Adversarial and Perceptual Refinement for Compressed Sensing MRI Reconstruction

Maximilian Seitzer[1,2(✉)], Guang Yang[3,4], Jo Schlemper[2], Ozan Oktay[2],
Tobias Würfl[1], Vincent Christlein[1], Tom Wong[3,4], Raad Mohiaddin[3,4],
David Firmin[3,4], Jennifer Keegan[3,4], Daniel Rueckert[2], and Andreas Maier[1]

[1] Pattern Recognition Lab, Friedrich-Alexander-Universität, Erlangen, Germany
`maximilian.seitzer@fau.de`
[2] Biomedical Image Analysis Group, Imperial College, London, UK
[3] National Heart and Lung Institute, Imperial College, London, UK
[4] Cardiovascular Research Centre, Royal Brompton Hospital, London, UK

Abstract. Deep learning approaches have shown promising performance for compressed sensing-based Magnetic Resonance Imaging. While deep neural networks trained with mean squared error (MSE) loss functions can achieve high peak signal to noise ratio, the reconstructed images are often blurry and lack sharp details, especially for higher undersampling rates. Recently, adversarial and perceptual loss functions have been shown to achieve more visually appealing results. However, it remains an open question how to (1) optimally combine these loss functions with the MSE loss function and (2) evaluate such a perceptual enhancement. In this work, we propose a hybrid method, in which a visual refinement component is learnt on top of an MSE loss-based reconstruction network. In addition, we introduce a semantic interpretability score, measuring the visibility of the region of interest in both ground truth and reconstructed images, which allows us to objectively quantify the usefulness of the image quality for image post-processing and analysis. Applied on a large cardiac MRI dataset simulated with 8-fold undersampling, we demonstrate significant improvements ($p < 0.01$) over the state-of-the-art in both a human observer study and the semantic interpretability score.

1 Introduction

Compressed sensing-based Magnetic Resonance Imaging (CS-MRI) is a promising paradigm allowing to accelerate MRI acquisition by reconstructing images from only a fraction of the normally required k-space measurements. Traditionally, sparsity-based methods and their data-driven variants such as dictionary learning [9] have been popular due to their mathematically robust formulation for perfect reconstruction. However, these methods are limited in acceleration

G. Yang and J. Schlemper/D. Rueckert and A. Maier share second/last coauthorship.

factor and also suffer from high computational complexity. More recently, several deep learning-based architectures have been proposed as an attractive alternative for CS-MRI. The advantages of these techniques are their computational efficiency, which enables real-time application, and that they can learn powerful priors directly from the data, which allows higher acceleration rates. The most widely adopted deep learning approach is to perform an end-to-end reconstruction using multi-scale encoding-decoding architectures [7,14]. Alternative approaches carry out the reconstruction in an iterative manner [15], conceptually extending traditional optimization algorithms. Most previous studies focus on exploring the network architecture; however, the optimal loss function to train the network remains an open question.

Recently, as an alternative to the commonly used MSE loss, adversarial [2] and perceptual losses [5] have been proposed for CS-MRI [14]. As these loss functions are designed to improve the visual quality of the reconstructed images, we refer to them as *visual loss functions* in the following. So far, approaches using visual loss functions still rely on an additional MSE loss for successful training of the network. Directly combining all these losses in a joint optimization leads to a suboptimal training process resulting in reconstructions with lower peak signal to noise ratio (PSNR) values. In this work, we propose a two-stage architecture that avoids this problem by separating the reconstruction task from the task of refining the visual quality. Our contributions are the following: (1) we show that the proposed refinement architecture improves visual quality of reconstructions without compromising PSNR much, and (2) we introduce the semantic interpretability score as a new metric to evaluate reconstruction performance, and show that our approach outperforms competing methods on it.

