 11.53]**	12.98 [10.72, 17.19]
3D noTTA noEnsembling	0.83 [0.75, 0.88]	0.90 [0.86, 0.93]	0.81 [0.76, 0.84]	19.36 [13.64, 49.93]	10.40 [8.55, 12.63]	13.30 [10.35, 17.78]
3D TTA noEnsembling	0.83 [0.76, 0.89]	0.90 [0.85, 0.93]	0.81 [0.76, 0.85]	14.64 [12.0, 19.3]	10.04 [8.25, 11.3]	12.46 [10.27, 16.31]
3D noTTA ensembling	**0.85 [0.79, 0.89]**	**0.91 [0.87, 0.94]**	0.81 [0.77, 0.85]	14.20 [11.83, 18.46]	10.03 [7.87, 11.9]	12.43 [10.39, 15.91]
3D TTA ensembling	**0.85 [0.79, 0.89]**	**0.91 [0.87, 0.94]**	**0.82 [0.77, 0.85]**	**13.72 [11.51, 18.39]**	9.98 [7.2, 10.87]	**12.35 [10.2, 15.83]**

employing TTA and ensembling, the best performance is achieved. The 2D approach performs considerable worse than the 3D approach. In Fig. 1 we can see that the performance depends on the imaging centre. Example segmentations are provided in Fig. 2.

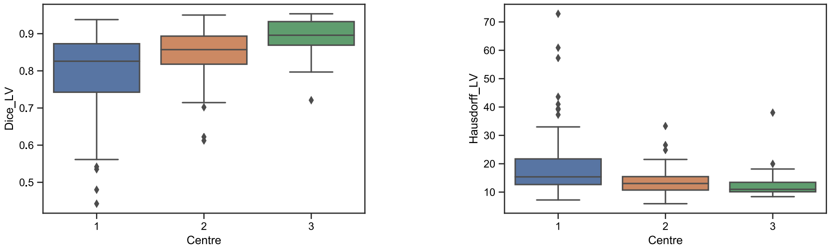

Fig. 1. Dice score (left) and Hausdorff distance (right) performance for the three imaging centres. The evaluation has been performed across the whole training set using the 3D U-Net with TTA and ensembling over all five folds.

5 Discussion

We have described an approach for automated segmentation of left ventricle, right ventricle and myocardium in cardiac MR images. Our results show that generally, a good performance is achieved for a variety of imaging protocols. The 3D segmentation approach is clearly superior to the 2D variant. Interestingly, even when 2D and 3D are combined in an ensemble, the 3D-only approach is superior. Furthermore we found that especially for the left ventricle that tends to be rather challenging to segment, TTA and ensembling strongly improve the performance in terms of both median and interquartile ranges. This indicates that TTA and ensembling are an effective tool for improving generalization. Manually inspecting the segmentations of our approach, we found that most segmentation errors arise at the boundaries of the cardiac structures along the long axis. In this area, we assume that even the ground truth is ambiguous and subject to high intra- and inter-rater variability, especially across institutions and acquisition protocols.

Moreover, the performance is still dependent on the imaging center. Baumgartner et al. achieved a higher segmentation performance on a different dataset using a 2D U-Net [1]. The same applies for [3] using an ensemble of 2D and 3D U-Nets.

In future work, our approach will be refined by further morphological postprocessing. In addition, more thorough evaluations should assess to what extends certain sequences lead to segmentations with high and low performance, respectively.

