

Brain Segmentation from k -space with End-to-end Recurrent Attention Network

Qiaoying Huang^{1*}, Xiao Chen², Dimitris Metaxas¹, and Mariappan S. Nadar²

¹ Rutgers University, Department of Computer Science, Piscataway, NJ, USA

² Siemens Healthineers, Digital Technology and Innovation, Princeton, NJ, USA

Abstract. The task of medical image segmentation commonly involves an image reconstruction step to convert acquired raw data to images before any analysis. However, noises, artifacts and loss of information due to the reconstruction process is almost inevitable, which compromises the final performance of segmentation. We present a novel learning framework that performs magnetic resonance brain image segmentation directly from k -space data. The end-to-end framework consists of a unique task-driven attention module that recurrently utilizes intermediate segmentation estimation to facilitate image-domain feature extraction from the raw data, thus closely bridging the reconstruction and the segmentation tasks. In addition, to address the challenge of manual labeling, we introduce a novel workflow to generate labeled training data for segmentation by exploiting imaging modality simulators and digital phantoms. Extensive experimental results show that the proposed method outperforms several state-of-the-art methods.

1 Introduction

Most image segmentation tasks start from existing images. While this might seem self-evident for natural image applications, many medical imaging modalities do not acquire data in the image space. Magnetic Resonance Imaging (MRI), for example, acquires data in the spatial-frequency domain (the so called k -space) and the MR images need to be reconstructed from the k -space data before further analysis. The traditional pipeline of image segmentation treats reconstruction and segmentation as separate tasks. Image noises, residual artifacts and potential loss of information on the imperfect reconstructed images are almost inevitable, even with advanced image reconstruction methods. On the other hand, these algorithms are usually designed to recover images for optimal visual quality to be used by physicians, rather than the “task” quality, namely the “segmentation quality” here. Without the final segmentation quality as a target, the reconstruction algorithm may discard image features that are critical for segmentation but less influential to image quality. Meanwhile, the algorithm may spend most of the resources (e.g. reconstruction time) to recover image features that are less important to segmentation accuracy improvement. It is thus highly desirable to use an end-to-end approach to predict segmentation directly from k -space.

*Work done while intern at Siemens Healthineers.

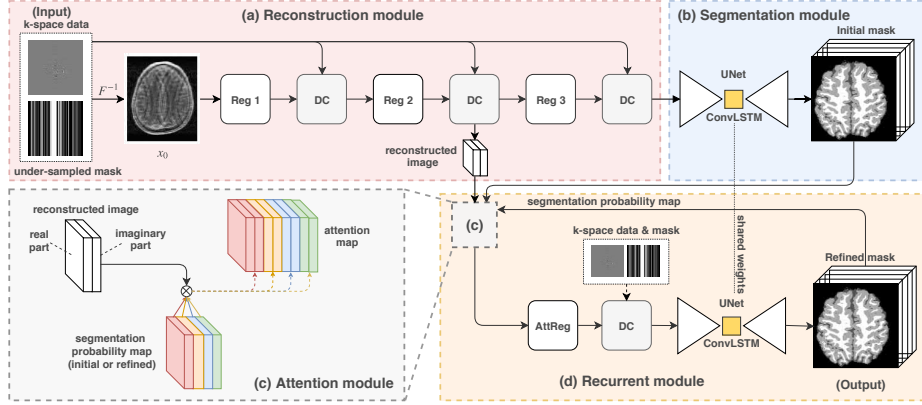


Fig. 1. The proposed model takes the under-sampled k -space data as input and outputs segmentation masks. It consists of (a) reconstruction module, (b) segmentation module, (c) attention module and (d) recurrent module.

Several end-to-end learning frameworks have been proposed for various applications. Caballero *et al.* [2] propose an unsupervised brain segmentation method that treats both reconstruction and segmentation simultaneously using patch-based dictionary sparsity and a Gaussian mixture model. This method is not suitable for complicated scenarios such as the case where there are no clear boundaries between different tissues. Schlemper *et al.* [10] propose two neural networks (LI-net and Syn-net) that predict cardiac segmentation from under-sampled k -space data. The Syn-net uses UNet [8] to map zero-filling image (inverse Fourier transform of under-sampled k -space) directly to segmentation maps. The LI-net exploits latent space features and requires fully-sampled images during training. Huang *et al.* [5] propose a Joint-FR-Net that simply optimizes a joint reconstruction and segmentation network by a combined loss function. The most relevant work to ours is [11] where the authors present a deep neural network architecture termed SegNetMRI. Specifically, the reconstruction and the segmentation sub-networks are pretrained and fine-tuned with shared encoders. The final segmentation depends on the combination of the intermediate segmentation results. The segmentation sub-network in SegNetMRI is trained on reconstructed images, which contain artifacts and noises and may influence the performance of segmentation. Different from SegNetMRI, we propose a new feature sharing method that overcomes the interference with noisy data.