2   Background

Deep Learning-Based CS-MRI Reconstruction. Let $x \in \mathbb{C}^N$ denote a complex-valued MR image of size N to be reconstructed, and let $y \in \mathbb{C}^M$ ($M << N$) represent undersampled k-space measurements obtained by $y = F_u x + \varepsilon$, where F_u is the undersampling Fourier encoding operator and ε is complex Gaussian noise. The linear inversion $x_u = F_u^H y$, also called zero-filled reconstruction, is fundamentally ill-posed and generates an aliased image due to violation of the Nyquist-Shannon sampling theorem. Therefore, it is necessary to add prior knowledge into the reconstruction to constrain the solution space, traditionally formulated as the following optimization problem:

$$\arg \min_{x} \ \mathcal{R}(x) + \lambda \|y - F_u x\|_2^2 \tag{1}$$

Here, $\mathcal{R}$ expresses a regularization term on x (e. g. ℓ_0/ℓ_1-norm for CS-MRI), and λ is a hyper-parameter reflecting the noise level. In deep learning approaches, one learns the inversion mapping directly from the data. However, rather than learning the mapping from Fourier directly to image domain, it is common to formulate this problem as *de-aliasing* the zero-filled reconstructions x_u in the

image domain [12,14]. Let $\mathcal{D}$ be our training dataset of pairs (x, x_u) and $\hat{x} = R(x_u)$ be the image generated by the reconstruction network R. Given $\mathcal{D}$, the network is trained by minimizing the empirical risk $\mathcal{L}(R) = \mathbb{E}_{(x,x_u)\sim\mathcal{D}}\, d(x, \hat{x})$, where d is a distance function measuring the dissimilarities between the reference fully-sampled image and the reconstruction.

For the choice of the reconstruction network R, most previous approaches [7, 14] relied on an encoder-decoder structure (e.g. U-Net [10]), but our preliminary experiments showed that these architectures performed subpar in terms of PSNR. Instead, we use the architecture proposed in [12], as it performed well even for high undersampling rates. This network consists of n_c consecutive de-aliasing blocks, each containing n_d convolutional layers. Each de-aliasing block takes an aliased image $x^{(i)} \in \mathbb{R}^{2N}$ as the input and outputs the de-aliased image $x^{(i+1)} \in \mathbb{R}^{2N}$, with $i \in \{0, \ldots n_c - 1\}$ and $x^0 = x_u = F_u^H y$ being the zero-filled reconstruction. Interleaved between the de-aliasing blocks are data consistency (DC) layers, which enforce that the reconstruction is consistent with the acquired k-space measurements by replacing frequencies of the intermediate image with frequencies retained from the sampling process. This process can be seen as an unrolled iterative reconstruction where de-aliasing blocks and DC layers perform the role of the regularization step and data fidelity step, respectively [12].

Loss Functions for Reconstruction. In deep learning-based approaches to inverse problems, such as MR reconstruction and single image super-resolution, a frequently used loss function [7,15] is the MSE loss $\mathcal{L}_{\mathrm{MSE}}(R) = \mathbb{E}_{(x,x_u)\sim\mathcal{D}}\|x - \hat{x}\|_2^2$. Though networks trained with MSE criterion can achieve high PSNR, the results often lack high frequency image details [1]. Perceptual loss functions [5] are an alternative to the MSE loss. They minimize the distance to the target image in some feature space. A common perceptual loss is the VGG loss $\mathcal{L}_{\mathrm{VGG}}(R) = \mathbb{E}_{(x,x_u)\sim\mathcal{D}}\|f_{\mathrm{VGG}}(x) - f_{\mathrm{VGG}}(\hat{x})\|_2^2$, where f_{VGG} denotes VGG feature maps [13].

Another choice is an adversarial loss based on Generative Adversarial Networks (GANs) [2,3]. A discriminator and a generator network are setup to compete against each other such that the discriminator is trained to differentiate between real and generated samples, whereas the generator is encouraged to deceive the discriminator by producing more realistic samples. For us, the discriminator D learns to differentiate between fully-sampled and reconstructed images, and the reconstruction network, playing the role of the generator, reacts by changing the reconstructions to be more similar to the fully-sampled images. The discriminator loss is then given by $\mathcal{L}_{\mathrm{GAN}}(D) = -\mathbb{E}_{(x,x_u)\sim\mathcal{D}}\log(D(x)) + \log(1 - D(R(x_u)))$. During training, the reconstruction network minimizes $\mathcal{L}_{\mathrm{adv}}(R) = -\mathbb{E}_{(x,x_u)\sim\mathcal{D}}\log(D(R(x_u)))$, which has the effect of pulling the reconstructed images closer towards the distribution of the training data.

