A generalizable deep-learning approach for cardiac magnetic resonance image segmentation using image augmentation and attention U-Net*

Fanwei Kong[1][0000−0003−1190−565X] and Shawn C. Shadden[1][0000−0001−7561−1568]

Department of Mechanical Engineering, University of California, Berkeley, CA 94720
USA {fanwei_kong,shadden}@berkeley.edu

Abstract. Cardiac cine magnetic resonance imaging (CMRI) is the reference standard for assessing cardiac structure as well as function. However, CMRI data presents large variations among different centers, vendors, and patients with various cardiovascular diseases. Since typical deep-learning-based segmentation methods are usually trained using a limited number of ground truth annotations, they may not generalize well to unseen MR images, due to the variations between the training and testing data. In this study, we proposed an approach towards building a generalizable deep-learning-based model for cardiac structure segmentations from multi-vendor,multi-center and multi-diseases CMRI data. We used a novel combination of image augmentation and a consistency loss function to improve model robustness to typical variations in CMRI data. The proposed image augmentation strategy leverages un-labeled data by a) using CycleGAN to generate images in different styles and b) exchanging the low-frequency features of images from different vendors. Our model architecture was based on an attention-gated U-Net model that learns to focus on cardiac structures of varying shapes and sizes while suppressing irrelevant regions. The proposed augmentation and consistency training method demonstrated improved performance on CMRI images from new vendors and centers. When evaluated using CMRI data from 4 vendors and 6 clinical center, our method was generally able to produce accurate segmentations of cardiac structures.

Keywords: MRI segmentation · Generalization · Image augmentation · Deep learning

1 Introduction

Cine cardiac MRI (CMRI) is considered a reference standard for assessments of the function and morphology of the heart. Analysis of the heart from CMRI can be essential in disease diagnosis and treatment planning. This analysis is greatly facilitated by proper identification of the left ventricular blood pool, myocardium, and the right ventricular blood pool at both end diastolic and end

* This works was supported by the National Science Foundation, Award No. 1663747.

systolic phases. Recent developments in deep learning (DL) are accelerating this previously time-consuming identification process. This has been accomplished by supervised learning of deep neural networks with previously annotated data [2, 1]. However, it is well known that DL methods are prone to over-fitting training data and in turn under-performing on real-world data, especially when the amount of training data is limited. Prior DL-based CMRI segmentation methods were usually trained using small datasets obtained from only one or two sources [6]. However, CMRI data are sensitive to a number of factors, including differences in vendor, magnetic coil types and/or acquisition protocols. Thus, the performance of DL based methods can drop significantly when tested on images that differ from the training data [9, 3]. An outstanding challenge has been to develop generalizable DL based methods that can perform consistently well across different centers, making them useful for real-world clinical applications.

Recent works have helped to improve the generalization capabilities of DL based models. Tao et. al. trained a conventional U-Net model on a large multi-vendor, multi-center training set from patients with various cardiovascular diseases[9]. Chen et. al. showed that training models using a single yet large data source with appropriate data normalization and augmentation could also achieve promising performance on data from other sources [3]. . However, the collection and labeling of such large or diverse datasets are extremely expensive, which limits real-world applicability or adaptability to other segmentation tasks. Therefore, several studies have sought to use unsupervised domain adaptation techniques to optimize the model on an unannotated target dataset [5, 4]. Such methods require images from new sources and their generalization capabilities were usually tested with only one new data source with a limited number of samples. Thus, building generalizable DL models that can be reliably and efficiently applied to data from new clinical centers and scanner vendors remains to be demonstrated.

In this study, we proposed an approach towards building a generalizable DL based model for cardiac structure segmentations from multi-vendor, multi-center and multi-diseases CMRI data. We develop a fully automated segmentation model based on an attention-gated U-Net model [8]. To improve model robustness to typical spatial and intensity variations of cardiac MR images, we propose a novel combination of image augmentation and consistency loss. The proposed image augmentation strategy leverages un-labeled data by a) using CycleGAN to generate images in different styles and b) exchanging the low-frequency features of images between different vendors. A consistency loss is introduced to coerce the model to generate consistent predictions on images with the same anatomical features but different appearances. Our framework demonstrated improved segmentation performance on CMRI images from new vendors and clinical centers.