One challenge for end-to-end segmentation learning is the preparation of the training data with ground truth segmentation. Most current studies simulate the raw k -space data from DICOM images using direct Fourier transform. However, realistic k -space data can rarely be recovered from the images alone due to the complex-value nature of MR and common MR post-processing practices that may alter the acquisition. In addition, it is hard to obtain the ground truth segmentation for training. Manual labeling which is performed on images is prone to imperfect reconstruction and human error for small anatomy structures.

We present here an end-to-end architecture: Segmentation with End-to-end Recurrent Attention Network (*SERANet*) featuring a unique recurrent attention module that closely connects the reconstruction and the segmentation tasks. The intermediate segmentation is recurrently exploited as anatomical prior that provides guidance to recover *segmentation-driven* image features from the raw data, which in turn improves segmentation performance. Our contributions include three folds: a) We propose an approach that recurrently performs image segmentation directly from under-sampled k -space data; b) We introduce a novel attention module that guides the network to generate segmentation-driven image features to improve the segmentation performance; c) We present a novel workflow to generate under-sampled k -space data with oracle segmentation maps by exploiting an MRI simulator and digital brain phantoms.

2 Methods

In this section, we begin with a brief introduction to the problem of achieving segmentation directly from k -space. Then we describe our attention module and end-to-end recurrent framework.

2.1 Background

Segmentation from raw data can be generally divided into two subproblems: reconstruction and segmentation. For deep learning based reconstruction on under-sampled k -space: $x = f_{Rec}(y, m)$, where $x \in \mathbb{R}^{2 \times w \times h}$ is the reconstructed image (real and imaginary parts concatenated in the first dimension), y is the under-sampled k -space data and m is the under-sampling mask indicating the position of sampling. f_{Rec} is a deep neural network that can be optimized by reconstruction loss, such as the l_2 loss: $l_2(x, x_{gt}) = \|x - x_{gt}\|_2$. Specifically, as shown in Figure 1 (a), the reconstruction module consists of two basic components. One component is a Data Consistency (*DC* layer) [9] that compensates the difference between the estimated and the measured k -space data. The other component is a regularization block (*Reg* block) that takes as input the zero-filled fast Fourier transform reconstructed image x_0 , or the output from the *DC* layer x_{dc} , and outputs an image x . Cascaded CNN [9] and UNet [4,11] are two popular choices for *Reg* block. Deep learning based iterative reconstruction, motivated by the recent success of compressed sensing in MR image reconstruction, is realized by cascading a series of *Reg* blocks and *DC* layers. The second step is to use the reconstructed image x to predict segmentation probabilities: $s = f_{Seg}(x)$. f_{Seg} is also a deep neural network that can be optimized by segmentation loss, such as the cross entropy loss: $l_{ce}(s, s_{gt}) = -\sum_{s^i \in S} s_{gt}^i \log(s^i)$. As shown in Figure 1 (b), the segmentation module is usually a UNet shape [8].

2.2 Attention module

In end-to-end training, it is beneficial to share information among different tasks. Therefore, we propose an attention module, as shown in Figure 1 (c) that

bridges the gap between reconstruction and segmentation to facilitate learning segmentation-aware features in the image domain. We consider a brain segmentation map as anatomical prior and use it to guide the image reconstruction such that the segmentation information is explicitly utilized to extract segmentation-aware image features from the raw k -space data. We use an attention network to facilitate the segmentation-aware learning. Different from the traditional attention mechanism that only considers two classes in one forward pass [6], we propose to generate multi-class attention maps simultaneously to distinguish features among four classes in the human brain: cerebrospinal fluid (CSF), gray matter (GM), white matter (WM) and background. After one forward pass through the image reconstruction module and the segmentation module, an initial segmentation result is obtained (see Figure 1 part (b)). The segmentation maps $s \in \mathbb{R}^{4 \times w \times h}$ have four tissue maps in separate channels, which are concatenation along the first dimension: $s = s^1 \oplus s^2 \oplus s^3 \oplus s^4$, where s^i indicates the i^{th} class prediction, \oplus represents concatenating along the first dimension. The segmentation map itself is already a probability map. After a softmax layer $\sigma(\cdot)$ that ensures the sum of the four different classes to be 1, the maps can be utilized directly for attention. Each of the four segmentation probability maps are element-wise multiplied with the input image features $x_{t-1} \in \mathbb{R}^{2 \times w \times h}$ to generate new features $x_t \in \mathbb{R}^{8 \times w \times h}$:

$$x_t = (s_{t-1}^1 \odot x_{t-1}) \oplus (s_{t-1}^2 \odot x_{t-1}) \oplus (s_{t-1}^3 \odot x_{t-1}) \oplus (s_{t-1}^4 \odot x_{t-1}), \quad (1)$$

where subscript t represents the t^{th} intermediate result (explained in the next section). As shown in Figure 1 (c), the new image features x_t go through one *Reg* block for attention features (referred as *AttReg*) and one *DC* layer, in order to extract image features. The difference between *AttReg* and *Reg* in the reconstruction module is the input channel size: *AttReg* has 8 instead of 2 channels. The output of the attention-assisted image feature extraction is then fed to the same segmentation module with shared weights to generate a new segmentation estimation s_t . Formally, s_t can be expressed as follows.

$$s_t = (f_{Seg} \circ f_{DC} \circ f_{AttReg})(x_t), \quad (2)$$

where \circ denotes function composition and s_t is the new segmentation estimation. By explicitly utilizing intermediate segmentation results for reconstruction, or more precisely image feature extraction, segmentation-driven features will be generated, which in turn improves segmentation performance during training with back-propagation algorithm. It can be seen from Figure 3 that clear boundaries are generated using the proposed *SERANet* from under-sampled k -space data, while the ground truth reconstruction from fully-sampled k -space data contains noise.

2.3 Recurrent framework

We treat segmentation feature learning as a recurrent procedures that final result is achieved by iterating the attention module several times. Formally, given

Algorithm 1: *SERANet*: Segmentation with Recurrent Attention Network

input : under-sampled k -space data y , under-sampling mask m , N , T

1 $x_0^{(N-1)}, x_0^{(N)} \leftarrow f_{Rec}(y, m)$; // initial reconstruction feature x_0

2 $s_0 \leftarrow f_{Seg}(x_0^{(N)})$; // initial segmentation result s_0

3 **if** $t \leq T$ **then**

4 $x_t \leftarrow f_{DC} \circ f_{AttReg}(x_0^{(N-1)}, s_{t-1})$; // Attention module

5 $s_t \leftarrow f_{Seg}(x_t)$; // Recurrent segmentation

output: s_T

under-sampled k -space data y and mask m , *SERANet* learns to segment the brain in T iterations. N is the number of *Reg* blocks and *DC* layers. As described in Algorithm 1, line 1 and 2 generate the initial reconstructed image x_0 and brain tissue map s_0 . Then line 3 to 5 represent a recurrent segmentation-aware reconstruction and segmentation process. The attention module f_{AttReg} takes the initial reconstruction feature x_0^{N-1} (feature from the $N - 1$ reconstruction block) and segmentation probability maps s_{t-1} as input and generates new image x_t , as illustrated in Figure 1 (c) and (d). To capture and memorize the spatial information at different recurrences, a ConvLSTM layer [12] is integrated into the UNet for segmentation. The objective function of the whole model is defined as $l_{ce}(s_T, s_{gt})$, where s_T denotes the output of the final iteration. By doing so, the reconstruction module in our method does not see nor need any ground truth reconstruction image during training. The recovered image content is guided by the segmentation error solely, which suffices the aim to recover image domain features from the raw data that best suits the segmentation task, rather than the conventional reconstruction task. The usage of “reconstruction” to name the module is just for conceptual simplicity.

2.4 Generate k -space data with oracle segmentation maps

We propose here a novel method to generate realistic k -space data with ground truth segmentation map. Specifically, a widely utilized MRI scanner simulator MRiLab [7][1] is adopted to provide a realistic virtual MR scanning environment, which includes scanner system imperfection, MR acquisition pattern and MR data generation and recording. We use publicly available digital brain phantoms from BrainWeb [3] as the object “scanned” in the MRI simulator. Each brain is consisted of 12 tissue types with known spatial distributions. Each tissue type has a unique set of values of MR physical parameters such as T1 and T2 that are needed for the MR scan simulation. Fully-sampled k -space data is then simulated by scanning the digital brain in MRiLab. Under-sampling is performed retrospectively by keeping a subset of the full-sampled data. To mimic realistic MR scanning, white Gaussian noises are added to the k -space data at multiple levels. The network never sees the fully-sampled data. The spatial distributions of the tissues are the oracle segmentation maps.