Perceptual losses are known to increase textural details [6], but also to introduce high frequency artifacts [2], whereas adversarial losses can produce realistic, high frequency details [6]. As perceptual and adversarial losses complement each other, it is sensible to combine them into a single visual loss $L_{\mathrm{vis}}(R) = \mathcal{L}_{\mathrm{adv}}(R) + \mathcal{L}_{\mathrm{VGG}}(R)$. For MR reconstruction, previous attempts [14]

further combined adversarial and/or perceptual loss with the MSE loss to sta-
bilize the training. This simultaneous optimization yields acceptable solutions,
typically however with low PSNR. We argue this is because the different training
objectives compete with each other, leading to the network ultimately converging
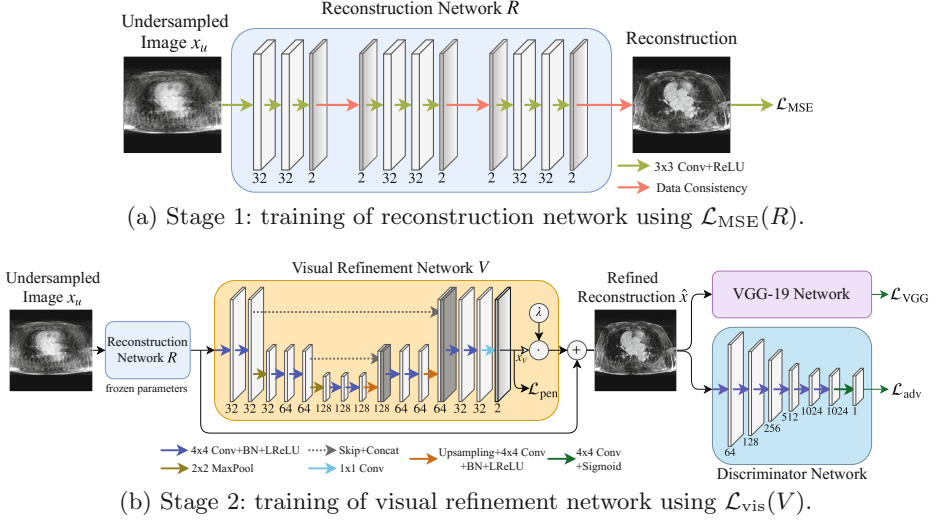to a suboptimal local maximum.

(a) Stage 1: training of reconstruction network using $\mathcal{L}_{\mathrm{MSE}}(R)$.

(b) Stage 2: training of visual refinement network using $\mathcal{L}_{\mathrm{vis}}(V)$.

Fig. 1. Overview of proposed method.

3   Method

The observation above motivates our approach: instead of directly training a
reconstruction network with all loss functions jointly, we use a two-stage proce-
dure, detailed in Fig. 1. In the first stage, the reconstruction network R is trained
with $\mathcal{L}_{\mathrm{MSE}}(R)$. In the second stage, we fix the reconstruction network and train
a visual refinement network V on top of R by optimizing $\mathcal{L}_{\mathrm{vis}}(V)$. The final
reconstruction is then given by $\hat{x} = R(x_u) + V(R(x_u))$, i.e. V learns an additive
mapping which refines the base reconstruction. In this setup, discriminator and
VGG network still receive the full reconstruction $\hat{x}$ as input.

The decoupling of the refinement step from the reconstruction task has sev-
eral benefits. The discriminator begins training by seeing reasonably good recon-
structions, which avoids overfitting it to suboptimal solutions during the train-
ing process. Furthermore, compared to training from scratch, the optimization
is easier as it starts closer to the global optimum. Finally, the visual refinement
step always starts out from the best possible MSE solution achievable with R,
whereas this guarantee is not given when jointly training R with $\mathcal{L}_{\mathrm{MSE}}$ and $\mathcal{L}_{\mathrm{vis}}$.