2 Methods

2.1 Image Dataset Information and Pre-Processing

Image data from the 2020 Multi-Centre, Multi-Vendor& Multi-Disease Cardiac Image Segmentation Challenge (M&Ms) was utilized. This dataset contains CMRI scans from 4 vendors and 6 clinical centers. The training set consisted of 150 annotated images from 2 vendors (Vendor A and B) and 25 unannotated images from another vendor (Vendor C). The ground truth annotations include the left and right ventricular blood pools and left ventricular myocardium. Trained models were validated using a separate validation set, which contained a collection of 80 CMRI scans from all vendors/centers. The final model was evaluated on the M&Ms test set containing 160 CMRI scans. We sliced each stacked 3D image volume into 2D short-axis images and resampled each 2D slice to the same size of 256×256 and a pixel spacing of 1.2 mm. For each scan, we clipped the pixel intensity values between 0 and the 99th percentile to reduce bright artifacts, and normalized the pixel intensity for each slice to zero mean and unit variance.

2.2 Image Augmentation

Image augmentation was used to improve the robustness of deep neural network models to certain image variations. We considered two categories of common variations for cardiac MR images, spatial and appearance. For spatial variations of the heart, we randomly scaled the training images by a factor of 0.8 to 1.2. We also randomly rotated the images clockwise or counter-clockwise by 90 degrees and then applied a small amount of random rotation by up to 10 degrees. For appearance variations of the images, we used frequency-domain augmentation and Cycle GAN to change pixel intensities, as described in more detail below.

Frequency Domain Augmentation (FDA) The intensity of MR images does not have fixed meaning; tissue intensities can vary significantly across different vendors and clinical centers even after applying intensity normalization as described above. To better handle this factor, we propose perturbing the low-frequency contents of an MR image. Namely, as shown in figure 1, we augmented Vendor A and B images with the low-frequency features of the unlabeled Vendor C images, to introduce the low-level statistics of Vendor C images into our training data, similar to a strategy proposed in [10]. The two image slices are extracted at the same relative location from their image stacks so that they show a similar region of the hearts. Namely, for a labeled image slice x^U from Vendor A or B and its corresponding unlabeled Vendor C image slice x^L showing the similar location of the heart, the augmented image slice $x^{L \longrightarrow U}$ can be obtained by equation 1, where $\mathcal{F}_\mathcal{A}$ and $\mathcal{F}_\mathcal{P}$ denote the amplitude and phase components of the Fouirer transform $\mathcal{F}$, and M is a mask with zero values except for the center square as illustrated in figure 1. We used a dimension ratio of 0.02 between the swapped region and the full image.

$$x^{L\rightarrow U} = \mathcal{F}^{-1}\left(\left[M \circ \mathcal{F}_{\mathcal{A}}(x^U) + (1-M) \circ \mathcal{F}_{\mathcal{A}}(x^L),\right.\right. \\ \left.\left. M \circ \mathcal{F}_{\mathcal{P}}(x^U) + (1-M) \circ \mathcal{F}_{\mathcal{P}}(x^L)\right]\right) \quad (1)$$

Our method differed from [10] in that we swapped both the amplitude and phase information in the frequency domain; swapping only the amplitude, as in [10], led to numerous artifacts.