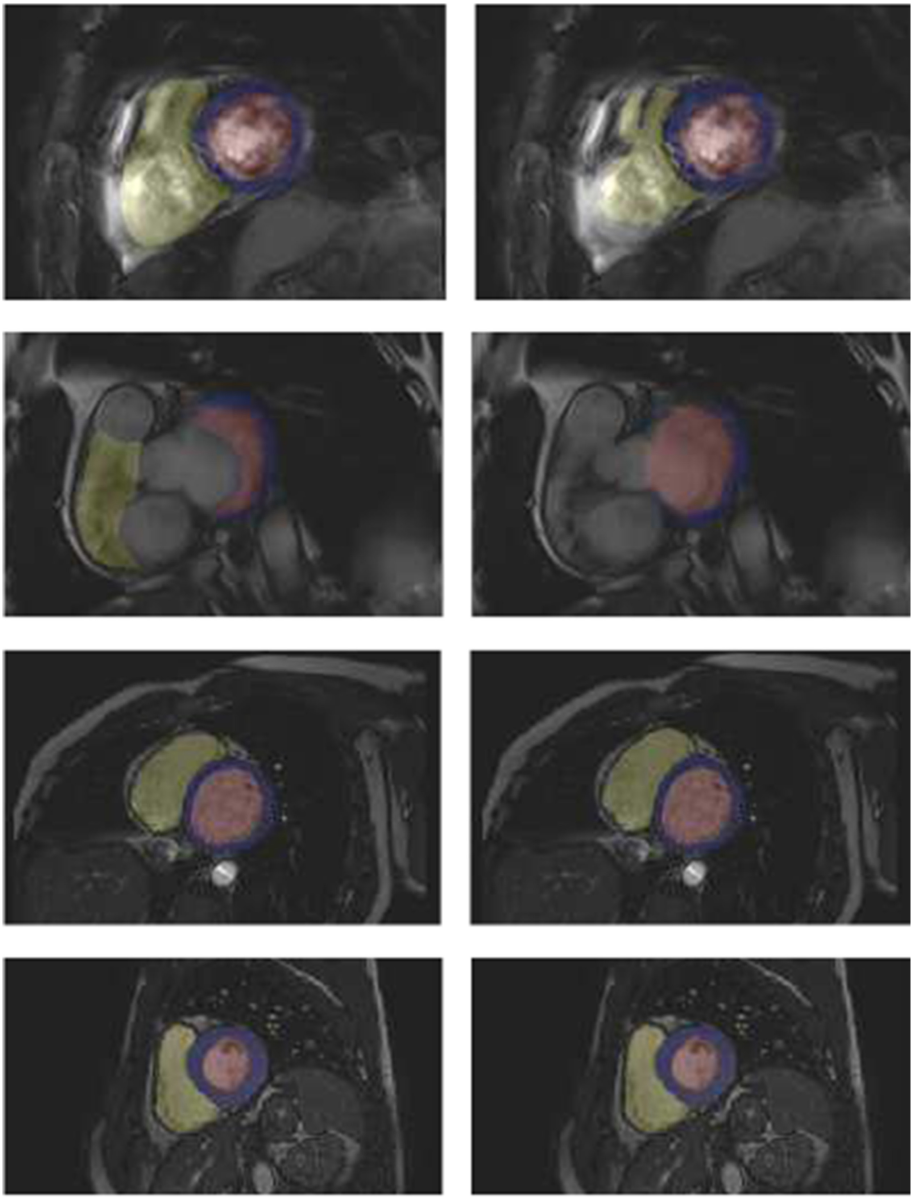

Fig. 2. Ground truth segmentations (left column) and segmentations provided by the 3D U-Net with TTA and ensembling (right column). The first and second row show poorly performing cases, whereas the third and fourth row show examples for cases with good performance.

6 Conclusion

We have shown an approach for automated segmentation of left ventricle, right ventricle and myocardium on cardiac MRI images. Our method is based on an adaptive preprocessing that takes voxel geometry and acquisition parameters for the parametrization of the U-Net architecture and preprocessing pipeline. Furthermore, we have shown that TTA and ensembling improve the performance and generalization.

Acknowledgements. The authors declare that the segmentation method implemented for participation in the M&Ms challenge has not used any pre-trained models nor additional MRI datasets other than those provided by the organizers.

References

1. Baumgartner, C.F., Koch, L.M., Pollefeys, M., Konukoglu, E.: An exploration of 2d and 3d deep learning techniques for cardiac mr image segmentation (2017)
2. Bernard, O., et al.: Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? IEEE Trans. Med. Imaging **37**(11), 2514–2525 (2018). https://doi.org/10.1109/tmi.2018.2837502, https://doi.org/10.1109/tmi.2018.2837502
3. Isensee, F., Jaeger, P., Full, P.M., Wolf, I., Engelhardt, S., Maier-Hein, K.H.: Automatic cardiac disease assessment on cine-mri via time-series segmentation and domain specific features (2017). https://doi.org/10.1007/978-3-319-75541-0
4. Isensee, F., et al.: nnu-net: self-adapting framework for u-net-based medical image segmentation (2018)
5. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI, pp. 234–241. Springer International Publishing (2015)
6. Simard, P.Y., Steinkraus, D., Platt, J.C., et al.: Best practices for convolutional neural networks applied to visual document analysis. In: Icdar vol. 3 (2003)