DEEP LEARNING BASED METHODS FOR ULTRASOUND IMAGE SEGMENTATION AND MAGNETIC RESONANCE IMAGE RECONSTRUCTION

by

Puyang Wang

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

March, 2021

Abstract

In this thesis, we develop various deep learning (DL) based approaches to address two medical image analysis problems. In the first problem, we focus on computer assisted orthopedic surgery (CAOS) applications that use ultrasound as intra-operative imaging modality. This problem requires an automatic and real-time algorithm to detect and segment bone surfaces and shadows in order to provide guidance for the orthopedic surgeon to a standardized diagnostic viewing plane with minimal artifacts. Due to the limitation of relatively small datasets and image differences from multiple ultrasound machines, we develop DL-based frameworks that leverage a local phase filtering technique and integrate it into the DL framework, thus improving the robustness.

Finally, we propose a fast and accurate Magnetic Resonance Imaging (MRI) image reconstruction framework using a novel Convolutional Recurrent Neural Network (CRNN). Extensive experiments and evaluation on knee and brain datasets have shown its outstanding results compared to the traditional compressed sensing and other DL-based methods. Furthermore, we extend this method to enable multi sequence-reconstruction where T2-weighted MRI image can provide guidance and improvement to the reconstruction of amid proton transfer-weighted MRI image.

Thesis Readers

Vishal M. Patel (Primary Advisor), Assistant Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Rama Chellappa, Bloomberg Distinguished Professor
Department of Electrical and Computer Engineering
Johns Hopkins University

Acknowledgments

Pursuing a doctoral degree over a period of five years took a lot of effort and dedication. I would not have made it without the support and encouragement from many people around me.

First and foremost I would like to thank my advisor Vishal Patel for guiding me through this journey and all the helps he provided. I'm very lucky and grateful to have an advisor who is not only friendly and patient but also professional and knowledgeable as Dr. Patel is. His style of advising inspired me a lot during my entire graduate study and is surprisingly effective. He has been a very thoughtful and understanding when I had to go through various milestones and all the highs and lows in my academic life. Thank you Vishal for all the support and guidance given through the years.

I wish to express my gratitude to Prof. Alan Yuille, Prof. Shanshan Jiang, Prof. Trac Tran and Prof. Shinji Watanabe who officiated in my Graduate Board Examination panel. I would like to pay a special thanks to Prof. Rama Chellappa, Prof. Carlos Castillo and Prof. Ilker Hacihaliloglu for officiating

ACKNOWLEDGMENTS

in my dissertation committee, acting as dissertation readers, and forgiving me valuable feedback throughout the process.

Finally, I would like to thank my family for their continued support after all these years. I would like to thank my Farther Zuo Wang who always supported me through my years abroad. Especially, I would like to thank my mother Yun Lei who has been the pillar of my life over the years.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this thesis, we address two fundamental problems in medical image analysis. In particular, we address problems regarding ultrasound (US) image segmentation and Magnetic Resonance (MR) image reconstruction. We develop various deep learning-based solutions to address these problems. In this chapter, we first give a brief background of the problems being addressed.

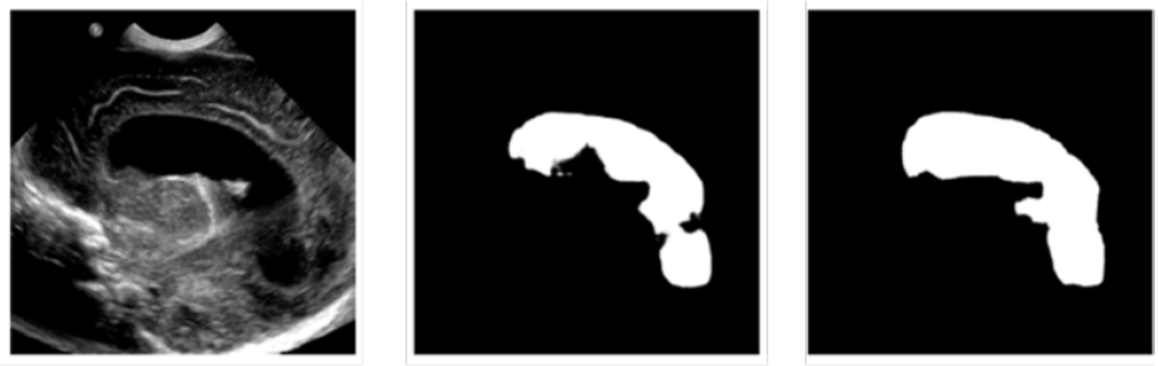
1.1 Ultrasound Image Segmentation

1.1.1 Ultrasound Brain Ventricle Segmentation

Very low birth weight ($< 1,500g$) premature babies account for 1.4 percent of all births in the United States. Of those, more than 16,000 babies each year will develop intraventricular hemorrhage (IVH) [7]. These hemorrhages result in ventricle dilation, which can lead to serious brain damage if not properly treated. Monitoring of ventricle volume change in neonates is clinically important in order to determine the correct intervention. Two-dimensional (2D) US is currently the main imaging modality used in the diagnosis and monitoring of IVH. However, irregular shape deformation of ventricles, high levels of noise and various imaging artifacts present in the acquired ultrasound data results in the inability to localize the site and extent of brain injury, or to predict neurologic outcomes in identifying IVH from US data. Due to these difficulties, quantitative assessment of anatomical information is mostly performed manually or using semi-automated methods [8]. In [9], a fully automated atlas-based segmentation pipeline was developed for segmenting 3D volumetric US data. Validation results performed on 30 3D US scans achieved a mean Dice similarity coefficient (DSC) and maximum absolute distance of 76.5% and 1 mm, respectively. The reported computation time for segmenting a single 3D volume was 54 mins [9].

In recent years, deep learning-based methods have shown to produce state-of-the-art results in many computer vision and medical image analysis tasks [10], [11], [12]. Inspired by [13], various Convolutional Neural Network (CNN)-based encoder-decoder networks have been proposed for different computer vision tasks in the literature. A typical encoder-decoder structure starts with an encoder network which decreases spatial resolution while learning a high-dimensional representation, followed by a decoder that recovers the original input resolution and outputs low-dimensional predictions. In particular for biomedical image segmentation, one such network, called SegNet was recently proposed in [1] and has been widely used for medical image segmentation. A conditional generative adversarial network-based method called pix2pix was proposed in [3]. U-Net is another popular network that has been widely used in the medical imaging community for segmentation [2]. The advantage of U-net structure comes from the symmetric contracting and expanding path which is capable of leveraging contextual information of different scales.

However, it often lacks the ability to capture complex context information in images due to its relative shallow network design. This negatively impact U-nets segmentation performance of objects with blurry boundaries and low contrast compared to background. For example, brain ventricles in US images can suffer from artifacts and noise in the ventricle region which may result into unsatisfied segmentation if a standard U-net is used. Thus, a more robust and accurate segmentation network that can tackle the problem of segmenting low contrast objects in US images needs to be proposed.



(a) US image (b) U-net segmentation (c) Ground truth

Figure 1.1: Ultrasound brain ventricle segmentation.

1.1.2 Ultrasound-guided Computer Assisted Orthopedic Surgery

In order to provide a radiation-free, real-time, cost effective imaging alternative for intra-operative fluoroscopy ultrasound has been incorporated into various computer assisted orthopedic surgery (CAOS) procedures such as percutaneous scaphoid fixation and pelvic ring facture surgery [14]. US-based guidance systems for non-surgical procedures such as epidural anesthesia and/or spinal blocks have also been developed [15, 16].

Nonetheless, problems such as high levels of noise, imaging artifacts, limited field of view and bone boundaries appearing several millimeters (mm) in thickness have hindered the wide spread adaptability of US-guided CAOS systems. These difficulties prohibited the use of US as a stand-alone intra-operative imaging modality and focus was given on developing automated bone segmentation, enhancement [17–20] and intra-operative US-based image registration [21] methods. Our main focus is the development of an US-based CAOS system where automatically extracted bone surfaces are used for continuous real-time guidance. Therefore, complete, accurate and robust segmentation of bone surfaces is of paramount importance.

Early work for segmenting bone surfaces from US, utilized image intensity and gradient information [14]. However, these methods are not robust for processing low contrast bone surfaces and are affected by acquisition settings, image artifacts, and body mass index (BMI) of the patient. To address this challenge, local phase-based bone surface enhancement methods have been proposed [14]. Local phase information is extracted by filtering the B-mode US data in frequency domain using bandpass quadrature filters. Most common filters include Log-Gabor filter, monogenic filter, and local phase tensor filter [14]. The enhanced bone surfaces were localized using post-processing methods such as dynamic programming [14] or simple bottom up ray-casting. Although, phase-based approaches are more robust to image artifacts and low contrast bone surfaces, successful segmentation depends on the robustness of the post-processing method used. Furthermore, local phase-based methods require the optimization of band-pass quadrature filter parameters which requires large processing time making the methods not suitable for real-time processing [17, 22].

Due to some recent advances in deep learning, deep learning-based methods have shown to provide much better bone segmentation given enough training data. In [6], a modified version of U-net was used for localizing vertebra bone surfaces. However, low-quality bone surfaces were excluded from the validation and testing procedure. Recently, various filtered feature guided methods [18–20] were proposed. These methods propose to incorporate filtered features, such as local phase tensor image or enhanced bone shadow image, into to the CNN by either using early feature fusion or late fusion operations. Particularly, the multi-feature guided CNN in [18] takes US image, local phase tensor image, bone shadow enhanced image, and local phase image as concatenated input. And it demonstrated the state-of-the-art performance when testing on different US machines. However it was shown that average computational time for additional input local phase and shadow enhancement was 2 seconds, making real-time application impossible. In summary, despite the fact that methods based on deep learning produce robust and accurate results, the success rate is dependent on either: (1) Consistent and high quality US scans used for training and testing [6], or (2) additional computation time required for image filtering [18].

High quality US data refers to US images where bone surfaces appear sharp with high intensity followed by intensity dropout representing the bone shadow interface. During the data collection, quality of the US images and machine plays an important role in acquiring high quality US data. Most of the clinically available US machines are equipped with high quality transducers. However, this is not valid with the point-of-care portable low cost transducers. Furthermore, manual operation of the transducer introduces additional difficulties during data collection since a single-degree deviation angle by the operator can reduce the signal strength by 50% [14]. Most of previously proposed deep learning methods are trained on high-quality US data from single US machine. If the acquisition involves low bone surface contrast and image quality, the ability of complete and accurate segmentation always decreases dramatically.

Another important task for US-guided CAOS is to detect bone shadow regions. In the context of bone imaging, using US, bone boundaries have the highest intensity in the image followed by a region with low intensity values denoted as the shadow region. Shadow region is the result of a high acoustic impedance mismatch between the soft tissue and the bone boundary resulting in most of the US signal being reflected back to the transducer surface. This information is important that it represents unknown regions in the image. In order to improve the accuracy and robustness of bone segmentation, several groups have incorporated bone shadow information into their framework [14]. Bone shadow information can also be used in order to guide the orthopedic surgeon to a standardized diagnostic viewing plane with minimal artifacts.

In Fig.1.2, we demonstrate the mentioned notion using a B-mode US im-

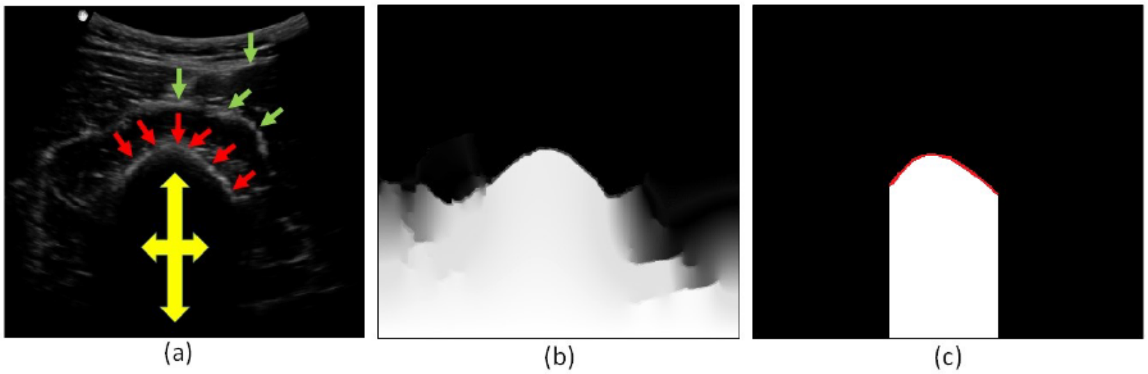


Figure 1.2: (a) B-mode US image of in vivo femur. Thick yellow arrows point to the bone shadow region. Red arrows point to the bone surface response. Green arrows point to soft tissue interface resembling bone response. (b) Bone shadow enhanced image obtained using [4]. (c) Gold standard bone shadow and surface obtained by expert manual segmentation. In both (a) and (c) regions corresponding to soft tissue are displayed with black color coding, regions corresponding to bone shadow are displayed with gray/white color coding and bone surface are displayed in red.

age of in-vivo femur. In summary, we aim to detect bone surfaces and shadow regions given a B-mode US image of bones with minimal errors by using information extracted from local phase image features.

1.2 Accelerated Magnetic Resonance Imaging

Magnetic resonance imaging (MRI) as a non-invasive approach has many advantages over other imaging techniques, such as good soft-tissue contrast and no radiation [23]. However, due to the physics limitations, the MRI data acquisition is inherently slow, which hinders MRI in many clinical applications [24]. One common approach to accelerate MRI data acquisition is to take fewer measurements, generating an undersampled k-space. To increase the signal-to-noise ratio, parallel MRI with multi-channel receiver coils is routinely used. Reconstructing images from the undersampled data is challenging and has been an active research field. In Fig.1.3, we demonstrate the MRI reconstruction problem using undersampling technique where only 25% of raw MRI signal/k-space data is needed for MR imaging thus resulting a 4-time acceleration. Note that the sampling pattern shown in the undersampled k-space is an example of Cartesian sampling which undersamples k-space along the phase encoding direction. As shown in the figure, undersampling produces images of lower spatial resolution and aliasing artifacts. In order to remove these artifacts and infer the true underlying spatial structure of the imaged subject, one may apply a effective reconstruction algorithm.