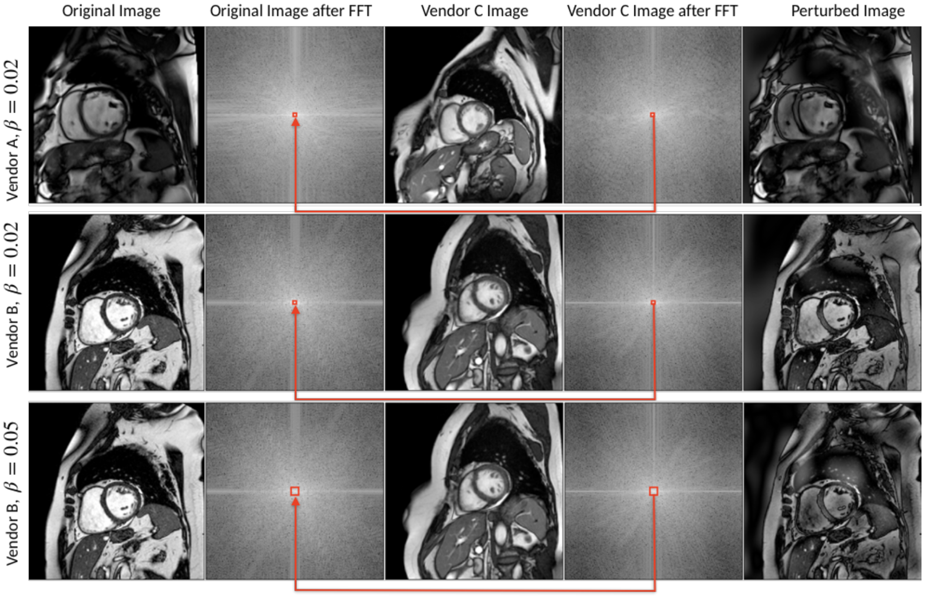

Fig. 1: Illustration of image augmentation in the frequency domain. An image slice (first column) from a labeled dataset (e.g., Vendor A in the first row or Vendor B in the second row) is selected and its magnitude spectrum from FFT is computed (second column). A corresponding image slice (third column) from the unlabeled Vendor C data is selected, and its magnitude spectrum from FFT is computed (fourth column), and subsequently used to augment the spectrum of the original image to generate a perturbed image (fifth column). The third row illustrates the influence of using a larger dimension ratio of 0.05 between the swapped region and the full image. Red boxes represent the center regions where the amplitude and phase components of fourier transformed images were swapped.

Appearance Augmentation Using Cycle GAN Although frequency domain augmentation perturbs image intensity while preserving anatomical features, it sometimes introduces unrealistic intensity inhomogeneity. Observing that MR images from different vendors are different in appearance, we used Cycle GAN [11] to transfer the appearance of images from Vendor B to Vendor A or C and vice versa. Briefly, Cycle GAN takes in two images from two styles and output the corresponding synthetic images that have the texture appearance in the other style, respectively. It consists of two generator networks to generate synthetic images and two discriminator networks that attempt to discriminate generated images from real images. Compared with the original implementation [11], we reduced the learning rate of the discriminator to 0.00002 to achieve a better balance between the generators and the discriminators and replaced the transpose convolution layer with bilinear upsampling followed by a convolution layer to reduce checkerboard artifacts [7]. We trained the Cycle GAN models for 30 epochs and saved the model weights at the end of each epoch after the 5th epoch. For each image in Vendor A or B, we augmented it with 10 CycleGAN models randomly picked from the saved ones. As shown in figure 2, the augmented images resemble the appearance of images from the other vendor and by using CycleGAN models saved at different epochs, we obtained further intensity variations among the generated images.

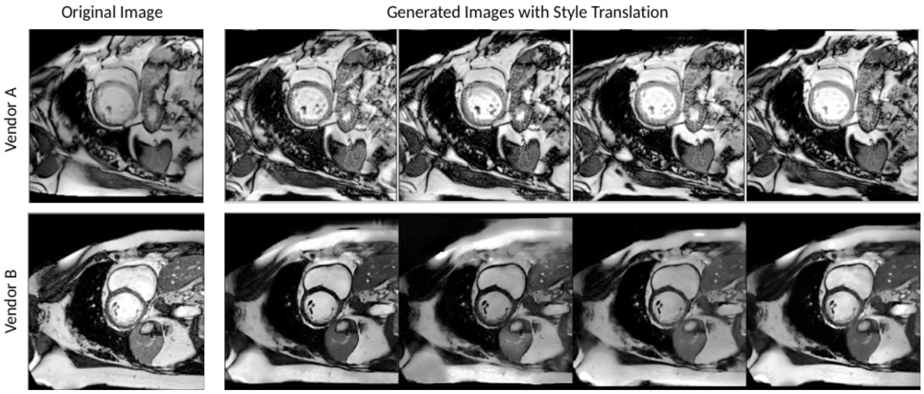

Fig. 2: Original images and the generated images after style translation.

2.3 Training with Consistency Loss

A robust segmentation model should generate consistent predictions for two images with the same anatomical features but different appearance. That is, after applying intensity augmentation to an image slice as described in the previous section, a robust model should predict similar probability maps between the augmented image and the original image. Therefore, we trained our models with two types of loss. One was a hybrid loss L_{seg}, accounting for cross entropy loss and the dice loss, which optimizes for the accuracy of the predicted segmentation.