The choice of the architecture for the visual refinement network is flexible,
and in this work we use a U-Net architecture. Within V, we gate the output

of the network by a trainable scalar λ, which improves the adversarial training dynamics during the early stages of training. If we initialize $\lambda = 0$, the discriminator receives $\hat{x} = R(x_u)$, and the gradient signal to V is forced to zero. This allows the discriminator to initially only learn on clean reconstructions from R, untainted by the randomly initialized output of V. For the refinement network, the impact of less useful gradients is reduced while the discriminator has not yet learned to correctly differentiate between the ground truth (i. e. fully-sampled data) and the reconstructions. We also scale $R(x_u)$ to the range of $(-1, 1)$ before using it as V's input and then scale $\hat{x}$ back to the original range after adding the refinement. In accordance to our goal of reaching high PSNR values, we constrain the output x_V of V (before gating with λ) with an ℓ^1-penalty $\mathcal{L}_{\text{pen}}(V) = \|x_V\|_1$. This guides V to learn the minimal sparse transformation needed to fulfill the visual loss, i. e. to change the MSE-optimal solution only in areas important for visual quality. In practice, this means that our approach yields higher PSNR values compared to joint training, as we show in Sect. 4.

We also utilize a couple of techniques known to stabilize the adversarial learning process. For the discriminator network, we use one-sided label smoothing [11] of 0.1, and an experience replay buffer [8] of size 80 with probability $p = 0.5$ to draw from it. For the refinement network, we add a feature matching loss [11] $\mathcal{L}_{\text{feat}}(V) = \mathbb{E}_{(x,x_u)\sim\mathcal{D}} \frac{1}{N} \sum_{i=1}^{N} \|f_D^{(i)}(x) - f_D^{(i)}(\hat{x})\|_1$, where $f_D^{(i)}$ denotes the i'th of N feature maps of the discriminator. The total loss for V is given by

$$\mathcal{L}(V) = \frac{1}{2}\left(\frac{\mathcal{L}_{\text{adv}}(V)}{M} + \frac{\mathcal{L}_{\text{feat}}(V)}{N}\right) + \frac{\mathcal{L}_{\text{VGG}}(V)}{O} + \alpha\mathcal{L}_{\text{pen}}(V) \qquad (2)$$

with α being the penalty strength, and M, N, O constants set such that $\frac{\mathcal{L}_{\text{adv}}}{M} = \frac{\mathcal{L}_{\text{feat}}}{N} = \frac{\mathcal{L}_{VGG}}{O} = 1$ in the first iteration of training, which amounts to assigning the two adversarial loss terms the same initial importance as $\mathcal{L}_{\text{VGG}}$. The penalty strength α is important for training speed and stability. Choosing α such that $\mathcal{L}_{\text{pen}} \approx 0.1$ in the first training iteration gave us sufficiently good results.

Semantic Interpretability Score. The most commonly used metrics to evaluate reconstruction quality are PSNR and the structural similarity index (SSIM). It has been shown that those two metrics do not necessarily correspond to visual quality for human observers, as e. g. demonstrated by human observer studies in [1,6]. Therefore, PSNR and SSIM alone are not sufficient in the evaluation of image reconstructions. This poses the question on how to evaluate reconstruction quality taking human perception into account. One possibility is to let domain experts (e.g. clinicians and MRI physicists) rate the reconstructions and average the results to form a mean opinion score (MOS). Obtaining opinion scores from expert observers is costly, hence cannot be used during the development of new models. However, if expert-provided segmentation labels are available, we can design a metric indicating how visible the segmented objects are in the reconstructed images, in the following referred to as *semantic interpretability score* (SIS). This metric is motivated by Inception scores [11] in GANs, which tells how well an Inception network can identify objects in generated images.

SIS is defined as the mean Dice overlap between the ground truth segmentation and the segmentation predicted by a pre-trained segmentation network from the reconstructed images. The scores are normalized by the average Dice score on the ground-truth images to obtain a measure of segmentation performance relative to the lower error-bound. We only consider images in which at least one instance of the object class is present, and ignore the background class. We argue that if a pre-trained network is able to produce better segmentations, the regions of interest are better visible (e.g. have clearly defined boundaries) in the images. Implementing SIS requires a segmentation network trained on the same distribution of images as the reconstruction dataset. In practice, the segmentation network is trained on the fully-sampled images used for training the reconstruction method. We trained an off-the-shelf U-Net architecture to segment the left atrium, achieving a Dice score of 0.796 on the ground truth images.