Compressed sensing (CS) lays theoretical foundations for MRI image reconstruction and includes three key ingredients. First, in order to reconstruct the image from undersampled data, the sampling of k-space must be incoherent such that the aliasing artifacts behave like random noise, which can be distinguished and removed from true signals. Second, CS requires the data to be sparse either in the original image or in a transformed domain [25]. Such sparsity requirement can be considered as a prior or regularization imposed to

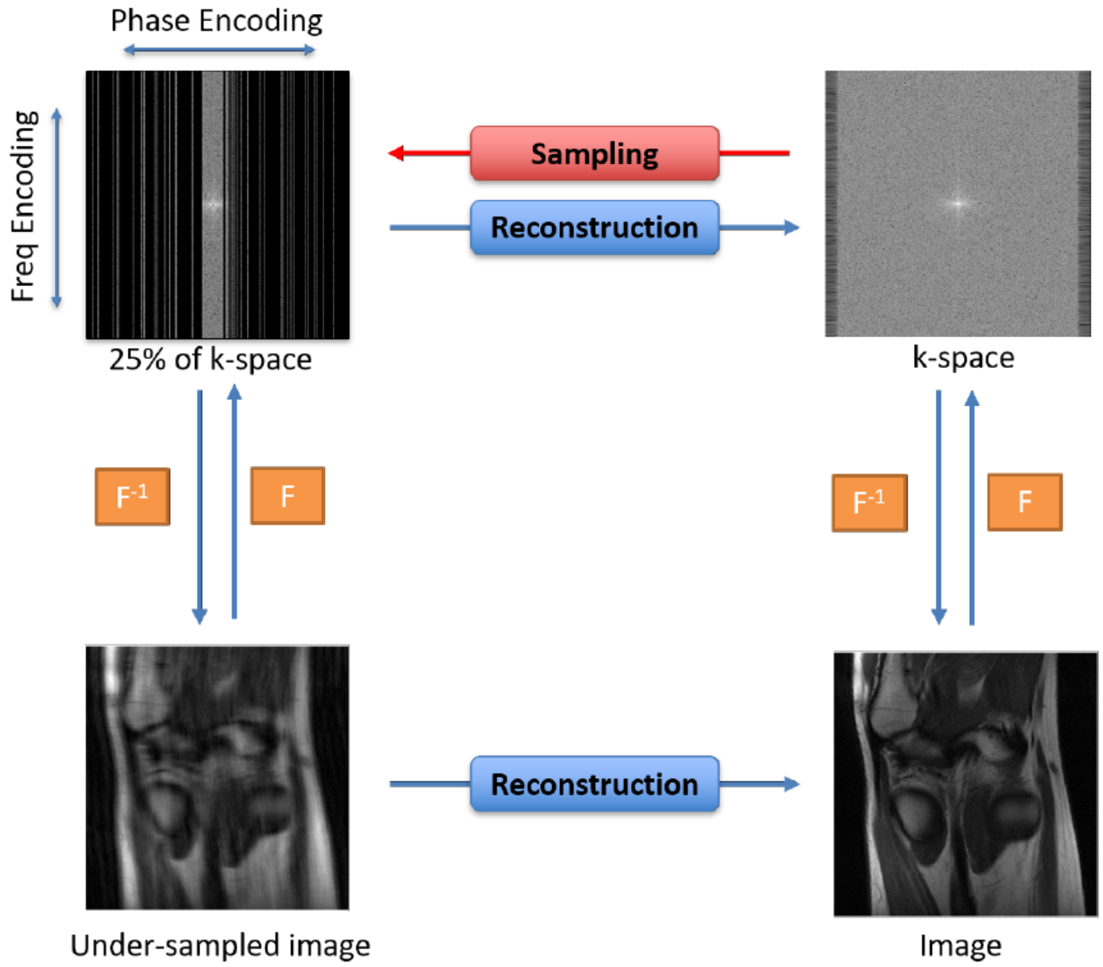


Figure 1.3: MRI Reconstruction with undersampled k-space. F and F^{-1} are 2D Fourier transform and its inverse transform.

solve the inverse problem for image reconstruction [26]. Since very few MRI image modalities are intrinsically sparse, identifying the optimal sparse transform is often the key component in CS. Those sparse transformations are often manually designed, such as total variation (TV) and wavelet transform [27,28] or a combination of transformations as in dictionary learning [29,30]. Third, the MRI reconstruction with the sparse constraint can be considered as an $\ell 1$

optimization problem. Iterative non-linear optimization algorithms are often used to search for optimal solutions [31,32].

However, those optimization algorithms are usually iterative by nature and take a relatively long computation time, which hinders or even prohibits MRI in certain clinical applications. In addition, CS based methods generally include some hyper-parameters that control the degree of the smoothness due to the sparse constraint and improper values of those hyper-parameters usually result in over-smoothed images or artifacts. Therefore, it takes great efforts to manually tune those hyper-parameters in real practice for CS based methods.

Recent advances in MRI pose new challenges to the implementations of associated algorithms in order to keep data acquisition and reconstruction times on par. While acquisition times decrease dramatically as in real-time MRI, at the same time the data grows in size when using up to 64 or even 128 independent receiver channels on modern MRI systems. In addition, new modalities such as model-based reconstructions for (dynamic) parametric mapping increase the computational complexity of reconstruction algorithms because they in general involve iterative solutions to nonlinear inverse problems. As a consequence, accelerators are increasingly used to overcome these challenges.

For parallel MRI reconstruction, there are two CS frameworks, SENSE [33] and GRAPPA [34], which perform reconstruction for multi-coil data in the image domain and k-space domain, respectively. SENSE utilizes coil sensitivity maps (CSM) for reconstruction, which are either obtained from a pre-scan, precomputed before reconstruction or estimated as part of the reconstruction algorithm [35]. Due to the computational complexity as well as the tendency to introduce artifacts, CS has taken some time to gain acceptance for clinical use.

Deep learning (DL)-based methods have been proposed for MRI reconstruction in recent years [36]. As a data-driven approach, deep learning can directly learn the optimal sparse transformation from the data. Additionally, it only takes one forward pass for the network to reconstruct the image in inference time and therefore is intrinsically faster than CS. However, it is a great challenge for current deep learning methods to recover high-frequency signals (i.e., fine details) from undersampled data, especially with a high acceleration rate. Nevertheless, the details in the reconstructed image are crucial for clinical diagnosis. Most current DL-based works tackle MRI reconstruction based on CS formulation, which focus on optimization algorithms [32] or regularization functions [27,28]. Few DL models have been proposed to specifically model the high-frequency information [37].

Chapter 2

Ultrasound Brain Ventricle Segmentation

2.1 Introduction

In this chapter, we present a novel network for segmenting brain ventricles from US images. Existing image segmentation networks such as U-Net perform poorly on US images due to low contrast and noise. To solve the issue of encoder-decoder network structure like U-net when segmenting low contrast objects with noise and artifacts in US images, in this chapter, we propose a new CNN encoder-decoder structure that combines the advantages of both SegNet and U-Net. The proposed method features an encoder that extracts deep features and a pyramid pooling decoder that leverages multi-scale information contained in the extracted deep features. We validate our proposed network against state-of-the-art segmentation methods. Figure 5.1 gives an overview of the proposed brain ventricles segmentation network.

2.2 Proposed Method

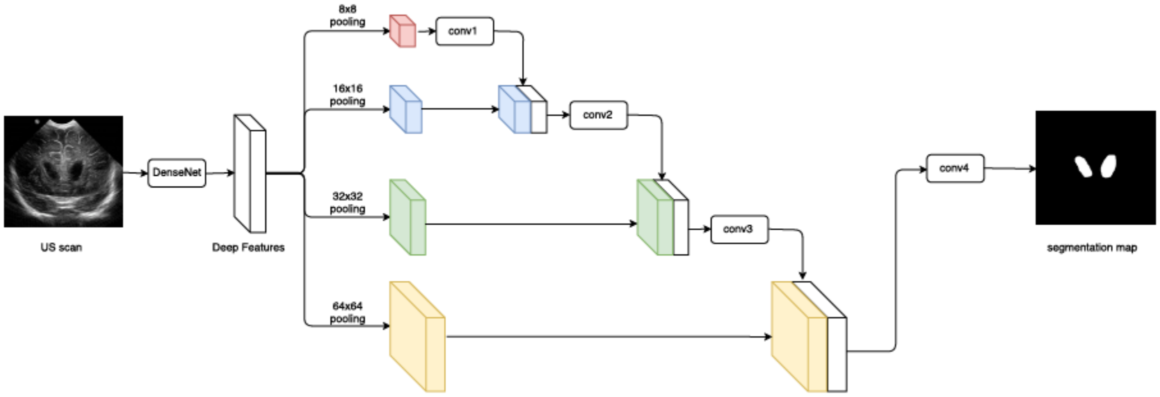


Figure 2.1: An overview of the proposed CNN architecture for brain ventricles segmentation from ultrasound images.

In this section, we provide details of the proposed US image segmentation method in which we aim to learn a mapping function between input US scans and the manual segmentation result using a specially designed CNN. The proposed method consists of two main components: deep feature extractor (i.e. encoder) and multi-scale pyramid pooling decoder.

For extracting features, we can use one of the many pretrained CNNs proposed in the literature. These CNNs often consist of either deep or shallow networks. However, increasing the depth of a network often results in an optimization difficulty. This problem has been partially solved by ResNets [38] and Highway Networks [39] with skip-connections. Recently, in [40], a novel deep neural network called DenseNet which connects each layer to every other layer in a feed-forward fashion is proposed and shown to outperform both ResNets and Highway Networks in image classification. In the proposed method, we use a pretrained DenseNet as the encoder to take the advantage of very deep neural network.

As for the decoder, the primary goal is to classify each pixel into one of two classes (ventricle or non-ventricle) given the extracted deep features. In other words, the decoder translates the input high-dimensional deep features into binary images (0: non-ventricle, 1: ventricle). As noted in U-Net [2], the key part of precise pixel-wise prediction for biomedical image segmentation task is to make good use of the multi-scale features. In our method, we accomplish this task by first pooling the feature maps into four different sizes followed by a series of transposed convolutions that transform lower dimensional feature maps into higher ones in steps.

The detailed architecture of the proposed method is illustrated in Figure 5.1, where *conv* denotes a sequence of transposed convolution, batch normalization and rectified linear unit (CONV-BN-ReLU), respectively. Note that the output of each transposed convolution is concatenated with existing feature maps of the same size and then fed into the next transposed convolution.

2.2.1 Network Architecture

As shown in Figure 5.1, we use pooling to downsample the feature maps extracted by the DenseNet into four different sizes: $8 \times 8, 16 \times 16, 32 \times 32$, and 64×64. Each of four pooled feature map has C channels, where C is number of channels in the previous original feature map. This pooling process is similar to the contracting path in U-Net but instead of the max pooling operation, we use adaptive average pooling which enables arbitrary large input size.

In order to generate the output of the same size as the input from all four different sized feature maps, upsampling is necessary. Despite many upsampling techniques, we choose transposed convolution. The decoder network starts with a transposed convolution (3×3 kernel size, stride 2 and padding 1) on the smallest pooled feature map ($C \times 8 \times 8$). As a result of stride 2, the size of the output feature map is doubled. A concatenation of the output and the corresponding pooled feature map of the same size is then fed into next the transposed convolution that has the same kernel size, stride and padding as the first one. Same process is repeated for the remaining pooled feature maps. However, because of concatenation, the number of feature channels reduce by half except for the first convolution.

Finally, in the last *conv* block as shown in right part of Figure 5.1, a number of L transposed convolutions transform the feature map ($C \times 64 \times 64$) into the final segmentation map. Here, L depends on the desired output size. The configuration of each convolution layer is given in Table 2.1. Here, $conv1$ to $conv4$ denote the sequence of Conv-BN-ReLU layers as depicted in Figure 5.1. Note that the desired final output image size is assumed to be 512×512. Hence, three transposed convolutions ($L = 3$) is needed for $conv4$.

Table 2.1: Network Configuration.

Block	Layer	Kernel Size	# Filters	Stride	Output Size
DenseNet					$C \times H \times W$
conv1	conv(1)	$C \times 3 \times 3$	C	2	$C \times 16 \times 16$
conv2	conv(2)	$2C \times 3 \times 3$	C	2	$C \times 32 \times 32$
conv3	conv(3)	$2C \times 3 \times 3$	C	2	$C \times 64 \times 64$
conv4	conv(4)	$2C \times 3 \times 3$	C	2	$C \times 128 \times 128$
	conv(5)	$C \times 3 \times 3$	C/2	2	$C/2 \times 256 \times 256$
	conv(6)	$C/2 \times 3 \times 3$	1	2	$1 \times 512 \times 512$

2.2.2 Training Details

In this section, we provide the details of training our proposed network including dataset, loss function and training parameters.

Data collection: After obtaining the approval from the Rutgers University Institutional Review Board, retrospective brain US scans were collected from subjects who were treated at the Robert Wood Johnson Medical Hospital. De-identification of the data is performed before using them for further processing. A total of 687 in vivo B-mode US images are collected. All the ventricles were manually segmented from the collected scans by an expert. Figure 4.3 shows a sample image and the corresponding ground truth image from this dataset.

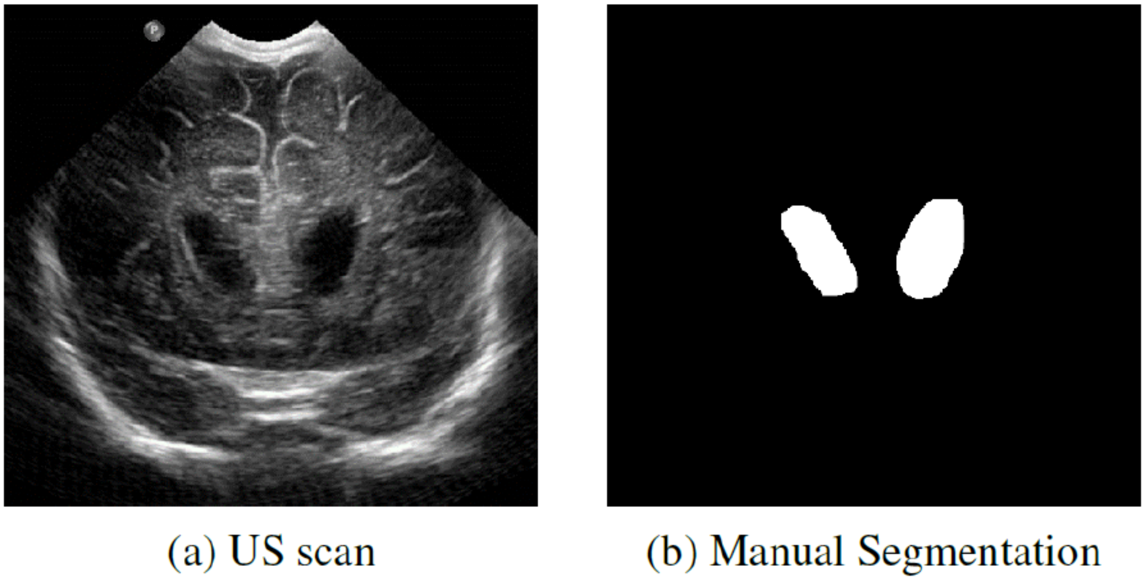


Figure 2.2: Sample brain US scan and the corresponding manual ventricles segmentation.