Table 1. Dice’s score of *SERANet* trained with different losses

Loss	10% noise				20% noise			
	CSF	GM	WM	Aver.	CSF	GM	WM	Aver.
$l_{ce}(s_T) + l_2(x_T)$	0.8048	0.8841	0.8518	0.8469	0.7995	0.8751	0.8092	0.8279
$\sum_{t=0}^T l_{ce}(s_t)$	0.8513	0.9082	0.8796	0.8797	0.8041	0.8733	0.8283	0.8352
$l_{ce}(s_T)$	0.8482	0.9102	0.8814	0.8799	0.8083	0.8762	0.8415	0.8423

3 Experiments

Implementation Details Total 20 healthy digital 3D brain volumes are utilized. The networks are trained on 2D axial slices. 969 slices of 17 brains are used for training and the rest 171 slices from 3 brains are reserved for testing only. Each digital brain is scanned by a spin echo sequence with Cartesian readout. Average TE = 80 ms and TR = 3 s are used and 5% variation of both TE and TR values are introduced for varying MR contrasts. Each slice has the corresponding tissue segmentation mask from BrainWeb. All slices have a unified size of 180×216 with 1 mm isotropic resolution. We use a zero-mean Gaussian distribution with a densely sampled k -space center to realize a pseudo-random under-sampling pattern, where 30% phase encoding lines are maintained with 16 center k -space lines. All k -space data are added with additional 10% and 20% white Gaussian noise. All models are implemented in Pytorch and trained on NVIDIA TITAN Xp. Hyperparameters are set as: a learning rate of 10^{-4} with decreasing rate of 0.5 for every 20 epochs, 50 maximum epochs, batch size of 12. Adam optimizer is used in training all the networks. We adopt Dice’s score as evaluation metric in all experiments. For the recurrent steps T , the segmentation performance of our model in terms of Dice’s score has converged after two recurrences. So we empirically set $T = 2$.

Effect of different losses We provide a comparison of training *SERANet* with different losses in Table 1. We observe that *SERANet* constrained by the reconstruction loss (l_2) performs the worst, as shown in the first row of Table 1. This may seem surprising at first but it actually verifies that reconstruction from the images with noise and artifacts may compromise segmentation results. The model that is optimized solely using segmentation loss (l_{ce}) on the final segmentation estimation s_T achieves the best result. These results demonstrate two key advantages of our proposed method. First, due to the efficiency of the attention module, *SERANet* automatically learns image feature that benefits the segmentation performance, without constraints on the reconstructed image. Second, our method does not require cumbersome tweaking of loss weights between reconstruction and segmentation tasks.

Effect of attention module We design two different baselines without the attention module: one is a *Two-step* model that contains separate reconstruction and segmentation modules, which are trained separately with reconstruction and segmentation losses. The other is a *Joint* model, which also contains these two

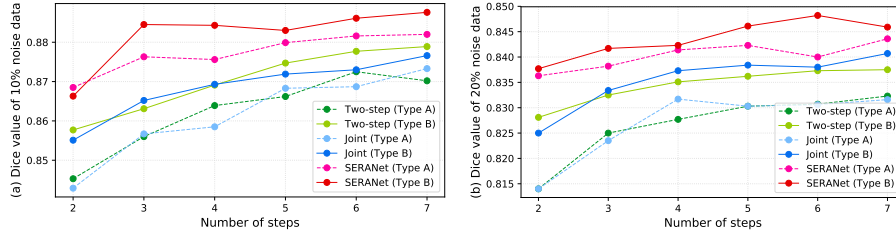


Fig. 2. Dice score as a function of the number of reconstruction blocks.

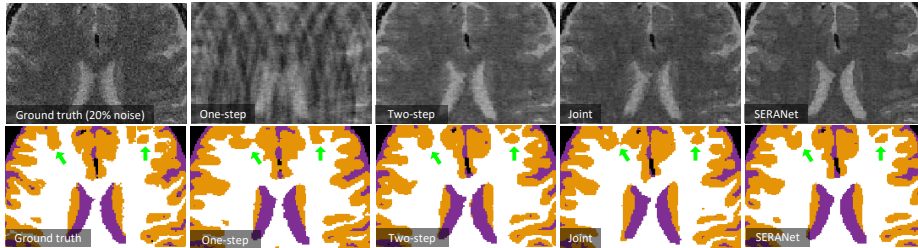


Fig. 3. Segmentation result of *SERANet* and the compared models. CSF, GM and WM parts in each brain are colorized with purple, orange and white, respectively.