4   Experiments

Datasets. We evaluated our method on 3D late gadolinium enhanced cardiac MRI datasets acquired in 37 patients. We split the 2D axial slices of the 3D volumes into 1248 training images, 312 validation images, and 364 testing images of size 512×512 pixels. For training, we generated random 1D-Gaussian masks keeping 12.5% of raw k-space data, which corresponded to an $8\times$ speed-up. During testing, we randomly generated a mask for each slice, which we kept the same for all evaluated methods.

Training Details and Parameters. For the reconstruction network, we used $n_c = 3$ de-aliasing blocks, and $n_d = 3$ convolutional layers with 32 filters of size 3×3. For the refinement network, we used a U-Net with 32, 64, 128 encoding filters and 64, 32 decoding filters of size 4×4, batch normalization and leaky ReLU with slope 0.1. The discriminator used a PatchGAN [4] architecture with 64, 128, 256, 512, 1024, 1024 filters of size 4×4, and channelwise dropout after the last 3 layers. The VGG loss used the final convolutional feature maps of a VGG-19 network pre-trained on ImageNet. The reconstruction network was trained for 1500 epochs with batch size 20, the refinement network for 200 epochs with batch size 5, both using the Adam optimizer with learning rate 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$. We found that the training is sensitive to the network's initialization. Thus, we chose orthogonal initialization for the refinement network and Gaussian initialization from $\mathcal{N}(0, 0.02)$ for the discriminator.

Evaluation Metrics. We use PSNR and SIS as evaluation metrics. To further evaluate our approach and assess how useful SIS is as a proxy for visual quality, we also asked a domain expert to rate all reconstructed images in which the left atrium anatomy and the atrial scars are visible. The rating ranges from 1 (poor) to 4 (very good), and is based on the overall image quality, the visibility of the atrial scar and occurrence of artifacts. To obtain an unbiased rating, the expert was shown all images from all methods in randomized order.

Results. We compared our approach against three other reconstruction methods: RecNet [12] (i.e. the proposed approach without refinement step), DAGAN[1] [14] using both adversarial and perceptual loss, and DLMRI[2] [9], a dictionary learning based method. No data augmentation was used for any of the methods.

Table 1. Quantitative results for 8-fold undersampling. Highest measures in bold.

Method	PSNR (dB)	MOS	SIS
Ground truth	∞	3.78 ± 0.45	1
RecNet [12]	$\mathbf{32.46} \pm 2.26$	2.75 ± 0.78	0.801
DLMRI [9]	31.45 ± 2.40	1.09 ± 0.29	0.842
DAGAN [14]	28.41 ± 1.91	2.61 ± 0.83	0.812
Proposed method	31.89 ± 2.18	$\mathbf{3.24} \pm 0.63$	$\mathbf{0.941}$

We show the results of our evaluation in Table 1, and a sample reconstruction in Fig. 2. RecNet performed best in terms of PSNR, which is expected as its training objective directly corresponds to this metric, but its reconstructions were over-smoothed. DLMRI had the lowest MOS, with its reconstructions showing heavy oil paint artifacts. DAGAN, combining MSE loss with a visual loss function without any further precautions, suffered from low PSNR. While its reconstructions also looked sharp, they were noisy and often displayed aliasing artifacts, which was reflected in a lower MOS compared to our method. Our proposed approach achieved significantly[3] higher mean opinion score than all other methods, while still maintaining high PSNR. Reconstructions obtained by our method appeared sharper with better contrast. Moreover, our method achieved the highest SIS close to segmentation performance on the ground truth data, which indicated that the segmented objects were clearly visible in the reconstructed images.

These results further demonstrate that PSNR alone is a subpar indicator for reconstruction quality, making our SIS a useful supplement to those metrics. For our method, SIS agreed with the quality score given by the expert user. Somewhat surprising is that the SIS of DLMRI is slightly higher than RecNet and DAGAN although DLMRI has the worst MOS. We conjecture this is because, although DLMRI reconstructed images lack textural details, areas belonging to the same organ have similar intensity values, which helps the segmentation task. While scoring through an expert user is thus still the safest way to evaluate reconstructions, we believe that in conjunction with PSNR, SIS is a helpful tool to quickly judge image quality during the development of new models.