Data augmentation: Since a deep CNN network often requires a large number of training samples, we perform data augmentation to generate extra training samples from the original data. The input data is augmented using horizontal flip and random crop. We perform horizontal flip to every samples, therefore the total number of data is doubled. Furthermore, all images in the dataset are first resized to 600×600 and then randomly cropped to the size of 512×512 during training. By applying this data augmentation strategy, we generate sufficient samples to eliminate as much dataset bias as possible.

Loss function: Given an image and segmented map pair (Y, X), where Y is the input US image and X is the corresponding segmentation ground truth, the per-pixel L1 loss is defined as

$$L(\phi) = \frac{1}{WH} \sum_{w=1}^{W} \sum_{h=1}^{H} \|\phi(Y^{w,h}) - X^{w,h}\|_1, \tag{2.1}$$

where ϕ is the learned network (parameters) and X and Y are assumed to have the same size of $W \times H$. By using this loss function, the network is trained to minimize the L1 distance between the output and the ground truth on the training set.

Training parameters: Among many different configurations of DenseNet, we choose the pretrained DenseNet121 as our encoder and re-train it along with the decoder. Since both the input and output images are of size 512×512, the number of transposed convolutions with stride 2 in the last *conv* block should be set to 3. The entire network is trained using the ADAM optimization method [41], with mini-batches of size 12 and learning rate of 0.0002.

2.3 Experimental Results

For evaluations, we randomly select 50 samples from the whole dataset of 687 samples as the test set. The remaining 637 samples are used as the training set. The network was trained for 100 epochs to ensure the convergence of the loss function. After the network was trained, we evaluate it on the test set. We compare the performance of our method with that of the following three recent methods: SegNet [1], U-net [2] and pix2pix [3] . For all the compared methods, parameters are set as suggested in their corresponding papers and trained using the same training dataset as used to train our network.

Experiments are carried out three times. The Dice coefficient, Intersection over Union (IoU) and pixel-wise accuracy (Pixel Acc.) are used to measure the segmentation performance of different methods. Average results corresponding to three randomized tests are shown in Table 5.1. As can be seen from this table, in all three metrics, our method provides the best performance compared to the other methods. This experiment clearly shows the significance of the proposed multi-scale decoder for image segementation.

Table 2.2: Comparison of the proposed method with SegNet [1], U-Net [2] and pix2pix [3].

Method	DICE	Mean IoU(%)	Pixel Acc.(%)
SegNet [1]	0.876±0.111	80.35±0.178	87.64±0.138
U-Net [2]	0.889±0.080	82.33±0.120	89.52±0.133
pix2pix [3]	0.869±0.103	79.89±0.137	88.64±0.130
Our	**0.908±0.053**	**84.84±0.078**	**92.14±0.063**

Apart from the quantitative comparison, we also compare our method with others qualitatively by visual inspection. The segmentation results corresponding to different methods on two input US scans are shown in Figure 2.3. The second to the fifth columns of Figure 2.3 show the segmentation maps corresponding to SegNet, U-Net, pix2pix and our method, respectively. The ground truth segmentation maps are shown in the last column of this figure. It can be observed that quantitative results are consistent with the visual results. No artifacts exist in our method while SegNet and pix2pix suffer from some noticeable artifacts for the first sample. It is also evident from the second row of Figure 2.3 that our method is capable of segmenting both small and large ventricles reasonably well compared with the other methods. This clearly demonstrates the effectiveness of the proposed multi-scale pyramid pooling decoder for US image segmentation.

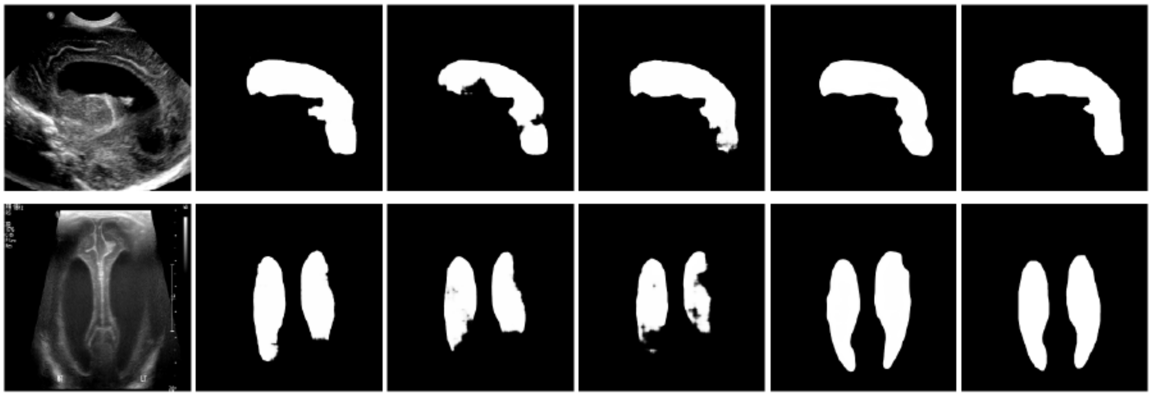


Figure 2.3: From left to right: B-mode US scans, SegNet [1], U-Net [2], pix2pix [3], our, manual segmentation.

Experiments were carried out with an Intel Xeon CPU at 3.00GHz and an Nvidia Titan-X GPU with 8GB of memory. On average our method takes about 22ms to segment an US image of size 512×512, which is sufficient for real-time applications.

2.4 Summary

The achieved results are promising for further investigation. The proposed CNN architecture achieves improved qualitative and quantitative results over previous state-of-the-art. The reported real-time computational time makes the method suitable for bedside investigation purposes. To the best of our knowledge, this work reports the first study on fully automatic real-time segmentation of ventricles from 2D US data.

Chapter 3

Simultaneous Segmentation and Classification of Bone Surfaces

3.1 Introduction

In this chapter, we propose a novel neural network architecture for simultaneous bone surface enhancement, segmentation and classification from US data. In discussing state-of-the-art we will limit ourselves to approaches that fit directly within the context of the proposed deep learning-based method. A detailed review of image processing methods based on the extraction of image intensity and phase information can be found in [14]. In [42], U-net architecture, originally proposed in [5], was investigated for processing in vivo femur, tibia and pelvis bone surfaces. Bone localization accuracy was not assessed but

0.87 precision and recall rates were reported. In [6], a modified version of the CNN proposed in [5] was used for localizing vertebra bone surfaces. Despite the fact that methods based on deep learning produce robust and accurate results, the success rate is dependent on: (1) number of US scans used for training, (2) quality of the collected US data for testing [6].

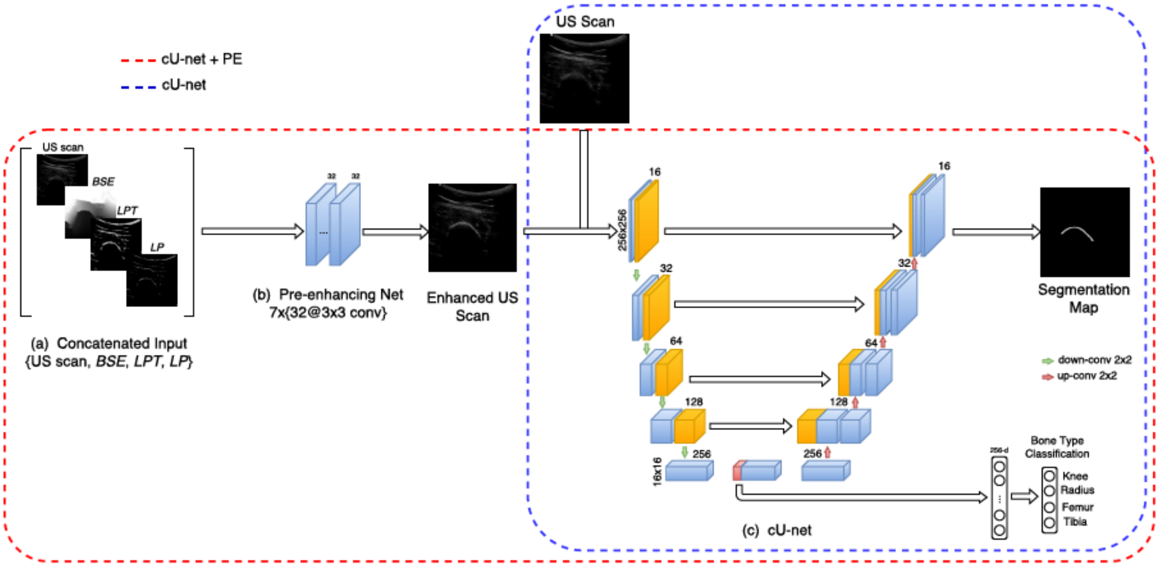
Our proposed network accommodates a bone surface enhancement network which takes a concatenation of B-mode US scan, local phase-based enhanced bone images, and signal transmission-based bone shadow enhanced image as input and outputs a new US scan in which only bone surface is enhanced. We show that the bone surface enhancement network, referred to as pre-enhancing (*PE*), improves robustness and accuracy of bone surface localization since it creates an image where the bone surface information is more dominant. As a second contribution, a deep-learning bone surface segmentation framework for US image, named classification U-net, *cU-net* for short, is proposed. Although *cU-net* shares the same basic structure with U-net [5], it is fundamentally different in terms of designed output. Unlike U-net, *cU-net* is capable of identifying bone type and segmenting bone surface area in US image simultaneously. The bone type classification is implemented by feeding part of the features in U-net to a sequence of fully-connected layers followed by a softmax layer. To take the advantages of both *PE* and *cU-net*, we propose a framework that can adaptively balance the trade-off between accuracy and running-time by combining *PE* and *cU-net*.

3.2 Proposed Method

Fig.4.1 gives an overview of the proposed joint bone enhancement, segmentation and classification framework. Incorporating pre-enhancing net, *cU-net+PE*, into the proposed framework is expected to produce more accurate results than using only *cU-net*. However, because of the computation of the additional input features and convolution layers, *cU-net+PE* requires more running time. Therefore, the proposed framework can be configured for both (i)real-time application using only *cU-net*, and (ii)off-line application using *cU-net+PE* for different clinical purposes. In the next section, we explain how the various filtered images are extracted.

3.2.1 Enhancement of Bone Surface and Bone Shadow Information

Different from using only B-mode US scan as input, the proposed pre-enhancing network, that enhances bone surface, takes the concatenation of B-mode US scan ($US(x,y)$) and three filtered image features which are obtained as follows: **Local Phase Tensor Image ($LPT(x,y)$):** $LPT(x,y)$ image is computed by



(a) Concatenated Input {US scan, *BSE*, *LPT*, *LP*}

(b) Pre-enhancing Net 7x{32@3x3 conv}

(c) cU-net

Figure 3.1: Overview of the proposed simultaneous enhancement, segmentation and classification network.

defining odd and even filter responses using [43]:

$$T_{even} = \left[\boldsymbol{H}(US_{DB}(x,y))\right]\left[\boldsymbol{H}(US_{DB}(x,y))\right]^T, \tag{3.1}$$

$$T_{odd} = -0.5 \times \left(\left[\nabla US_{DB}(x,y)\right]\left[\nabla\nabla^2 US_{DB}(x,y)\right]^T + \left[\nabla\nabla^2 US_{DB}(x,y)\right]\left[\nabla US_{DB}(x,y)\right]^T\right).$$

Here T_{even} and T_{odd} represent the symmetric and asymmetric features of $US(x,y)$. **H**, ∇ and ∇^2 represent the Hessian, Gradient and Laplacian operations, respectively. In order to improve the enhancement of bone surfaces located deeper in the image and mask out soft tissue interfaces close to the transducer, $US(x,y)$ image is masked with a distance map and band-pass filtered using Log-Gabor filter [43]. The resulting image, from this operation, is represented as $US_{DB}(x,y)$.

The final $LPT(x, y)$ image is obtained using:

$$LPT(x, y) = \sqrt{T_{even}^2 + T_{odd}^2} \times cos(\phi), \tag{3.2}$$

where ϕ represents instantaneous phase obtained from the symmetric and asymmetric feature responses, respectively [43].

Local Phase Bone Image ($LP(x, y)$): $LP(x, y)$ image is computed using: $LP(x, y) = LPT(x, y) \times LPE(x, y) \times LwPA(x, y)$, where $LPE(x, y)$ and $LwPA(x, y)$ represent the local phase energy and local weighted mean phase angle image features, respectively. These two features are computed using monogenic signal theory as [4]:

$$LPE(x, y) = \sum_{sc} |US_{M1}(x, y)| - \sqrt{US_{M2}^2(x, y) + US_{M2}^3(x, y)}, \tag{3.3}$$

$$LwPA(x, y) = \arctan \frac{\sum_{sc} US_{M1}(x, y)}{\sqrt{\sum_{sc} US_{M1}^2 + \sum_{sc} US_{M2}^2(x, y)}}, \tag{3.4}$$

where $US_{M1}, US_{M2}, US_{M3}$ represent the three different components of monogenic signal image ($US_M(x, y)$) calculated from $LPT(x, y)$ image using Riesz filter [4] and sc represents the number of filter scales.

Bone Shadow Enhanced Image ($BSE(x, y)$): $BSE(x, y)$ image is computed by modeling the interaction of the US signal within the tissue as scattering and attenuation information using [4]:

$$BSE(x, y) = [(CM_{LP}(x, y) - \rho)/[max(US_A(x, y), \epsilon)]^{\delta}] + \rho, \tag{3.5}$$

where $CM_{LP}(x, y)$ is the confidence map image obtained by modeling the propagation of US signal inside the tissue taking into account bone features present in $LP(x, y)$ image [4]. $US_A(x, y)$, maximizes the visibility of high intensity bone features inside a local region and satisfies the constraint that the mean intensity of the local region is less than the echogenicity of the tissue confining the bone [4]. Tissue attenuation coefficient is represented with δ. ρ is a constant related to tissue echogenicity confining the bone surface, and ϵ is a small constant used to avoid division by zero [4].

3.2.2 Pre-enhancing Network (*PE*)

A simple and intuitive way to view the three extracted feature images is viewing them as an input feature map of a CNN. Each feature map provides different local information of bone surface in an US scan. In deep learning, if a network is trained on a dataset of a specific distribution and is tested on a dataset that follows another distribution, the performance usually degrades significantly. In the context of bone segmentation, different US machines with different settings or different orientation of the transducer will lead to scans that have different image characteristics. The main advantage of multi-feature guided CNN is that filtered features can bring the US scan to a common domain independent of the image acquisition device. Hence, the bone surface in a US scan appears more dominant after the multi-feature guided pre-enhancing net regardless of different US image acquisition settings (Fig.5.5).