The other was a consistency loss $L_{consistency}$, which regularizes the differences between the predictions P of one image slice I and its intensity-augmented version $A(I)$. The hybrid loss and consistency loss are defined by the following equations, where G is the one-hot coded ground truth segmentation and N is the number of segmentation domains:

$$L_{seg}(I, G) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{x\in I} G_i(x)\log(P_i(x)) + N - \sum_{i=1}^{N}\frac{2\sum_{x\in I} G_i(x)P_i(x)}{\sum_{x\in I} G_i(x) + \sum_{x\in I} P_i(x)} \tag{2}$$

$$L_{consistency}(I, A(I)) = -\frac{1}{N}\sum_{i=1}^{N}|P_i(I) - P_i(A(I)))|. \tag{3}$$

As illustrated in figure 3, during each training iteration, our proposed pipeline takes in one image slice and its intensity-augmented version and generates predictions separately for the two inputs. The model parameters are then updated based on the sum of the hybrid loss computed for the two predictions and the consistency loss. Our model architecture is based on an attention-gated U-Net model that learns to focus on cardiac structures of varying shapes and sizes while suppressing irrelevant regions [8]. We used an Adam stochastic gradient descent algorithm with an initial learning rate of 0.0005. We randomly split the training dataset into five folds and used one fold as validation data and the rest as training data.We adopted a learning rate schedule where the learning rate was reduced by 20% if the validation dice score had not improved for 10 epochs.

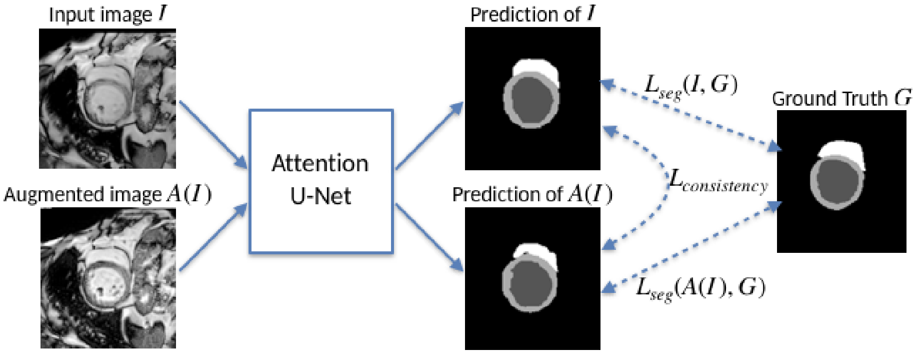

Fig. 3: Proposed training pipeline with both segmentation and consistency losses

2.4 Weighted Ensemble

Ensemble learning with deep neural networks can reduce DL model variance and thus better generalize to unseen data. We observed that our models were more likely to produce mistakes or inconsistent predictions on apical or basal slices. Indeed, the ventricles imaged in these slices usually appear very small in size or have large anatomical variation and noisy ground truth labels. Therefore, we augmented the apical and basal slices four times more than the others using the image augmentation methods described above. We computed a weighted average of the probability maps predicted by models trained with and without such extra augmentation on high-variance image slices. Namely, for apical or basal slices, we assigned a higher weight ratio of 3:2 between models trained with and without extra augmentation on high-variance image slices, respectively. While for other image slices, we assigned a lower weight ratio of 1:2. These weight ratios were experimentally determined by validation.

3 Results

We compared the performance of five segmentation models trained under different augmentation and ensemble settings 1) the baseline attention-UNet model trained without image augmentation (NoAug), 2) with only spatial augmentation(Sptl.Aug), 3) with spatial augmentation and intensity augmentation in the frequency domain (Sptl.Aug+FDA) 4) with spatial augmentation and intensity augmentation using CycleGAN (Sptl.Aug+CycleGan), and 5) ensemble of 3 sets of models trained with frequency domain augmentation (Sptl.Aug+FDA+Ens.). The segmentation predictions generated by these five models were evaluated on the validation dataset from the M&Ms challenge and Table 2 compares the segmentation accuracy for each vendor/center. Spatial augmentation consistently improved segmentation performance for all vendors and centers. Adding frequency domain and CycleGAN augmentation significantly improved the dice scores for the fourth unseen vendor, Vendor D. Compared with a single model, a weighted ensemble consistently improved the performance for all vendors/centers.