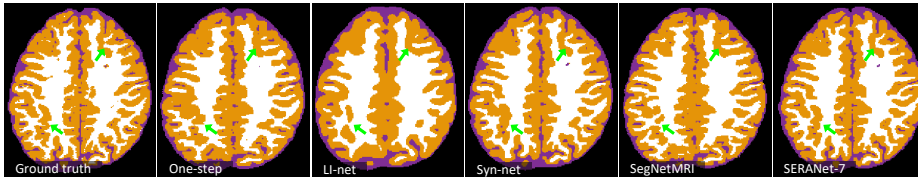
modules but is trained together with only segmentation loss on the final output. In order to evaluate the robustness under different settings, we compare performances of *Two-step*, *Joint* and *SERANet* with different number of reconstruction blocks. We also implement two different reconstruction blocks: one using cascading CNN (*Type A*) and the other using auto-encoder (*Type B*). The Dice's score against number of reconstruction block is plotted in Figure 2. We observe, as expected, that with more cascading reconstruction blocks the segmentation performances of all tested methods improve. However, *SERANet* outperforms the others in all settings which proves the benefit of using the attention module.

For visualization, we also train an *One-step* model that takes input as zero-filling images (inverse Fourier transform of under-sampled k -space data) and outputs the segmentation maps. We visualize segmentation results and reconstructed images of models trained by 20% noise data in Figure 3. Our method *SERANet* predicts more accurate anatomical segmentation details and clearer image contrasts compared to *Two-step* and *Joint*. This shows *SERANet* overcomes the interference with noisy input by using the attention module.

Comparisons to State-of-the-Art We provide qualitative and quantitative comparisons to three state-of-the-art algorithms: *LI-net*[10] *Syn-net*[10], and *SegNetMRI*[11]. We also compare to the *One-step* model. We list the performance of *SERANet-7* that with 7 reconstruction blocks and *SERANet-2* that with 2 reconstruction blocks. The results of all methods are reported in Table 2. We also list whether the method is pretrained and what loss the method uses to optimize in column 2 and 3, respectively. For *LI-net* and *Syn-net*, since they perform segmentation from fully-sampled data as a warm start, we consider

Table 2. Comparisons to State-of-the-Art

Method	Pretrain	Loss	10% noise				20% noise			
			CSF	WM	GM	Aver.	CSF	WM	GM	Aver.
One-step	No	l_{ce}	0.7677	0.8334	0.7900	0.7970	0.7600	0.8324	0.7911	0.7945
LI-net [10]	Yes	l_{ce}	0.6849	0.7576	0.7558	0.7328	0.6686	0.7276	0.7282	0.7081
Syn-net [10]	Yes	$l_{ce}+l_2$	0.7558	0.8256	0.7961	0.7925	0.7307	0.8095	0.7808	0.7737
SegNetMRI [11]	Yes	$l_{ce}+l_2$	0.8210	0.8905	0.8575	0.8563	0.7817	0.8472	0.7728	0.8006
SERANet-2	No	l_{ce}	0.8344	0.8977	0.8669	0.8663	0.8053	0.8706	0.8373	0.8377
SERANet-7	No	l_{ce}	0.8548	0.9175	0.8905	0.8876	0.8122	0.8798	0.8457	0.8459

**Fig. 4.** Segmentation result with different approaches.

this as a pretraining technique. We observe that *SERANet-2* and *SERANet-7* consistently outperform the three state-of-the-art approaches for both 10% and 20% noises. Additionally, the Dice’s scores drop more for the *SegNetMRI* when noise level increases compared to *SERANet*, which may be due to the fact that *SegNetMRI* contains information from the noisy ground truth images. Example segmentation results are shown in Figure 4. Improvements of *SERANet-7* on detailed anatomy structure are highlighted by the green arrows.

4 Conclusion

In this paper, we propose a novel end-to-end approach *SERANet* for MR brain segmentation, which performs segmentation directly on under-sampled k -space data via a segmentation-aware attention mechanism. Moreover, we design a training data generation workflow to simulate realistic MR scans on digital brain phantoms with ground truth segmentation maps. Extensive experiments are conducted and the results demonstrate the effectiveness and the superior performance of our model compared to the state-of-the-art methods.

5 Additional Implementation Details

In this paper, we consider a segmentation problem with four brain segmentation masks: 0-Background, 1-CSF, 2-Gray Matter and 3-White Matter. As mentioned in the paper, these four masks are adapted from original 11 brain tissues. The reason to use the four masks is that CSF, Gray Matter and White Matter cover most parts of the brain, as shown in Figure 5 (a) and (b). The other eight tissues, such as vessels, skulls and skins, are grouped as Background mask.