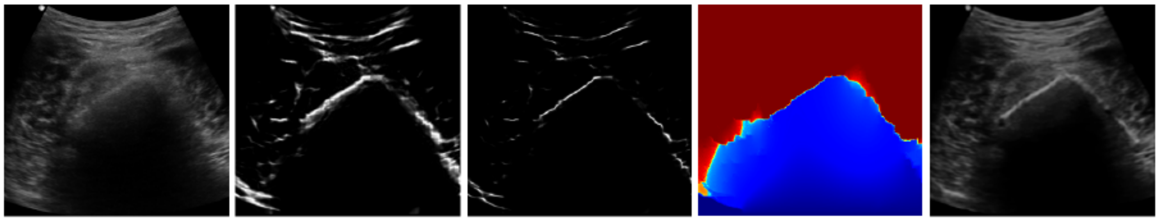


Figure 3.2: From left to right: B-mode US scan, *LPT*, *LP*, *BSE*, bone-enhanced US scan.

The input data consists of a $4 \times 256 \times 256$ matrix, i.e., each channel consists of a 256×256 image. The pre-enhancing network (*PE*) contains seven convolutional layers with 32 feature maps and one with single feature map (Fig.4.1 (b)). To balance the trade-off between the large receptive field, which can acquire more semantic spatial information and the increase in the number of parameters, we set the convolution kernel size to be 3×3 with zero-padding of size 1. The batch normalization (BN) [44] and rectified linear units (ReLU) are attached to every convolutional layer except the last one for faster training and non-linearity. Finally, the last layer is a Sigmoid function that transforms the single feature map to visible image of values between $[0, 1]$. Next we explain the proposed simultaneous segmentation and classification method.

3.2.3 Joint Learning of Classification and Segmentation

Although U-net has been widely used in many segmentation problems in the field of biomedical imaging, it lacks the capability of classifying medical images. Inspired by the observation that the contracting path of U-net shares similar structure with many image classification networks, such as AlexNet [10] and VGG net [11], we propose a classification U-net (*cU-net*) that can jointly learn to classify and segment images. The network structure is shown in Fig.4.1 (c).

While our proposed *cU-net* is structurally similar to U-net, three key difference of the proposed cU-net from U-net are as follows:

1. The MaxPooling layers and the convolutional layers in the contracting path are replaced by the convolutional layers with stride two. The stride of convolution defines the step size of the kernel when traversing the image. While its default is usually 1, we use a stride of 2 for downsampling an image similar to MaxPooling. Compared to MaxPooling, strided convolution can be regarded as parameterized downsampling that preserves positional information and are easy to reverse.

2. Different from [45], for the purpose of enabling U-net to classify images, we take only part of the feature maps at the last convolution layer of the contracting path (left side) and expand it as a feature vector. The resulting feature vector is input to a classifier that consists of one fully-connected layer with a final 4-way softmax layer.

3. To further accelerate the training process and improve the generalization ability of the network, we adopt BN and add it before every ReLU layers. By reducing the internal covariance shift of features, the batch normalization can lead to faster learning and higher overall accuracy.

Apart from the above two major differences, one minor difference is the number of starting feature maps. We reduce the number of starting feature maps from 32 to 16. Overall, the proposed *cU-net* consists of the repeated application of one 3×3 convolution (zero-padded convolution), each followed by BN and ReLU, and a 2×2 strided convolution with stride 2 (down-conv) for downsampling. At each downsampling step, we double the number of feature maps. Every step in the expansive path consists of an upsampling of the feature map followed by a dilated 2×2 convolution (up-conv) that halves the number of feature maps, a concatenation with the corresponding feature map from the contracting path, and one 3×3 convolution followed by BN and ReLU.

3.3 Data Acquisition and Training

After obtaining the institutional review board (IRB) approval, a total of 519 different US images, from 17 healthy volunteers, were collected using SonixTouch US machine (Analogic Corporation, Peabody, MA, USA). The scanned anatomical bone surfaces included knee, femur, radius, and tibia. Additional 131 US scans were collected from two subjects using a hand-held wireless US system (Clarius C3, Clarius Mobile Health Corporation, BC, Canada). All the collected data was annotated by an expert ultrasonographer in the preprocessing stage. Local phase images and bone shadow enhanced images were obtained using the filter parameters defined in [4]. For the ground truth labels we dilated the ground truth contours to a width of 1 mm.

We apply a random split of US images from SonixTouch in training (80%) and testing (20%) sets. The training set consists of a total of 415 images obtained from SonixTouch only. The rest 104 images from SonixTouch and all 131 images from Clarius C3 were used for testing. We also made sure that during the random split of the SonixTouch dataset the training and testing data did not include the same patient scans. Experiments are carried out three times on random training-testing splits and average results are reported. For training both *cU-net* and pre-enhancing net (*PE*), we adapt a 2-step training phase. In a total of 30,000 training iterations, the first 10,000 iterations were only performed on *cU-net* and we jointly train the *cU-net* and pre-enhancing net for another 20,000 iterations. We used cross entropy loss for both segmentation and classification tasks of *cU-net*. As for the pre-enhancing net, to force the network only enhance bone surfaces, we used Euclidean distance between output and input as the loss. ADAM stochastic optimization [41] with batch size of 16 and a learning rate of 0.0002 are used for learning the weights.

For the experimental evaluation and comparison, we selected two reference methods: original U-net [5] and modified U-net for bone segmentation [6] (denoted as TMI). For the proposed method, we included two configurations: *cU-net+PE* and *cU-net*, where *cU-net* is the trained model without pre-enhancing net (PE). To further validate the effectiveness of *cU-net* and *PE*, *U-net+PE* (*U-net* trained with enhanced images) and *U-net* trained using same input image features as *PE* (denoted as *U-net2*) were added to the comparison. All these methods were implemented and evaluated on segmenting several bone surfaces including knee, femur, radius, and tibia. To localize the bone surface, we threshold the estimated probability segmentation map and use the center pixels along each scanline as a single bone surface. The quality of the localization was evaluated by computing average Euclidean distance (AED) between the two surfaces. Apart from AED, we also evaluated the bone segmentation methods with regards to recall, precision, and their harmonic mean, the F-score. Since manual ground truths cannot be regarded as absolute gold standard, true positive are defined as detected bone surface points that are maximum 0.9 mm away from the manual ground truth.

3.4 Experimental Results

The AED results (mean± std) in Table 4.1 show that the proposed *cU-net+PE* outperforms other methods on test scans obtained from both US machines. Note that training set only contains images from one specific US machine (SonixTouch) while testing is performed on both. A further paired t-test between *cU-net+PE* and U-net at a 5% significance level with *p-value* of 0.0014 clearly indicates that the improvements of our method are statistically significant. The *p-values* for the remaining comparisons were also < 0.05 proving the achieved significance. The average recall and precision rates as well as F-scores are reported in Table 4.1. Although our method is not performing the best in term of average precision, the more practical measurement for detection tasks, F-score, shows the superiority of our method on bone detection performance. Further experiments of *U-net+PE* and *U-net2* yield 0.949/0.876 and 0.941/0.856 in term of F-score on both US machines. From the fact that *cU-net+PE* $>$ *U-net+PE* $>$ *U-net2*, the proposed *cU-net* and *PE* are shown to improve the segmentation result independently. Qualitative results in Fig.4.3 show that *TMI* method achieves high precision but low recall due to missing

Table 3.1: AED, 95% confidence level (CL), recall, precision, and F-scores for the proposed and state of the art methods.

	SonixTouch				Clarius C3			
	cU-net+PE	*cU-net*	U-net [5]	TMI [6]	*cU-net+PE*	*cU-net*	U-net [5]	TMI [6]
AED	**0.246±0.101**	0.338±0.158	0.389±0.221	0.399±0.201	**0.368±0.237**	0.544±0.876	1.141±1.665	0.644±2.656
95%CL	**0.267**	0.371	0.435	0.440	**0.409**	0.696	1.429	1.103
Recall	**0.97**	0.948	0.929	0.891	**0.873**	0.795	0.673	0.758
Precision	0.965	0.943	0.930	**0.963**	0.94	0.923	0.907	**0.961**
F-score	**0.968**	0.945	0.930	0.926	**0.906**	0.855	0.773	0.847

bone boundaries which is more important for our clinical application. It can be observed that quantitative results are consistent with the visual results. Average computational time for bone surface and shadow enhancement was 2 seconds (MATLAB implementation).

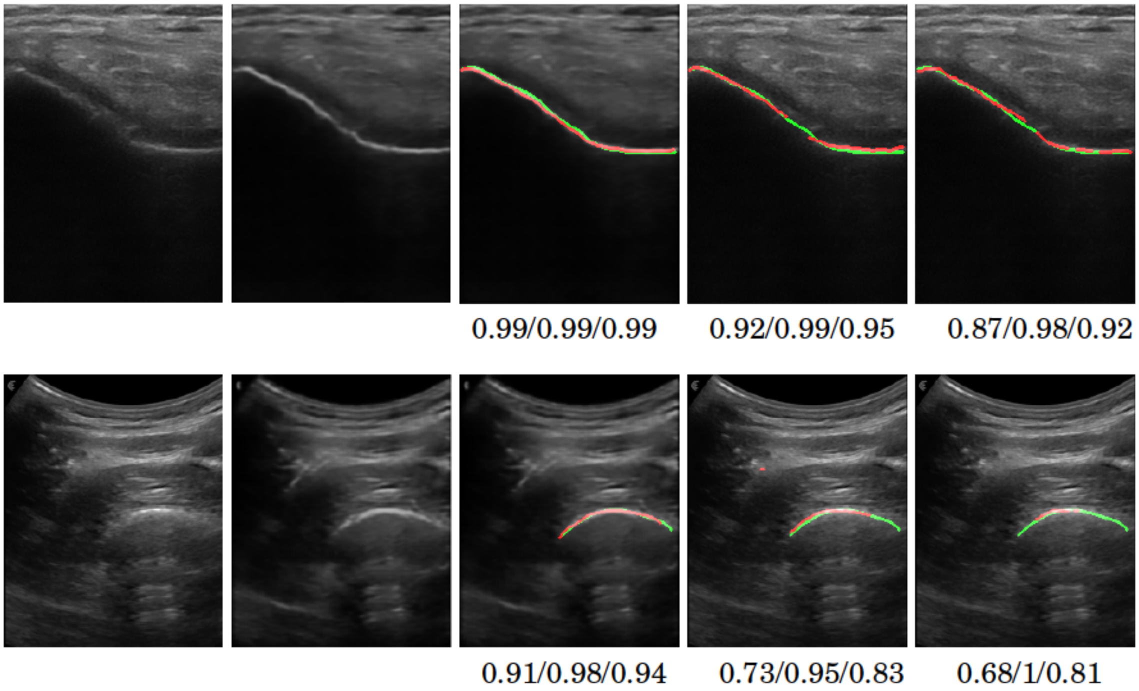


Figure 3.3: From left to right column: B-mode US scans, *PE*, *cU-net+PE*, U-net [5], TMI [6]. Green represents manual expert segmentation and red is obtained using corresponding algorithms. Recall/Precision/F-score are shown under segmentation results.

Moreover, we evaluate the classification performance of the proposed *cU-net* by calculating classification errors on four different anatomical bone types. The proposed classification U-net, *cU-net*, is near perfect in classifying bones for US images of SonixTouch ultrasound machine with an overall classification error of 0.001. However, the classification errors increase significantly to 0.389 when *cU-net* is tested on test images of Clarius C3 machine. We believe it is because of the imbalanced dataset and dataset bias since the training set only contains 3 tibia images and no images from Clarius C3 machine. Furthermore, Clarius C3 machine is a convex array transducer and is not suitable for imaging bone surfaces located close to the transducer surface which was the case for imaging distal radius and tibia bones. Due to suboptimal transducer and imaging extracted features were not representative of the actual anatomical surfaces.

3.5 Summary

We have presented a multi-feature guided CNN for simultaneous enhancement, segmentation and classification of bone surfaces from US data. To the best of our knowledge this is the first study proposing these tasks simultaneously in the context of bone US imaging. Validation studies achieve a 44% and 27% improvement in overall AED errors over the state-of-the-art methods reported in [6] and [5] respectively. In the experiments, our method yields more accurate and complete segmentation even under not only difficult imaging conditions but also different imaging settings compared to state-of-the-art. In this study the classification task involved the identification of bone types. However, this can be changed to identify US scan planes as well. Correct scan plane identification is an important task for spine imaging in the context of pedicle screw insertion and pain management. One of the main drawbacks of the proposed framework is the long computation time required to calculate the various phase image features. However, the proposed *cU-net* is independent of the *cU-net+PE*. Therefore, for real-time applications initial bone surface extraction can be performed using *cU-net* and updated during a second iteration using *cU-net+PE*. Future work will involve extensive clinical validation and real-time implementation of phase filtering.

Chapter 4

LPT-guided Real-time Bone Surfaces Segmentation

4.1 Introduction

In this chapter, in order to address the problem of segmenting bone data not only more robustly but also in real-time, we propose a novel local phase tensor [43] guided CNN architecture for bone surface segmentation from US data of various quality. In order to improve the computation time of multi-feature guided CNN [18], our proposed framework accommodates a Local Phase Tensor (LPT) network that is trained to capture contrast and noise invariant local phase information. To further improve the robustness and suppress non-bone responses of LPT network, a Global Context Tensor (GCT) network that focuses on learning global context is proposed. The two sub-networks share a common encoder and their outputs are fused to generate the final segmentation map. To take the full advantages of both LPT and GCT, we propose and evaluate three different fusion methods including addition, multiplication and concatenation. The fundamental difference of leveraging LPT image between multi-feature guided CNN [18] and proposed method is that our proposed method see LPT image as a supervision signal during training instead of a input feature to the network. This allows the further optimization of LPT image towards better bone segmentation. Thus, the proposed method can provide robust, accurate and real-time segmentation for bone US data.