Metric	Dice						ASSD					
Vendor	A		B		C	D	A		B		C	D
Center	1	6	2	3	4	5	1	6	2	3	4	5
NoAug	0.867	0.822	0.883	0.928	0.848	0.837	1.303	1.746	0.894	0.358	1.249	1.421
Sptl.Aug	0.874	0.851	0.889	**0.929**	**0.866**	0.828	1.340	1.531	0.780	**0.300**	1.246	1.782
Sptl.Aug+FDA	**0.883**	0.839	**0.899**	0.926	0.862	0.850	**1.133**	1.615	**0.605**	0.421	1.230	1.436
Sptl.Aug+CycleGan	0.863	**0.853**	0.898	0.923	0.863	**0.859**	1.330	**1.359**	0.609	0.363	**1.230**	**1.194**
Sptl.Aug+FDA+Ens.	**0.888**	0.849	**0.904**	**0.932**	**0.868**	**0.862**	**1.016**	1.460	**0.557**	**0.290**	1.246	1.211

Table 1: Dice scores and average surface distance errors (ASSD) of segmentations generated by different models. The dice scores and ASSD were calculated as the averages over the three cardiac structures. The yellow-colored cells represent clinical centers with no annotated training data and the bold numbers in black are the best scores among the models without using ensemble. The bold numbers in blue are when using ensemble achieves the best performance among all models.

Metric	Dice						ASSD					
Vendor	A		B		C	D	A		B		C	D
Center	1	6	2	3	4	5	1	6	2	3	4	5
Sptl.Aug+FDA	0.883	0.839	0.899	0.926	0.862	0.850	1.133	1.615	0.605	0.421	**1.230**	1.436
Sptl.Aug+FDA+Ens.	**0.888**	**0.849**	**0.904**	**0.932**	**0.868**	**0.862**	**1.016**	**1.460**	**0.557**	**0.290**	1.246	**1.211**

Table 2: Dice scores and average surface distance errors (ASSD) of segmentations generated by different models. The dice scores and ASSD were calculated as the averages over the three cardiac structures. The yellow-colored cells represent clinical centers with no annotated training data and the bold numbers in black are the best scores among the models without using ensemble. The bold numbers in blue are when using ensemble achieves the best performance among all models.

Metric	Dice						ASSD					
Vendor	A		B		C	D	A		B		C	D
Center	1	6	2	3	4	5	1	6	2	3	4	5
LV	0.924	0.904	0.900	0.919	0.891	0.903	1.010	1.168	1.168	0.913	1.374	1.184
RV	0.876	0.864	0.872	0.869	0.819	0.882	1.183	1.323	1.201	1.324	2.064	1.347
Myo	0.826	0.839	0.843	0.873	0.817	0.820	0.725	0.821	0.835	0.658	0.974	0.986

Table 3: Dice and ASSD values of the final model evaluated on the test set.

As ensemble learning improved the segmentation accuracy for most clinical centers, we selected the best three sets of models, with each set trained using different training/validation splits. Each set contains two models that were selected based on their performance–both on our own validation split and the M&Ms validation dataset. Specifically, the model ensemble consists of three models trained with FDA, two models trained with CycleGAN augmentation and one model trained with FDA and extra augmented apical and basal slices. Table 3 displays the segmentation accuracy of our final submission evaluated on the M&Ms test data. Overall, our method achieved promising results for most of the vendors and clinical centers, although our method was only trained with annotated data from two vendors and three centers. Figure 4 displays example segmentations of our method on Vendor C and D, which did not have annotated training data. Generally, our method predicted segmentations that closely resemble the ground truths. For some cases, our method tends to make mistakes on apical or basal slices, while generating better predictions in the middle part of the heart.

4 Conclusion

We presented a DL-based automatic cardiac segmentation framework that demonstrated promising performance across multi-scanner and multi-site CMRI scans. We showed that using image augmentation to simulate appearance variations of CMRI data while at the same time constraining the model to generate similar predictions on appearance-augmented images can lead to improved generalization to previously unseen samples from a new vendor or clinical center.