---

[1] https://github.com/nebulaV/DAGAN.
[2] http://www.ifp.illinois.edu/~yoram/DLMRI-Lab/DLMRI.html.
[3] Significance determined by a two-sided paired Wilcoxon signed-rank test at $p < 0.01$.

(a) Zero-filled   (b) DLMRI   (c) DAGAN

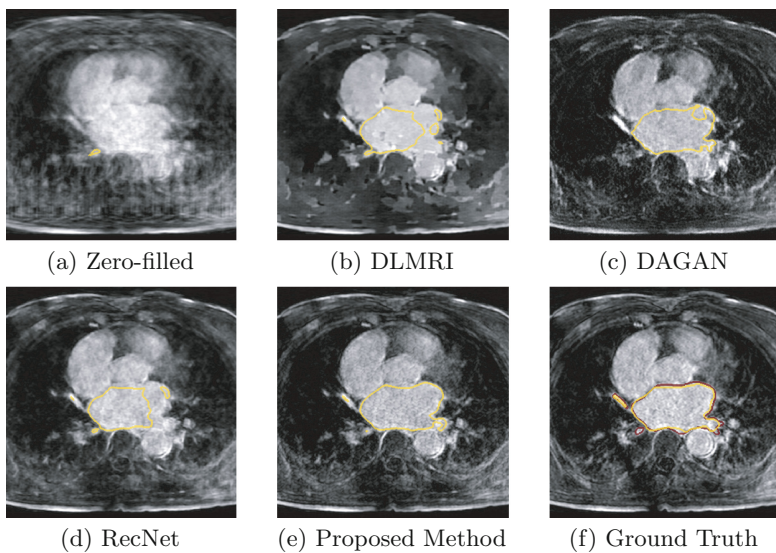(d) RecNet   (e) Proposed Method   (f) Ground Truth

Fig. 2. Qualitative visualization for 8-fold undersampling. Contour of predicted segmentation of left atrium in yellow, contour of ground truth segmentation in red.

5 Conclusion

In this work, we highlighted the inadequacy of previously proposed deep learning based CS-MRI methods using MSE loss functions in direct combination with visual loss functions. We improved on them by proposing a new refinement approach, which incorporates both loss functions in a harmonious way to improve the training stability. We demonstrated that our method can produce high quality reconstructions with large undersampling factors, while keeping higher PSNR values compared to other state-of-the-art methods. We also showed that the reconstruction obtained by our method can provide the best segmentation of the ROIs among all compared methods.

References

1. Dahl, R., et al.: Pixel recursive super resolution. In: ICCV, pp. 5449–5458 (2017)
2. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: NIPS (2016)
3. Goodfellow, I.J., et al.: Generative adversarial nets. In: NIPS (2014)
4. Isola, P., et al.: Image-to-image translation with conditional adversarial networks. In: IEEE CVPR, pp. 5967–5976 (2017)
5. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43

6. Ledig, C., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE CVPR, pp. 105–114 (2017)
7. Lee, D., et al.: Deep residual learning for compressed sensing MRI. In: IEEE 14th International Symposium on Biomedical Imaging, pp. 15–18 (2017)
8. Pfau, D., Vinyals, O.: Connecting generative adversarial networks and actor-critic methods. arXiv preprint arXiv:1610.01945 (2016)
9. Ravishankar, S., Bresler, Y.: MR image reconstruction from highly undersampled k-space data by dictionary learning. IEEE TMI **30**, 1028–1041 (2011)
10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
11. Salimans, T., et al.: Improved Techniques for Training GANs. In: NIPS (2016)
12. Schlemper, J., et al.: A deep cascade of convolutional neural networks for dynamic MR image reconstruction. IEEE TMI (2017)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
14. Yang, G., et al.: DAGAN: deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. IEEE TMI (2018)
15. Yang, Y., et al.: Deep ADMM-net for compressive sensing MRI. In: NIPS (2016)