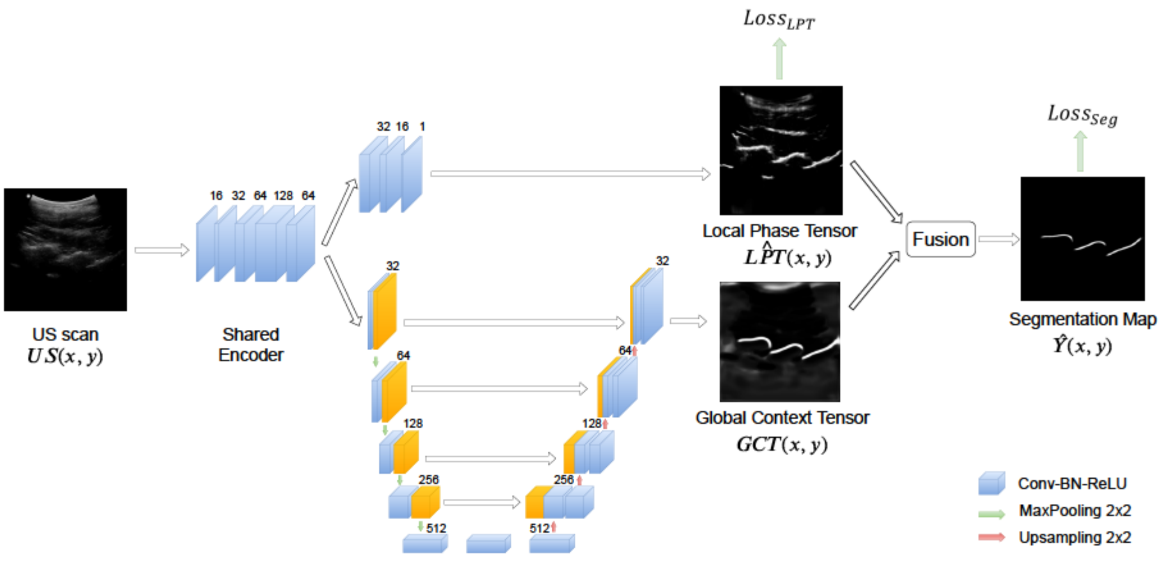


Figure 4.1: Overview of the proposed method. 3×3 zero-padded convolutions are used for all convolution layer. The network, including LPT and GCT, is trained jointly in an end-to-end fashion with $Loss_{LPT}$ and $Loss_{Seg}$.

4.2 Method

Terminology: In the following sections, we use LPT in italic font to represent the ground truth local phase tensor image, $L\hat{P}T$ for approximated LPT and LPT in non-italic font for general concept including both LPT and $L\hat{P}T$.

In our proposed robust real-time bone segmentation network architecture (Fig. 4.1), we first construct an encoder with a series of CNN blocks using Convolution (Conv), Batch Normalization (BN) [44] and Rectified Linear Unit (ReLU) [46] to capture mainly low level features from the input US scan. Then the network splits into two branches to generate the estimated local phase tensor images, denoted as $L\hat{P}T(x, y)$, and global context tensor images, denoted as $GCT(x, y)$, separately. Different from the original LPT in [43, 47], $L\hat{P}T(x, y)$ is the approximation of LPT using CNN. The LPT network output provides a local phase image response of contrast information that is independent of not only the US transducer but even the quality of the scans. Hence, the LPT network can be seen as a general boundary indicator. On the other hand GCT network provides bone-related global context information. Considering the expected feature level of each branch, the LPT and GCT subnetworks feature different network architectures to extract low and mid level features respectively. Finally, the two output are combined together using different fusion methods (addition, multiplication and concatenation) to generate the final segmentation map. To ensure the capability of LPT subnetwork outputs intensity invariant local phase features, we employ $Loss_{LPT}$ that takes pre-calculated local phase tensor images, denoted as $LPT(x, y)$, using its definition as the ground truth which we explain in the next section.

4.2.1 Local Phase Tensor

In the proposed work, local phase information is obtained using a gradient energy tensor filter. This information is used to construct local phase tensor image, denoted as $LPT(x, y)$, which highlights the contrast change and weak edges including bone surfaces in US scans. Because LPT is derived in a non data-driven way, the calculation of $LPT(x, y)$ is independent of imaging devices which provides a perfect feature for cross-machine bone segmentation. In the proposed framework, the LPT network is designed to learn a mapping function between the input image and its $LPT(x, y)$ image calculated by three Conv-BN-ReLU blocks (Fig. 4.1). This enables weakly-supervised training of the LPT subnetwork without manual annotations.

Given a B-mode US image, denoted as $US(x, y)$, $LPT(x, y)$ is obtained from odd (T_{odd}) and even (T_{even}) filter responses using [43]:

$$T_{even} = [\boldsymbol{H}(US_{DB}(x,y))] \, [\boldsymbol{H}(US_{DB}(x,y))]^T \,, \tag{4.1}$$

$$T_{odd} = -0.5 \times ([\nabla US_{DB}(x,y)] \left[\nabla\nabla^2 US_{DB}(x,y)\right]^T + \\ [\nabla\nabla^2 US_{DB}(x,y)] \, [\nabla US_{DB}(x,y)]^T),$$

where T_{even} and T_{odd} represent the symmetric and asymmetric features of $US(x,y)$, respectively. **H**, ∇ and ∇^2 represent the Hessian, Gradient and Laplacian operations, respectively. In order to improve the enhancement of bone surfaces located deeper in the image and mask out soft tissue interfaces close to the transducer, $US(x,y)$ image is masked with a distance map and band-pass filtered image using the Log-Gabor filter [43,47]. The resulting image, from this operation, is represented as $US_{DB}(x,y)$. The final $LPT(x,y)$ image is obtained using the instantaneous phase ϕ:

$$\phi = \text{ang}(s_{even}\sqrt{\text{Trace}(T_{even})} + i \cdot s_{odd}\sqrt{\text{Trace}(T_{odd})}) \tag{4.2}$$

$$LPT(x,y) = \sqrt{T_{even}^2 + T_{odd}^2} \times \cos(\phi), \tag{4.3}$$

where $s_{even} = -\,\text{sign}(\mathbf{o}^T[\mathbf{H}US_{DB}(x,y)]\mathbf{o})$, $s_{even} = -\,\text{sign}(\mathbf{o}^T[\nabla US_{DB}(x,y)])$ and **o** is the orientation vector obtained from gradient energy tensor (GET) filter [43, 47].

In order to regularize the network to approximate Eq. 4.3, we employ the following loss function between the estimated $\hat{LPT}(x,y)$ and the ground truth $LPT(x,y)$:

$$Loss_{LPT} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}|LPT(x,y) - \hat{LPT}(x,y)|, \tag{4.4}$$

where it is assumed that LPT and $\hat{LPT}$ are of size $W \times H$.

4.2.2 Global Context Tensor

As mentioned above, the LPT network can be seen as a general boundary indicator. Although, it can effectively enhance the bone surfaces in the US image, non bone boundaries such as soft tissues will also be highlighted (Fig.3). To overcome the drawbacks of the LPT network, we propose a Global Context Tensor (GCT) subnetwork for extracting the missing bone-related global context informatio n in LPT. Unlike locally computed LPT, the GCT network requires a larger receptive field to extract high-level features. Thus, we use the widely used contractive-expansive design with skip connections similar to U-net [5].

Given the output features from the shared encoder, it is fed through a 4-stage maxpooling and upsampling U-net with half the feature maps compared to the original U-net. The output of GCT network is a 2D image, denoted as

$GCT(x,y)$, of same size as $L\hat{P}T(x,y)$.

Although there is no direct supervision on GCT (no ground truth), GCT is indirectly supervised by the final segmentation loss which is discussed in following section and the goal is to refine the coarse segmentation of $L\hat{P}T(x,y)$ through a fusion layer.

4.2.3 LPT&GCT Fusion

Despite that the shared encoder and LPT network are guided by back propagating $Loss_{LPT}$, GCT must be properly regularized by incorporating it into the end-to-end bone segmentation framework. Therefore, some kind of fusion for LPT and GCT should be applied to generate the final predicted bone segmentation map $\hat{Y}$. The following three fusion methods (addition, multiplication and concatenation) to generate $\hat{Y}$ are proposed for our framework:

$$\hat{Y} = \text{Sigmoid}(L\hat{P}T + GCT) \tag{4.5}$$

$$\hat{Y} = \text{Sigmoid}(L\hat{P}T \cdot GCT) \tag{4.6}$$

$$\hat{Y} = \text{Sigmoid}(w_1 \cdot L\hat{P}T + w_2 \cdot GCT + b). \tag{4.7}$$

Note that for concatenation in Eq. 4.7, it is implemented by concatenation along channel dimension followed by a linear layer with learnable weight (w_1, w_2) and bias b that are optimized during the training process as part of the network. For all fusion methods, we define the segmentation loss using the binary cross entropy loss:

$$Loss_{Seg} = -\frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H} Y(x,y)\cdot\log\hat{Y}(x,y) + (1-Y(x,y))\cdot(1-\log\hat{Y}(x,y)), \quad (4.8)$$

where Y is ground truth segmentation mask.

4.2.4 Dataset and Experiments

With the institutional review board (IRB) approval 25 healthy volunteers were included in the study. We have collected a total of 1042 different US images using SonixTouch US machine (Analogic Corporation, Peabody, MA, USA) using 2D C5-2/60 curvilinear and L14-5 linear transducer. In order to collect new test data not used for training we have recruited 3 new subjects and collected a total of 185 US scans using a handheld wireless US system (Clarius C3, Clarius Mobile Health Corporation, BC, Canada). Image resolution varied between 0.1mm-0.15mm depending on the depth setting. Because of differences in transducer design and images reconstruction pipeline, the US scans from Clarius C3 have lower image quality in terms of bone imaging. The following bones were scanned: knee, femur, radius, and spine. Bone surfaces from the collected data were manually segmented by an expert ultrasonographer in order to generate the gold standard surfaces.

A random split of US images based on subjects from SonixTouch in training (80%) and testing (20%) sets was applied. The training set consists of a total of 834 images obtained from SonixTouch only. The remaining 208 images from SonixTouch and all 185 images from Clarius C3 were used for testing. During the random split of the SonixTouch dataset the training and testing data did not include the same patient scans. All images including ground truth LPT are normalized to [-1, 1] before feeding to the networks. For training, the overall loss is defined as:

$$Loss_{Seg} = Loss_{Seg} + \lambda \cdot Loss_{LPT}, \tag{4.9}$$

where λ is balancing weight of two losses. We searched for the optimal λ by varying it from 0 to 1 using 10% of training set as validation set. We observed that the segmentation performance will be severely impacted when λ is either less than 0.01 or larger than 0.25. But when λ is inside that range, the result stays relatively stable. Therefore, we empirically set it to 0.1 for all experiments. ADAM stochastic optimization [41] with batch size of 16 and a learning rate of 0.001 are used for learning the weights.

For validation and comparison, two reference methods were selected: original U-net [5] and state-of-the-art for bone segmentation [18] (MFGCNN). For the proposed method, we included four configurations with three different fusion method and one ablation study in which $Loss_{LPT}$ is not added. All these methods were implemented and evaluated by segmenting collected data. By thresholding the estimated segmentation map, we used the center pixels along each scanline as a single bone surface. The quality of the localization was evaluated by computing average Euclidean distance (AED) between the two surfaces along each scanline. We also evaluated the bone segmentation methods in terms of recall, precision, and their harmonic mean, the F-score. True positive are considered with 1mm tolerance. Bone surface point outside 1mm tolerance are excluded from AED error.

4.3 Results

4.3.1 Quantitative Results

The AED results in Table 4.1 show that all variations of the proposed method achieve comparable bone surface localization performance against the stat-of-the-art method, MFGCNN for both datasets. GCT-only represents the network without LPT branch and has a sigmoid activation layer added to the end. Note that training set only contains images from US machine (SonixTouch) while testing is performed on both including the low quality images from handheld

Table 4.1: AED (mm), standard deviation of AED, recall, precision, and F-scores for the proposed and state-of-the-art methods. All methods are trained only on SonixTouch data. CAT, ADD and MUL denotes three fusion methods, concatenation, addition and multiplication. LPT is the option of adding $Loss_{LPT}$. Best number across all methods is in **bold** font.

	Method	U-net [5]	MFGCNN [18]	GCT-only	Proposed					
	CAT				✓			✓		
	ADD					✓				✓
	MUL						✓		✓	
	LPT							✓	✓	✓
SonixTouch	AED	0.371	0.332	0.358	0.340	0.334	0.359	0.332	0.347	**0.325**
	std	0.172	0.164	0.169	0.158	0.152	0.173	0.148	0.167	0.158
	Recall	0.878	0.939	0.917	0.934	0.935	0.932	0.941	0.940	**0.948**
	Precison	0.809	0.809	0.808	0.81	0.805	0.812	0.806	0.811	**0.815**
	F-score	0.842	0.869	0.859	0.869	0.865	0.868	0.868	0.870	**0.877**
Clarious C3	AED	0.500	**0.401**	0.453	0.416	0.429	0.438	0.422	0.427	0.410
	std	0.351	0.201	0.310	0.217	0.228	0.243	0.237	0.235	0.215
	Recall	0.698	0.787	0.712	0.750	0.765	0.727	0.817	0.803	**0.847**
	Precison	0.880	0.902	0.894	0.897	0.882	0.906	0.898	0.911	**0.920**
	F-score	0.779	0.841	0.793	0.817	0.819	0.807	0.856	0.854	**0.882**

wireless US scanner (Clarius C3). It is worth mentioning that although AED results don't show significant difference across different methods, 0.3 mm (2 pixels) error for segmented bone surfaces can be very well accepted for US-based segmentation.

The average recall and precision rates as well as F-scores in Table 4.1 clearly demonstrate the robustness of our proposed method for both datasets. By adding $Loss_{LPT}$, the F-score on Clarius C3 are boosted from 0.817 to 0.882 compared to 0.841 for MFGCNN (paired t-test $p<0.05$). Similar significant improvements are also observed for precision and recall rates (paired t-test $p<0.05$). Although both MFGCNN and the proposed method utilize LPT features, MFGCNN does not have direct supervision on leveraging/extracting information of LPT features. LPT features are used as supervision signals to generate potentially better intermediate features for bone segmentation. This is done by optimizing the estimated $\hat{LPT}$ over final segmentation loss $Loss_{Seg}$.

It is also shown that fusion by addition outperforms multiplication and concatenation in terms of recall, precision and F score. The possible reason for concatenation fusion method is that it can easily let the estimated GCT dominate the final segmentation without leveraging the information in LPT. For example, in concatenation fusion equation Eq. 4.7, the network can solely depend on GCT by simply putting w_1 to near zero without increasing the training loss. This conjecture is supported by the observation that optimized $w_1 = 0.0031$ and $w_2 = 7.5725$ after training. As for multiplication fusion, it requires GCT subnetwork to precisely localized the false predicted pixels in $\hat{LPT}$ in order to correct it by flipping the sign since $\hat{LPT}$ is in the range of [-1, 1]. While for addition fusion, false predictions in $\hat{LPT}$ can be possibly corrected by uncertain prediction from GCT.

The box plot for recall rates of U-net, MFGCNN and the proposed method with addition as fusion in Fig. 4.2 further demonstrate the superior cross-machine bone surface segmentation ability of providing more complete result. Note that from the AED and recall rate results on SonixTouch and Clarious C3 data in Table 4.1 and Fig. 4.2, one can observed that all methods suffer performance drops when tested on Clarious C3 data. This is the result of low bone surface contrast of Clarious C3.