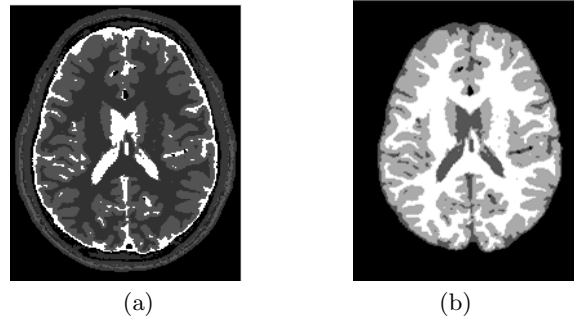


Fig. 5. (a) Original 11 tissues segment- masks. (b) Selected 4 tissues segment- masks.

5.1 Data generation

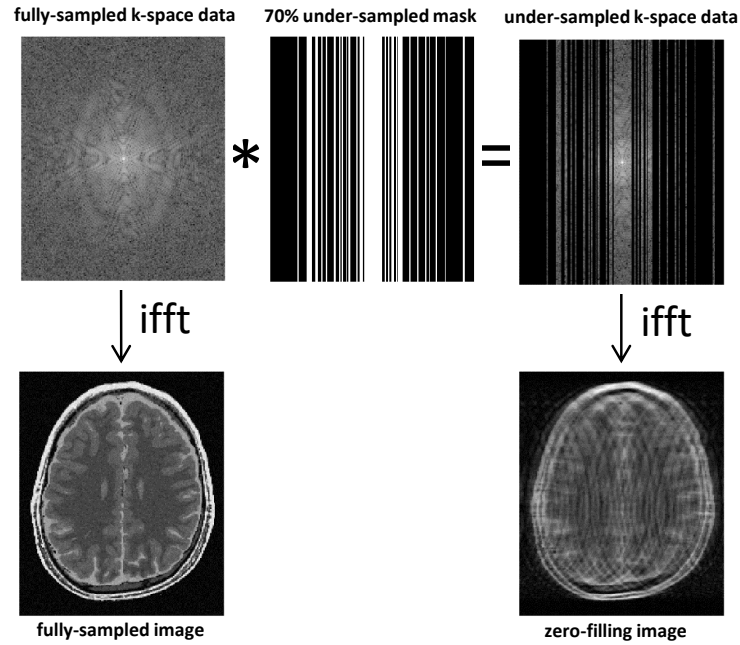


Fig. 6. The process to generate under-sampled k-space data.

The detailed data generation process is illustrated in Figure 6. Given fully-sampled k-space data (top-left), we first randomly generate under-sampled mask, e.g. with 70% sampling rate (top-middle). Under-sampled k-space data (top-right) is obtained by employing the mask on the fully-sampled k-space data. Then, fully-sampled image (bottom-left) can be generated from fully-sampled k-

space data via inverse fast Fourier transform. Note that the fully-sampled image is used as ground truth by some existing algorithms, however, it may contain noise and comprise the segmentation performance. Similarly, zero-filling image (bottom-right) is generated from under-sampled k-space data via inverse Fourier transform, and is taken as the input in all models.

5.2 Network architectures

In our SERANet, we implement two types of regularization blocks: cascaded CNN (Type A) [9] and auto-encoder (Type B) [11], which are two popular choices for image reconstruction using deep learning. Their architectures are respectively shown in Figure 7 (a) and 7 (b). We also show the architecture of the UNet adopted in the paper for the segmentation module in Figure 7 (c).