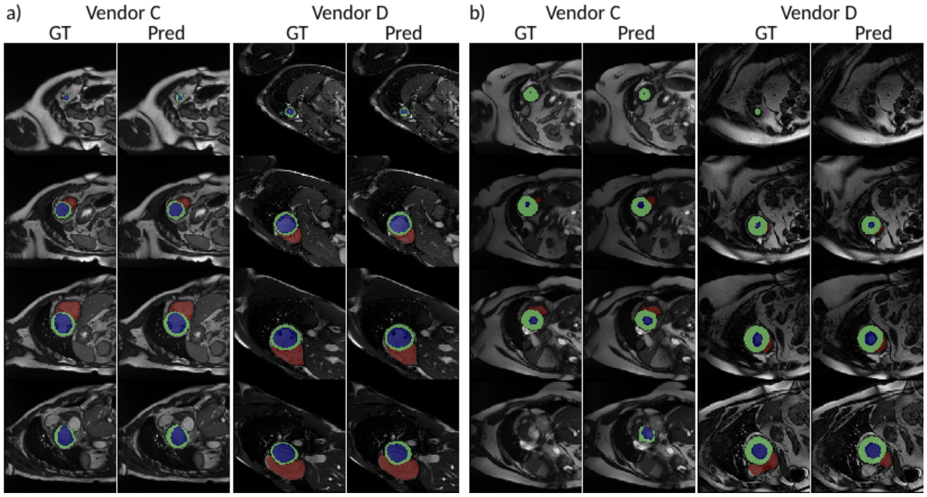

Fig. 4: Examples of predicted segmentations that were better (a) and worse (b) than average compared with predictions by other participants. Segmentations are overlaid with the corresponding image data. Predictions (Preds) are compared with ground truths (GTs) for Vendor C and D. From top row to bottom row are apical, middle and basal slices.

We also explored and compared two effective appearance augmentation techniques, frequency-domain augmentation and CycleGAN based augmentation, that can leverage information from an unlabeled data source to enrich the training dataset. The proposed method can be applied not only to CMRI segmentation but may be readily adapted to other segmentation tasks. Future work will aim to combine the proposed augmentation and consistency training methods with semi-supervised learning to further leverage the unlabeled data.

Acknowledgement This works was supported by the National Science Foundation, Award No. 1663747. The authors declare that the segmentation method implemented for participation in the M&Ms challenge has not used any pretrained models nor additional MRI datasets other than those provided by the organizers.

References

1. Avendi, M., Kheradvar, A., Jafarkhani, H.: A combined deep-learning and deformable-model approach to fully automatic segmentation of the left ventricle in cardiac mri. Med. Image Analy. **30** (12 2015)
2. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. pp. 253–260 (09 2017)

3. Chen, C., Bai, W., Davies, R.H., Bhuva, A.N., Manisty, C.H., Augusto, J.B., Moon, J.C., Aung, N., Lee, A.M., Sanghvi, M.M., Fung, K., Paiva, J.M., Petersen, S.E., Lukaschuk, E., Piechnik, S.K., Neubauer, S., Rueckert, D.: Improving the generalizability of convolutional neural network-based segmentation on cmr images. Frontiers in Cardiovascular Medicine **7**, 105 (2020)
4. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. AAAI **33**, 865–872 (07 2019)
5. Dou, Q., Ouyang, C., Chen, C., Chen, H., Heng, P.A.: Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. pp. 691–697 (07 2018)
6. Isensee, F., Jaeger, P.F., 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. In: STACOM. pp. 120–129. Springer, Cham (2018)
7. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill (2016). https://doi.org/10.23915/distill.00003, http://distill.pub/2016/deconv-checkerboard
8. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M.C.H., Heinrich, M.P., Misawa, K., Mori, K., McDonagh, S.G., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D.: Attention u-net: Learning where to look for the pancreas. ArXiv **abs/1804.03999** (2018)
9. Tao, Q., Yan, W., Wang, Y., Paiman, E., Shamonin, D., Garg, P., Plein, S., Huang, L., Xia, L., Sramko, M., Tintera, J., de Roos, A., van der Geest, R.: Deep learning–based method for fully automatic quantification of left ventricle function from cine mr images: A multivendor, multicenter study. Radiology **290**, 180513 (10 2018)
10. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: CVPR (June 2020)
11. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)