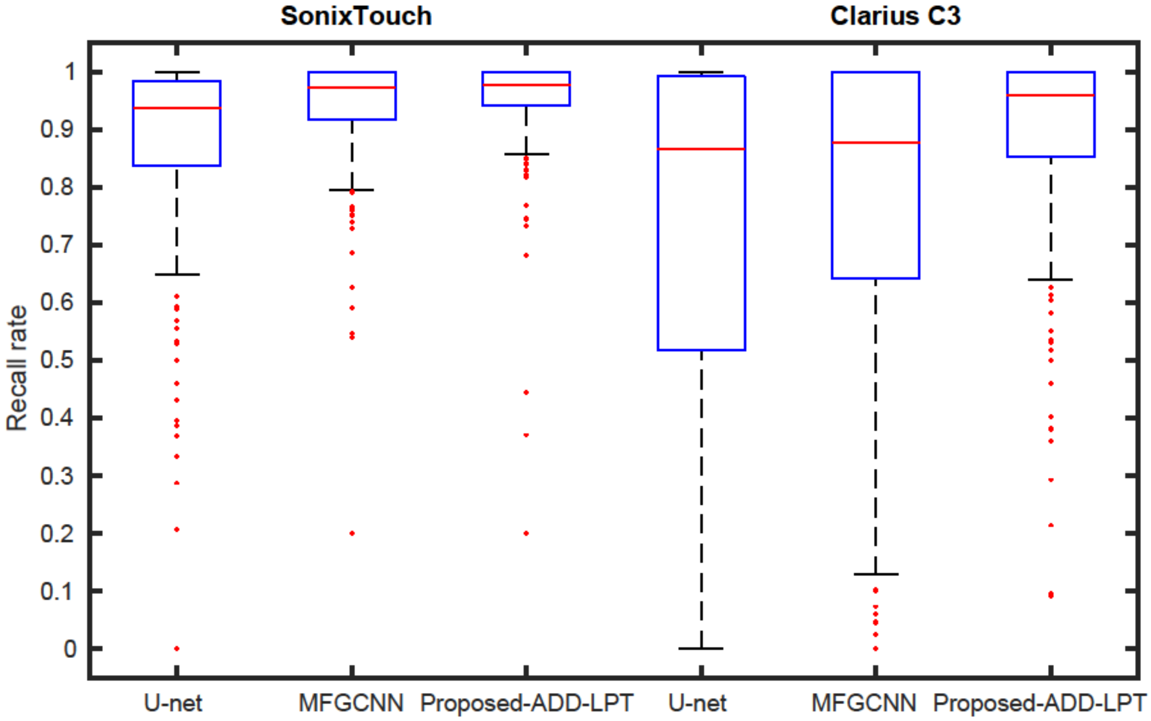


Figure 4.2: Recall rates reflects the completeness of segmentation results which is the main drawback of existing bone segmentation methods when testing on data from different US machine.

4.3.2 Qualitative Results

Qualitative results in Fig.4.3 show that our method achieves improved and complete segmentation results while U-net and MFGCNN suffer from missing bone segments which is crucial for US-based intra-operative guidance using features extracted from US data. Average computational time for our method is 30 ms compared to 2 seconds for the complete MFGCNN framework. This is an improvement of 98.5 % over MFGCNN [18]. The comparison in term of

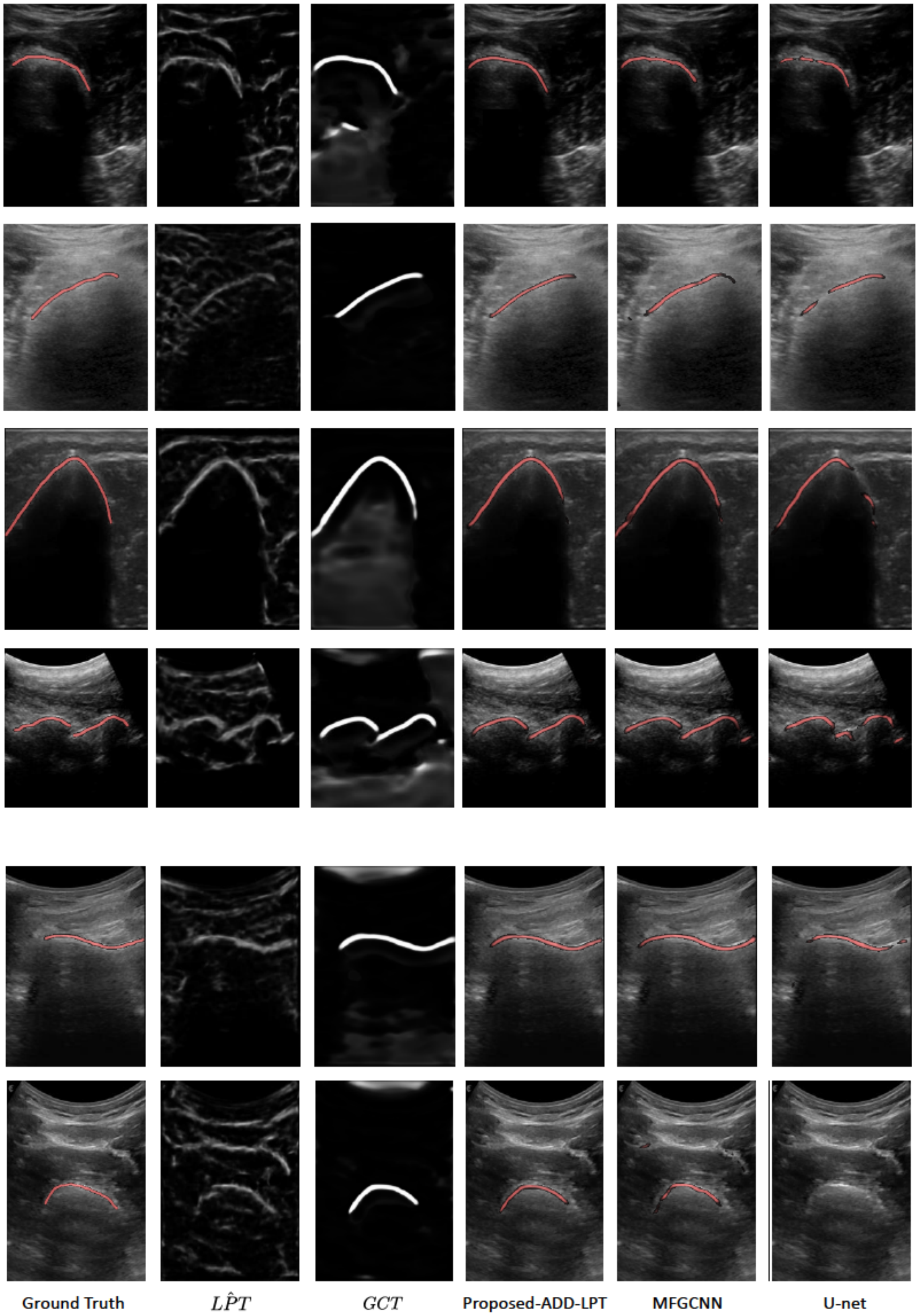


Figure 4.3: Bone segmentation results. Top four and bottom two rows show in vivo B-mode US scans from SonixTouch and Clarius C3 US machines respectively.

image quality between two US machines can be made by comparing first and last row in Fig.4.3 that both are femur bone US scans. The US image (last row) from Clarius C3 shows lower bone surface contrast compared to SonixTouch. However, both estimated $L\hat{P}T$ images can enhance bone surfaces very well despite of intensity/gradient difference in two US images.

Although LPT is used as supervision signal to regulate the estimated $L\hat{P}T$, we observed that the estimated LPT in our proposed network can produce a better coarse bone segmentation inside bone regions. We demonstrated this in Fig. 4.4 by showing both LPT and ground truth bone segmentation. It is clear to see that estimated $L\hat{P}T$ has more clear and complete bone enhanced signal around bone surfaces.

4.4 Summary

In order to make US an essential imaging modality in orthopedics clinically acceptable accuracy and robustness of guidance system needs to be ensured. Therefore, complete, accurate and robust bone segmentation is of paramount importance for US-based orthopedic surgical and non-surgical procedures where automatically extracted bone surfaces are used for continuous real-time guidance.

We proposed an end-to-end local phase guided framework that enables ro-

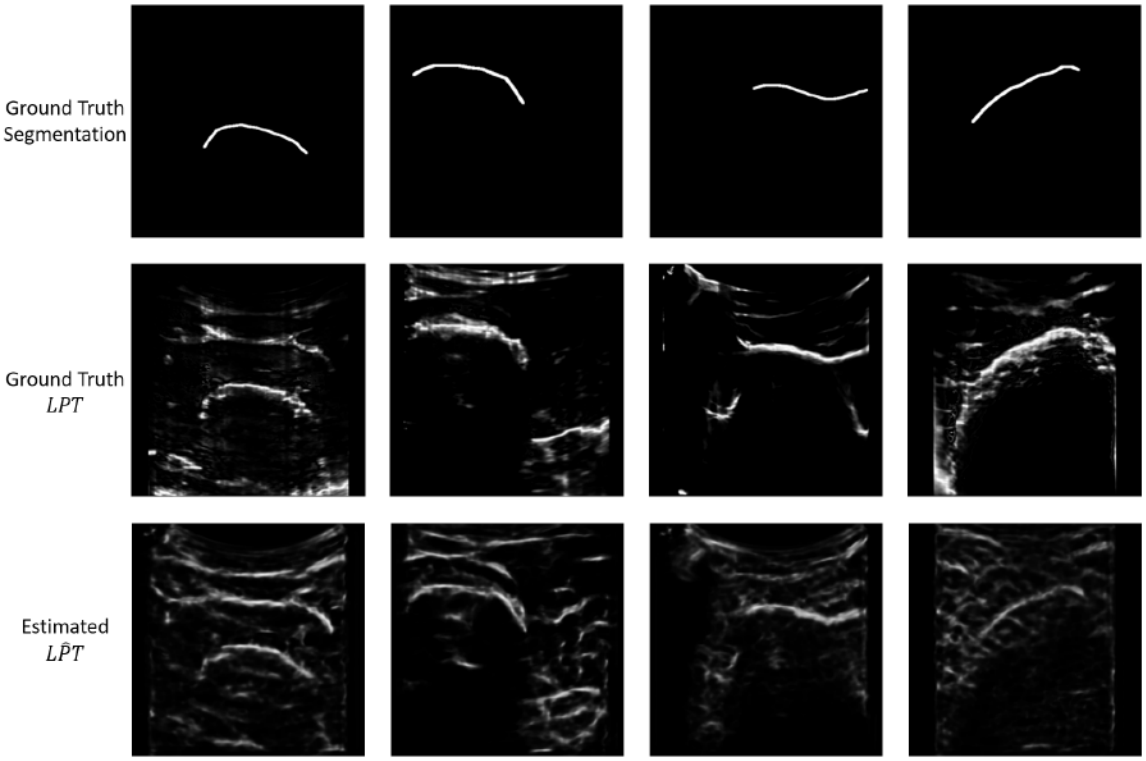


Figure 4.4: Comparison of ground truth LPT and estimated $L\hat{P}T$. Ground truth bone segmentation is shown at the top row for reference.

bust and accurate bone surfaces segmentation for US-based computer assisted orthopedic procedures. The main novelty of our work lies in (1) the integration of learning Local Phase Tensor (LPT) and Global Context Tensor (GCT) into a single network, (2) the design of fusion method of LPT and GCT to improve cross-machine segmentation performance of various bone imaging quality, and (3) the first systematic design of a fully automatic real-time framework for robust multi-machine LPT-guided bone surface segmentation from US images. It is critical for an automatic US segmentation algorithm to maintain robust performance on various US machines without any modification. Furthermore, this is the first study proposing a CNN-based local phase image generation network which we believe is an important contribution in the field of US-based orthopedic procedures.

Through validation, we demonstrate the state-of-the-art sensor adaption capability of our proposed method by separating training and testing data into two US machines. our proposed method achieves an AED of 0.32mm and 0.41mm respectively with a processing time of 30ms (98.5% improvement over state-of-the-art in processing speed [18]).

Chapter 5

Bone Shadow Segmentation Through Task Decomposition

5.1 Introduction

As shown in previous chapters, by incorporating local phase-based image features into deep learning framework, we can significantly improve the bone surface segmentation result. However, applying such local phase information to bone shadow segmentation remains an open problem.

Our goal is to improve bone shadow segmentation by proposing a deep learning-based method which yields better performance over other methods. The motivation and contribution of the proposed method are as follows:

- In Chapter 2, we proposed a novel CNN segmentation network with pyramid pooling decoder and it shows promising segmentation performance compared to common used U-net. We use it as our base segmentation network for its great ability at segmenting shapes with large variations.

- Because of US imaging principle and anatomy of bone structures, bone shadows share some common shape profiles. In Fig.1.2 (c), the gold standard bone shadow will ideally have sharp horizontal cut-off for non-bone area and certain bone surfaces on top. Thus we propose an adversarial network to implicitly impose the shape regularization.

- Expert manual annotations of medical images are expensive and time consuming. We leverage the bone shadow image features extracted using the method proposed in [4] and use it as surrogate ground truth to not only provide additional supervision on intermediate results, but also enable the semi-supervised learning for US bone shadow segmentation.

- By using only left and right boundary of bone, one can create a horizontal bone interval mask and apply it on bone shadow enhanced image, Fig.1.2 (b), to output bone shadow segmentation results that are close to ground truth. we propose a subnetwork that estimate the bone regions horizontally by only learning from manually annotated bone landmarks which has lower annotation cost than full segmentation. This could lead to larger scale dataset for training.

Despite the fact that each of the above three contributions can improve the segmentation performance individually, we integrate them all together into one end-to-end deep learning framework in which the three parts can strengthen each other towards better final segmentation. In Sec.5.2, we discuss the details of each components and the overall network structure. Extensive experiments are conducted to demonstrate the effectiveness of the proposed method. Furthermore, in Sec.5.4 an ablation study is conducted to show the contribution brought by different components in our method.