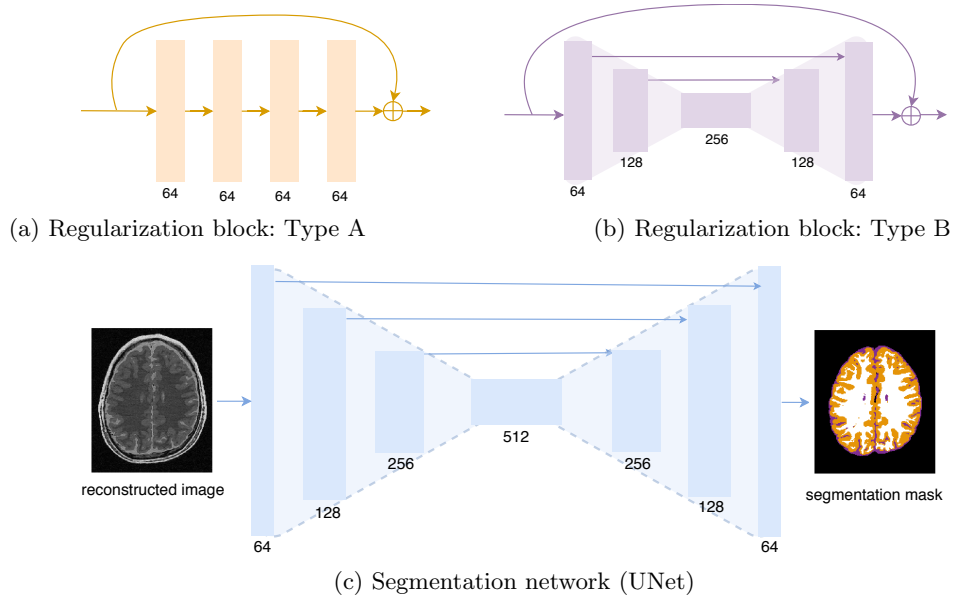


Fig. 7. Architectures of two regularization blocks and one segmentation network utilized in the paper.

6 Additional Quantitative Result

In this section, we provide additional quantitative results. We demonstrate the comparison results of our SERANet and other approaches on data with 10% (Figure 6) white Gaussian noise and data with 20% noise (Figure 6). For LI-net and Syn-net, we only show their segmentation results since they bypassed the

reconstruction step. For our SERANet, we present the results of SERANet-2, SERANet-4 and SERANet-7 here.

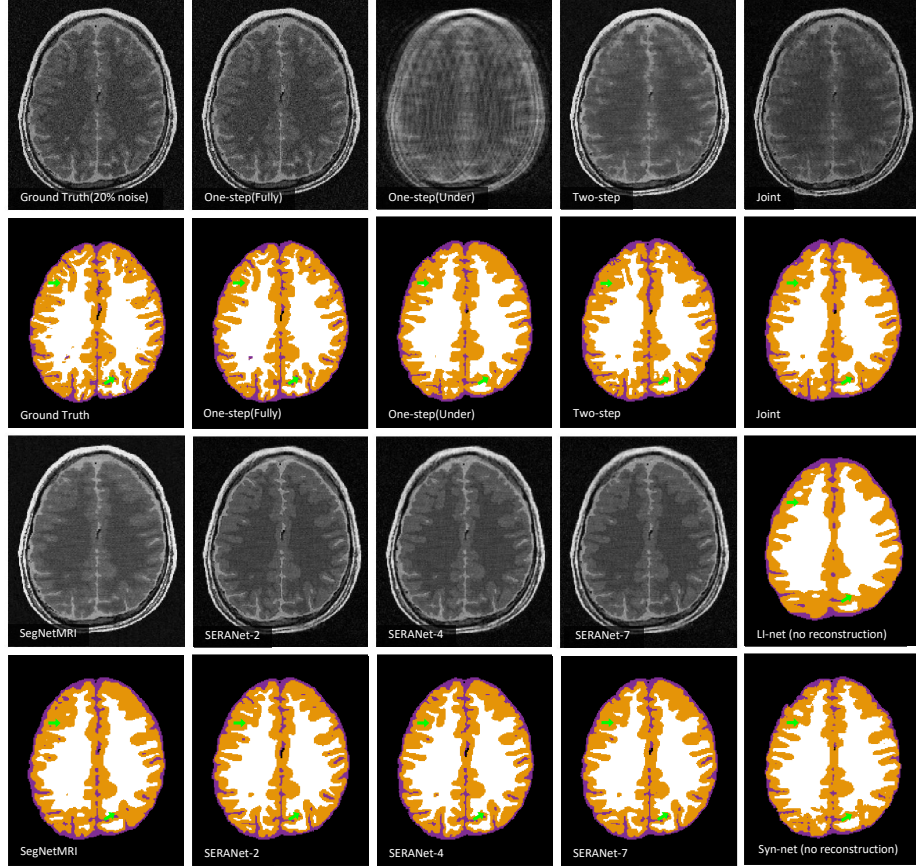


Fig. 8. Segmentation performance on input data with 10% white Gaussian noise. Since LI-net and Syn-net bypass the reconstruction, we only show their segmentation results here.