5.2 Proposed Method

In the proposed method, two subnetworks are first trained separately to produce a coarse bone shadow enhancement (BSE) and horizontal bone interval mask (HBIM). After obtaining both coarse BSE and HBIM, a masking operation is used to generate the final bone shadow. As a result, we provide a joint trainable end-to-end deep learning model for robust bone shadow segmentation. The proposed CNN model consists of one shared encoder and two independent multi-scale decoders for coarse BSE and HBIM estimation. To further regularize the shape of the output bone shadow, we introduce a conditional shape discriminator which can guide the training of bone shadow segmentation network by adding the adversarial loss on the shape information. Fig.5.1

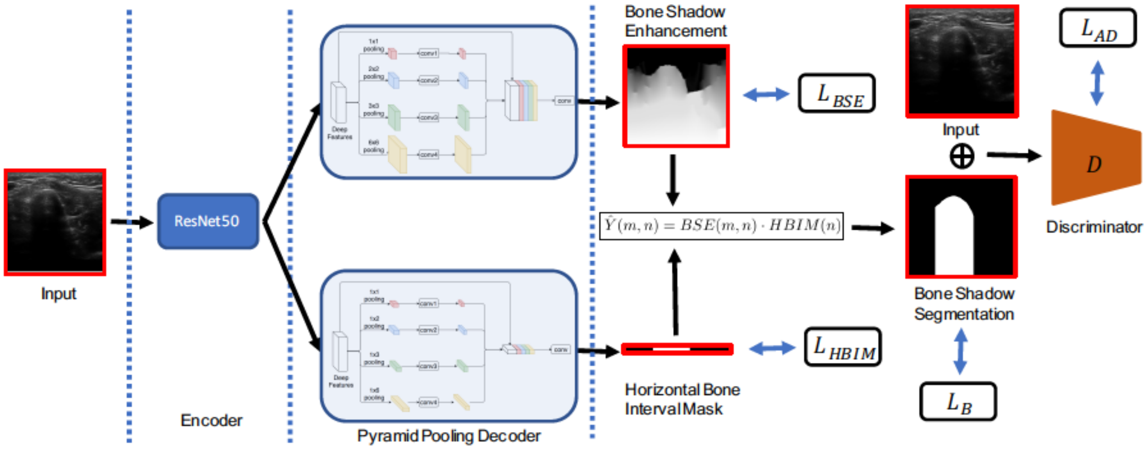


Figure 5.1: An overview of the proposed multi-task learning-based method for bone shadow segmentation from US images.

provides an overview of our framework.

5.2.1 Conditional Shape Discriminator

Unlike other semantic segmentation tasks, bone shadow segmentation is different in many ways. One major difference of the output segmentation map is the general shape. The type of bones (knee, fibia , femur, etc), view planes (longitudinal and transverse), and most importantly the orientation of the US transducer with respect to the imaged bone anatomy would affect the bone shadow shape individually.

To ensure specific shape on the estimated bone shadows by a CNN, a conditional shape discriminator D is added in the training stage and designed following a conditional Generative Adversarial Network (cGAN) framework [48].

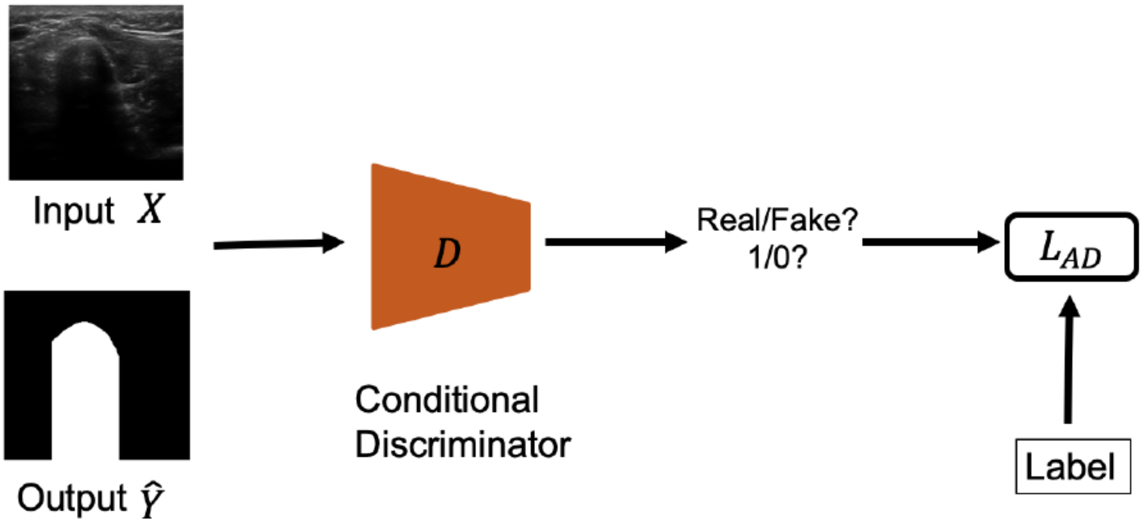


Figure 5.2: Conditional Shape Discriminator.

It takes both the input image and its corresponding bone shadow segmentation (segmentation from proposed network or ground truth) to identify if the segmentation is ground truth on the basis of binary images. From the perspective of segmentation network, it regularizes N output $\hat{Y}$ using the binary cross entropy loss:

$$L_{AD} = -\frac{1}{N}\sum_{i=1}^{N}[\log(1 - D(X_i, \hat{Y}_i)) + \log(D(X_i, Y_i))], \tag{5.1}$$

where X_i is input image and Y_i is the corresponding ground truth. Because, for binary segmentation task, the output segmentation is binary which varies in different shapes, this adversarial loss can effectively enforce the output segmentation map to follow a reasonable shape even with different types of bones and view planes.

5.2.2 Coarse Bone Shadow Enhancement

One of the main challenges in deep learning-based medical image analysis is the generalization ability of the trained model due to the lack of large amounts of manually annotated data. However, recent studies have shown that, by training the model through semi supervised learning on automatic annotated or weakly labelled data, the model gains better generalization ability and improves the overall performance even for different imaging modalities [49].

In this work, we propose to use Bone Shadow Enhancement (BSE) method, proposed in [4], to filter the US image and generate a coarse estimation of the bone shadow regions. BSE image signal at position (x, y) is computed by modeling the interaction of the US signal within the tissue as scattering and attenuation information using:

$$BSE(x,y) = [(CM_{LP}(x,y) - \rho)/[max(US_A(x,y), \epsilon)]^{\delta}] + \rho, \tag{5.2}$$

where $CM_{LP}(x, y)$ is the confidence map image obtained by modeling the propagation of US signal inside the tissue taking into account bone features present in local phase bone image $LP(x, y)$ [4]. $US_A(x, y)$ maximizes the visibility of high intensity bone features inside a local region and satisfies the constraint that the mean intensity of the local region is less than the echogenicity of the tissue confining the bone [4]. Tissue attenuation coefficient is represented by δ. ρ is a constant related to tissue echogenicity confining the bone surface, and ϵ is a small constant used to avoid the division by zero [4].

5.2.3 Horizontal Bone Interval Mask

As shown in Fig.1.2 (b) and (c), the previously defined BSE image can be regarded as a coarse estimation of bone shadow regions. While the sharp boundary of the bone surface is usually well preserved, it can also have high confidence shadows leaking into non-bone regions horizontally. To solve this shadow leakage problem, image processing technique that can remove shadows corresponding to non-bone structure while keeping the bone shadow needs to be applied on the BSE image. From the observation that the shadow leakage usually happens below the bone surface and expands horizontally, a Horizontal Bone Interval Mask (HBIM) is proposed to mask out the non-bone shadows. Given a US image $X(m, n)$ of size $N \times M$, its corresponding BSE image $BSE(m, n)$ and the manually segmented bone shadow $Y(m, n)$, HBIM is defined as follows:

$$HBIM(n) = \begin{cases} 1, & \text{if } \exists\, m,\ Y(m,n) > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

HBIM can be seen as a vector in which 1 indicates the presence of bone surface along corresponding vertical line in US image. Thus we can derive the final fine bone shadow segmentation $\hat{Y}$ using HBIM as follows,

$$\hat{Y}(m,n) = BSE(m,n) \cdot HBIM(n). \tag{5.4}$$

As a result, one is able to calculate a high quality bone shadow segmentation using only the input US image and the horizontal location information of the bone in the US image. Moreover, as will be shown later, this leads to a much more robust and predictable bone shadow segmentation than a simple end-to-end training scheme.

5.2.4 Network Structure

The proposed framework features three tasks: 1) Coarse BSE estimation, 2) HBIM estimation, and 3) final bone shadow segmentation. Noticeably, with three different tasks, our proposed framework is a multi-task learning (MTL) model.

We view the first two tasks as intermediate tasks that are highly correlated with the final task. In the proposed method, we use a ResNet50 [38] pretrained on ImageNet [50] as the shared encoder to take the advantage of very deep neural network. The first convolutional layer is modified to take a single channel input. While the ResNet50 encoder is shared across all tasks for deep feature extraction, each of the intermediate tasks has its own decoder. As noted in U-Net [2], the key part of precise pixel-wise prediction for biomedical image segmentation task is to make good use of the multi-scale features. In our network, we adopt the decoder that was first proposed in [51]. For HBIM estimation, the desired output is a one-dimensional row vector. In order to achieve that, we changed the pyramid pooling to $(1,1),(1,2),(1,3),(1,6)$ and add another average pooling layer between the input deep feature and concatenation to align the feature size.

Finally, we complete the final bone shadow segmentation by using the estimated HBIM and BSE of previous two tasks following Eq.5.4. Given the proposed MTL model for these three tasks, it turns out helpful to extract comprehensive image features by sharing a shared encoder and then branching out for task-specific losses for each task. To further maximize the synergy across all the tasks, we propose a combined loss function containing four task-specific losses: $L = L_{BSE} + L_{HBIM} + L_B + \lambda L_{AD}$, where L_{HBIM} and L_B are binary cross entropy loss of estimated HBIM and bone shadow, and L_{BSE} is the L_1 loss of estimated BSE. L_{AD} represents the adversarial loss (loss from the discriminator D) as defined in Eq.5.1 with weight λ. As for the structure of the discriminator D, we follow the structure that was proposed in [52].

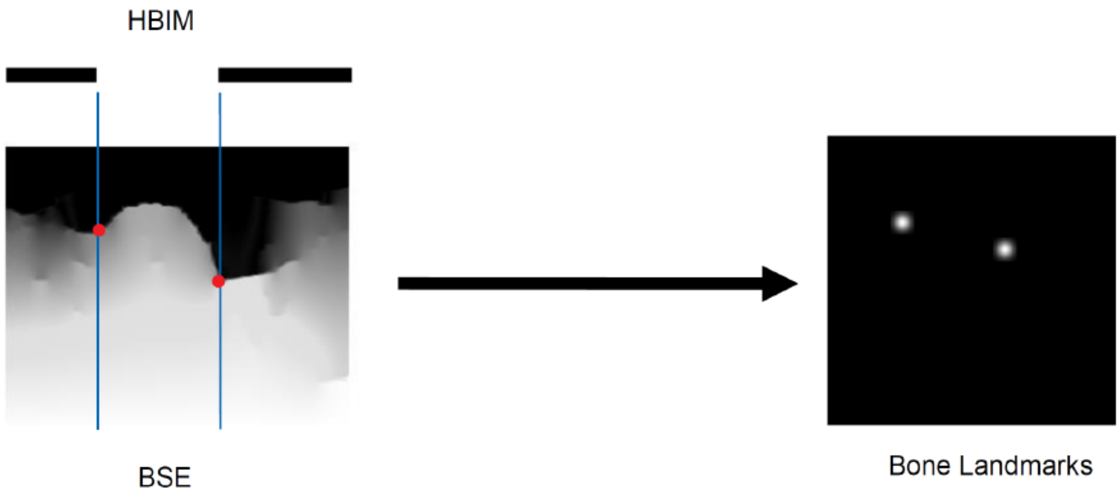


Figure 5.3: Bone landmark detection from HBIM.

5.2.5 Landmark Detection from HBIM

In this subsection, we demonstrate how HBIM enables bone landmark detection. First we define the bone landmark as two horizontally furtherest pixels of a bone segment in bone US image. In the case where multiple bone segments exist in an image (i.e spine), there will be more than two landmarks. Although, from Eq.5.3, HBIM is only a one-dimensional mask which represents horizontal location of a bone, we can fully recover the coordinates of the boundaries by leveraging the estimated HBIM and BSE from the outputs of the proposed model. Consider a set of K bone landmarks $\{(s_1, t_1),\ \ldots, (s_K, t_K)\}$, each (s_k, t_k) can be derived as,

$$s_k = \min m \mid BSE(m, h_k) > \theta, \qquad t_k = h_k, \tag{5.5}$$

where h_k denotes the horizontal coordinate of the *k-th* boundary line in HBIM which can be easily derived by finding the local maxima of the absolute derivative and θ is the threshold for identifying bones. Compared to the traditional CNN-based landmark detection methods which have a fixed number of landmarks, our landmark detection from HBIM can adapt as many landmarks as the number of bones present in the US image without modifying the network.

5.3 Dataset and Training

After obtaining the institutional review board (IRB) approval, a total of 814 different US images, from 20 healthy volunteers, were collected using Sonix-Touch US machine (Analogic Corporation, Peabody, MA, USA). The scanned anatomical bone surfaces include knee, femur, radius, and tibia. All bone shadows of the collected data were manually annotated by an expert ultrasonographer in the preprocessing stage. The BSE images were obtained using the filter parameters defined in [4] and the HBIMs were obtained using Eq.5.3 with bone shadow annotations. The datasets were randomly separated on the subject level into training and testing sets by an 60%/40% split (573/241 in images level). Any subject with data included in the training set were excluded from the testing set. During preprocessing, the images were resampled into 0.15mm isotropic resolution, and resized to 256×256.

The coarse BSE and HBIM estimation tasks are trained first with a batch size of 32 for 100 epochs in which only L_{BSE} and L_{HBIM} are used to train the network. The base network is optimized by the Adam optimizer with a learning rate of 10^{-4}. A joint training using all four losses is applied afterwards with a batch size of 32 for 50 epochs with $\lambda = 0.1$. During testing, the image can be forwarded though the network for all tasks by one shot. The experiments are performed on a Linux workstation equipped with an Intel 3.50 GHz CPU and a 12GB NVidia Titan Xp GPU using the PyTorch framework. The average running time of our model for single testing image is around 0.03 seconds which makes real-time application possible.

5.4 Experimental Results

5.4.1 Bone Shadow Segmentation

We compare the performance of our method with that of the following four methods: Unet [2], PSPnet [51], PSPGAN and PSPnet-MTL. PSPGAN denotes the method that combines the proposed conditional shape discriminator in Sec.2.1 and PSPnet. PSPnet-MTL is the multi-task version of PSPnet without conditional shape discriminator. The comparison between PSPGAN, PSPnet-MTL and PSPGAN-MTL is for the purpose of ablation study. For all

Table 5.1: Bone shadow segmentation and bone surface localization comparison of methods on various metrics. The proposed PSPGAN-MTL achieves statistically significant improvements using two-tailed t test with p values < 0.05.