References

1. Mrilab: A numerical mri simulator. <http://mrilab.sourceforge.net/> (2018)
2. Caballero, J., Bai, W., Price, A.N., Rueckert, D., Hajnal, J.V.: Application-driven mri: Joint reconstruction and segmentation from undersampled mri data. In: MIC-CAI. pp. 106–113. Springer (2014)
3. Chris A. Cocosco, Vasken Kollokian, R.K.S.K.A.C.E.: Brainweb: Online interface to a 3d mri simulated brain database (1997), <http://brainweb.bic.mni.mcgill.ca/brainweb/>

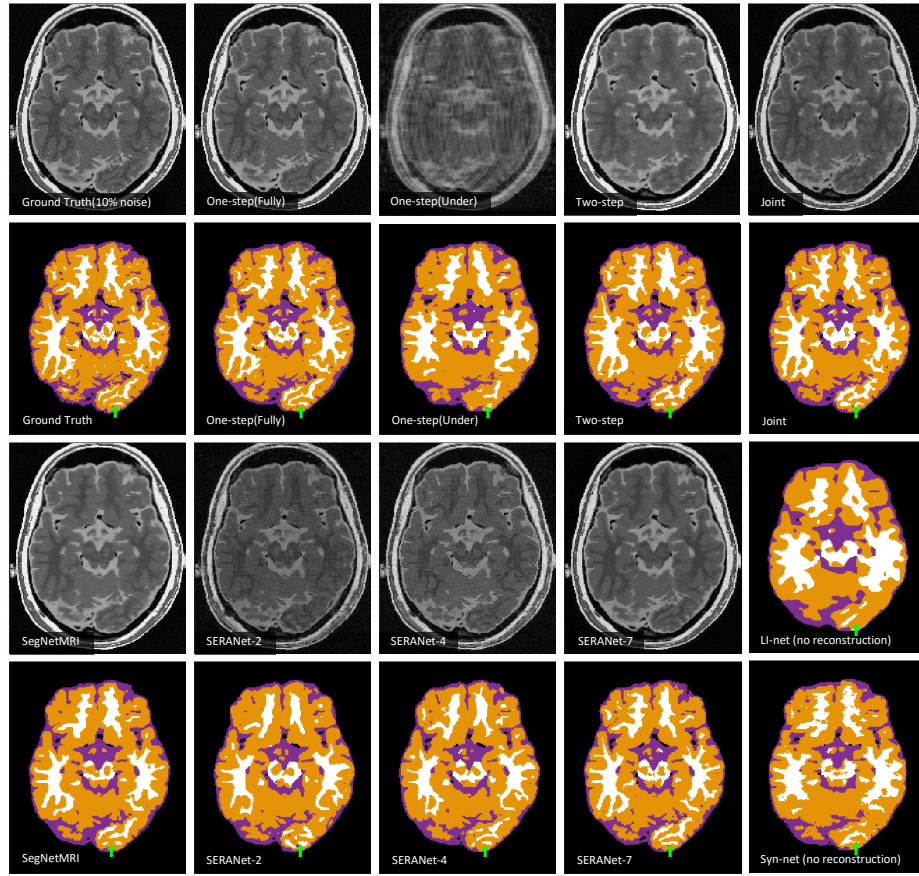


Fig. 9. Segmentation performance on input data with 20% white Gaussian noise. Since LI-net and Syn-net bypass the reconstruction, we only show their segmentation results here.

4. Huang, Q., Yang, D., Wu, P., Qu, H., Yi, J., Metaxas, D.: Mri reconstruction via cascaded channel-wise attention network. In: ISBI. pp. 1622–1626. IEEE (2019)
5. Huang, Q., Yang, D., Yi, J., Axel, L., Metaxas, D.: Fr-net: Joint reconstruction and segmentation in compressed sensing cardiac mri. In: International Conference on Functional Imaging and Modeling of the Heart. pp. 352–360. Springer (2019)
6. Li, K., Wu, Z., Peng, K.C., Ernst, J., Fu, Y.: Tell me where to look: Guided attention inference network. In: CVPR. pp. 9215–9223 (2018)
7. Liu, F., Velikina, J.V., Block, W.F., Kijowski, R., Samsonov, A.A.: Fast realistic mri simulations based on generalized multi-pool exchange tissue model. *IEEE transactions on medical imaging* **36**(2), 527–537 (2017)
8. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
9. Schlemper, J., Caballero, J., Hajnal, J.V., Price, A., Rueckert, D.: A deep cascade of convolutional neural networks for mr image reconstruction. In: IPMI. pp. 647–

658. Springer (2017)
10. Schlemper, J., Oktay, O., Bai, W., Castro, D.C., Duan, J., Qin, C., Hajnal, J.V., Rueckert, D.: Cardiac mr segmentation from undersampled k-space using deep latent representation learning. In: MICCAI. pp. 259–267. Springer (2018)
 11. Sun, L., Fan, Z., Huang, Y., Ding, X., Paisley, J.: Joint CS-MRI reconstruction and segmentation with a unified deep network. In: IPMI. Springer (2019)
 12. Xingjian, S., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c.: Convolutional lstm network: A machine learning approach for precipitation nowcasting. In: NIPS. pp. 802–810 (2015)