	Bone shadow segmentation			Bone surface localization			
	Dice	mIoU(%)	pAcc.(%)	AED	Recall	Precision	F-score
U-net [2]	0.890±0.068	80.97	87.86	2.11±1.05	0.625	0.616	0.620
PSPnet [51]	0.911±0.062	85.69	92.76	1.36±1.41	0.730	0.825	0.775
PSPGAN	0.927±0.056	86.98	92.83	1.49±1.69	0.727	**0.826**	0.774
PSPnet-MTL	0.956±0.052	92.08	96.47	0.25±0.19	0.894	0.748	0.918
PSPGAN-MTL	**0.962±0.046**	**92.97**	**96.63**	**0.19±0.13**	**0.907**	0.775	**0.934**

the compared methods, parameters are set as suggested in their corresponding papers and trained using the same training dataset as used to train our network.

The Dice coefficient, mean Intersection over Union (mIoU) and pixel-wise accuracy (Acc.) are used to measure the segmentation performance of different methods. Average results of all test scans are shown in Table 5.1. As can be seen from this table, in all three metrics, our method provides the best performance compared to the other methods. Going directly from PSPGAN to PSPGAN-MTL provides implicit data augmentation and bone shadow prior for the tasks with limited data, thus results in a much more robust and accurate bone shadow segmentation. By adding proposed conditional shape discriminator, both PSPGAN and PSPGAN-MTL can outperform their counterparts, PSPnet, PSPnet-MTL. These experiments clearly show the significance of each component of proposed method, integrating coarse BSE estimation and HBIM for bone shadow segmentation and conditional shape discriminator.

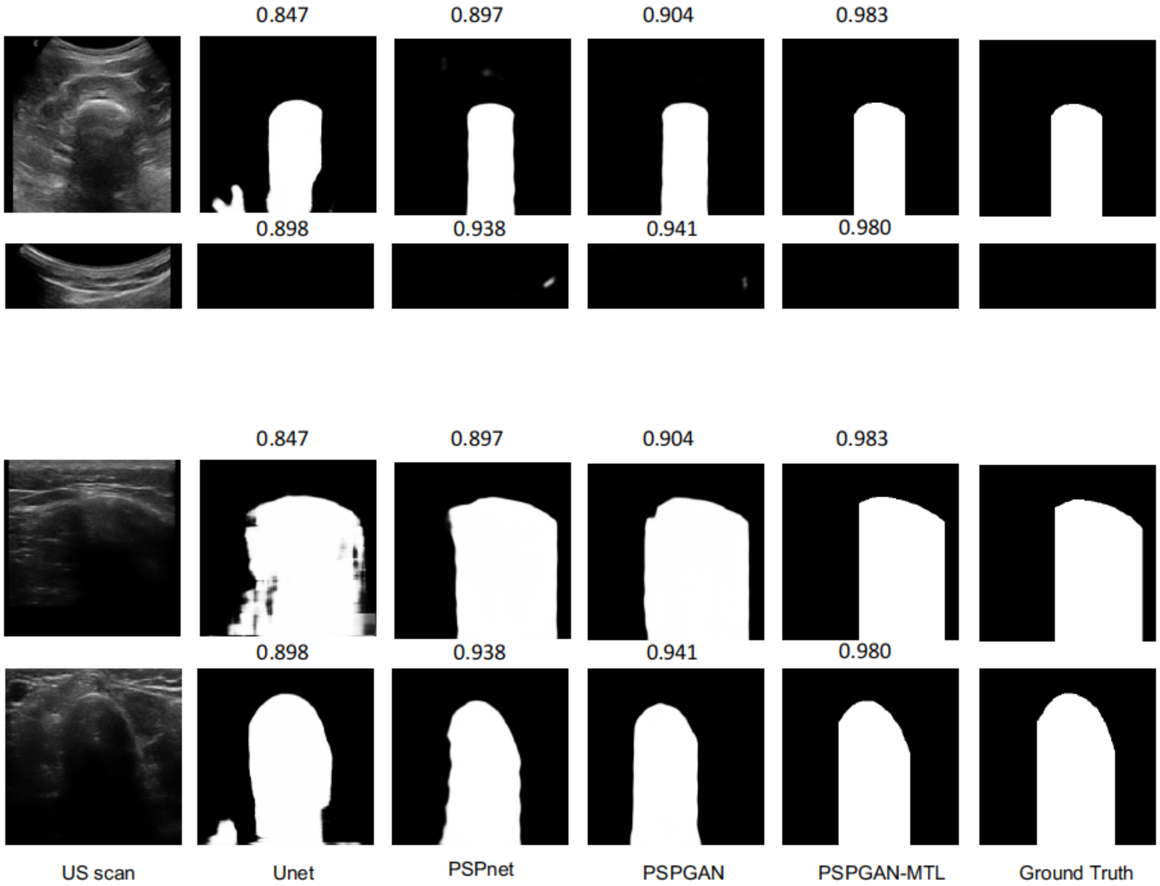


Figure 5.4: Bone shadow segmentation results for in vivo tibia, distal radius, knee and femur. Dice coefficients computed against the ground truth are shown on top of the each result.

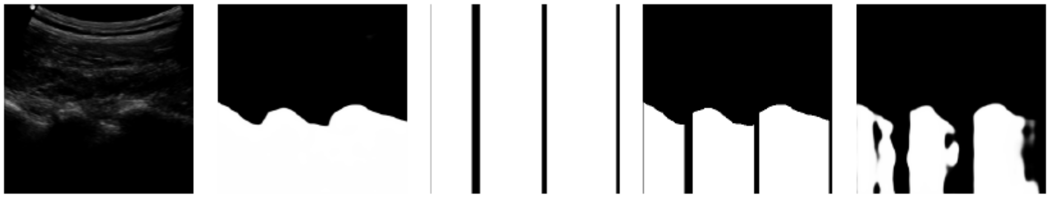


Figure 5.5: From left to right: In vivo US scan of spine, estimated BSE, estimated HBIM, PSPGAN-MTL, PSPGAN.

Apart from the quantitative comparison of Dice, mean IoU and pixel accuracy, we also compared our method PSPGAN-MTL with others qualitatively by visual inspection. The segmentation results corresponding to different methods and the intermediate outputs of PSPGAN-MTL are shown in Fig.5.4. The more shape alike PSPGAN result shows the effect of the proposed conditional shape discriminator comparing to PSPnet.

For the final experiment of bone shadow segmentation, we compare two methods: PSPGAN-MTL and PSPGAN, in term of their ability to correctly segment spine (multiple bones) which is not present in the dataset. From the results shown in Fig.5.5, it is clear that with the help of the proposed multi-task bone shadow segmentation, PSPGAN-MTL suffers no mis-segmentation and provides a more complete segmentation compared with PSPGAN.

5.4.2 Landmark Detection

Moreover, we evaluate the bone landmark detection performance of our method proposed in Sec.5.2.5 using HBIM. We compare our method with a simple landmark detection model with PSPnet as the base network and modify the final convolution layer to generate the bone landmark probability map. The average localization errors in mm for our method and the baseline landmark detection method are 1.15±1.35 mm and 3.47±3.38 mm with the p-value of 0.007. The results corresponding to the proposed method not only show better localization error but also show lower miss rate. Sample landmark detection results are also shown in Fig.5.6.

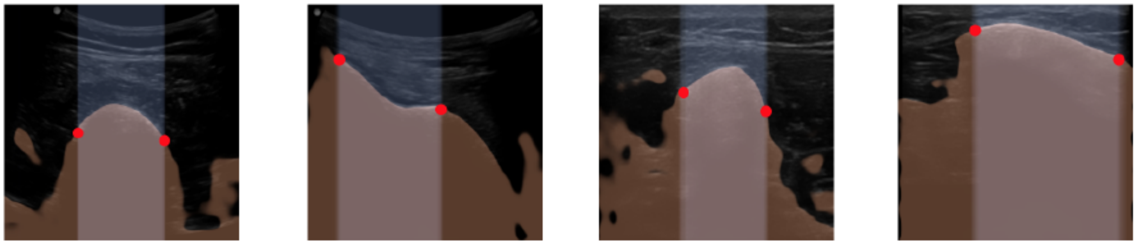


Figure 5.6: Landmark detection results. The input US scan is overlaid with the estimated BSE and HBIM. Red dots correspond to the detected bone landmarks.

5.4.3 Bone Surface Localization

One main application of bone shadow segmentation is bone surface localization from bone shadow in which accurate and robust localization is important for the improved guidance in US-based CAOS procedures. In this experiment, we applied raycasting method to perform bone surface localization from bone shadows.

The Average Euclidean Distance (AED) results (mean+std) in Table 5.1 show that the proposed PSPGAN-MTL outperforms the other methods on test scans by a large margin. Note that the bone surface localization experiment was carried out using previous bone shadow segmentation results for all methods. Therefore, the networks are not trained specificly on the bone surface localization task. A further paired t-test between PSPGAN-MTL and PSPGAN at a 5% significance level with p-value of 0.0009 clearly indicates that the improvements of our method are statistically significant.

5.5 Summary

In this chapter, we proposed an end-to-end deep learning framework that enabled robust and accurate bone shadow segmentation for bone ultrasound examination. The main novelty lies in (1) the introduction of conditional shape discriminator to shape specific image segmentation problem, (2) the design of two subtasks, coarse bone shadow enhancement and horizontal bone interval mask to improve the performance of each task and (3) the integration of the highly-related homogeneous tasks into a single unified bone shadow segmentation network. Formulating the network with a single powerful encoder based on Resnet50 and two pyramid pooling decoders, the proposed network brings strong synergy across all tasks when extracting shared deep features.

Chapter 6

Pyramid Convolutional RNN for MRI Image Reconstruction

6.1 Introduction

In this chapter, we propose a pyramid convolutional RNN (PC-RNN) model for MRI reconstruction, which learns multi-scale features. We formulate the MRI reconstruct as an inverse problem and solve it by iteratively optimizing a regularized objective function. The optimization is learned by three convolutional RNN (ConvRNN) modules at different scales and the reconstructed images in coarse to fine scales are combined by a final CNN module in a pyramid fashion. The proposed model can also be interpreted as learning multi-scale image priors. Each ConvRNN module features a downsample convolution block with different downscale rates, a residual convolution block as the recurrent unit and a deconvolution block to recover the original image size. The output from the recurrent residual convolution block acts as the hidden state and passes into the next iteration. Each ConvRNN and final CNN have data consistency layers to enforce the data consistency at each scale. Our method can recover fine details while preserving data consistency.

For multi-coil data generated from parallel MRI, it is common to utilize CSM and combine multiple coils to generate a combined image. Several DL-based approaches have been proposed to handle the multi-coil data either by using pre-computed CSM or learning CSM during the construction. However, we found that using CSM is slow since it takes time to calculate those maps. This is especially true for models that learn the maps, which includes separated networks for CSM estimation [53,54]. Such models explicitly using CSM are often very large models due to the extra components for CSM and thus are less efficient for GPU memory and training time. On the other hand, the reconstruction quality relies on the accuracy of the estimated CSM. Inaccurate CSM often leads to artifacts in the final combined image [55]. Therefore, to avoid the above limitations, our model takes the multi-coil data as multi-channel input without explicitly using CSM. For data with different numbers of coils, we use coil compression [56] to standardize the coil dimension.

In summary, our main contributions are:

- We propose a novel pyramid ConvRNN model to learn the optimization process for MRI reconstruction.
- To best of our knowledge, this is the first work proposing a DL model with pyramid architecture to explicitly reconstruct MRI images in multiple scales.
- We propose to model multi-coil MRI data without CSM and apply coil compression to standardize the coil dimension.
- We evaluate our model on the fastMRI knee and brain datasets. Our proposed method is one of the winner solutions in the 2019 fastMRI competition.

6.2 Related work

In this section, we will review DL-based MRI reconstruction methods. Please refer to [28] for CS-based approaches. Most DL approaches borrow the basic ideas from CS. Some approaches use neural networks to represent the regularization functions used in CS. Such functions parameterized by neural networks are learned from training data. For example, Tezcan *et al.* [57] applied a VAE model pre-trained on patches of fully sampled images as the prior. Aggarwal *et al.* [58] proposed a CNN-based prior for MRI reconstruction. GANCS [59] was proposed to learn the manifold of high-quality images and then project the aliased images onto the learned manifold to remove the artifacts. Yang *et al.* [60] proposed a conditional GAN model to learn the image distribution from large training data. He *et al.* [37] proposed a model based on denoising autoencoder to learn the high-frequency component as the prior.

On the other hand, some methods employ neural networks to learn the optimization algorithms used in CS. For example, ADMM-Net [61] substitutes the operations in the ADMM optimization framework by neural network layers, in which parameters can be learned during training. Similarly, ISTA-Net [62] replaces all the manually designed parameters in the ISTA algorithm with neural networks. VS-Net [63] decomposes the optimization problem into several sub-problems using the variable split method and solves each sub-problem with a neural network. Chen *et al.* [64] proposed a similar method based on the split Bregman iterative algorithm. Hammernik *et al.* [65] generalized optimization formulation in CS with a variational network.

Since the optimization algorithms are often iterative, many DL models utilize similar architectures to iteratively update the reconstruction results. The most common iterative models are cascaded networks [54,58,66–74] and RNN models [75]. Recently, Putzky *et al.* introduced an iterative inverse model (i-RIM) based on invertible networks and applied the model to MRI reconstruction [76]. The neural ODE model has also been proposed to model the continuous optimization trajectory in MRI reconstruction as solving ODE equations [77]. The data consistency layers [66] are often adopted into the iterative network design to enforce the data fidelity during optimization.

Some DL approaches perform MRI reconstruction by learning the direct mapping between undersampled images and fully sampled images [78]. For instance, [79,80] proposed to learn this mapping in k-space domain. The recently proposed AUTOMAP [81] learns the mapping between k-space and image domains using fully connected layers, which may consume a lot of GPU memory for large images. dAUTOMAP [82] was later proposed to improve the computational burden by replacing the fully connected layers with convolutional layers.

In the case of parallel MRI, one strategy is to combine the multi-coil undersampled images and feed the network with coil combined images for reconstruction [83]. Another strategy is to include CSM along with multi-coil data into the model. For example, VS-Net pre-computes CSM from undersampled data and includes them in the reconstruction [63]. Blind-PMRI-Net jointly learns the image prior and explicitly estimates CSM with an iterative reconstruction algorithm [84]. E2E-VN includes two networks, where one network combines multi-coil data using CSM and refines the combined image, and the other network estimates CSM used in the refinement network [53]. Chen *et al.* introduced a filter operator on the multi-coil data, which is implemented by convolutional layers, and performs reconstruction without explicitly calculating CSM [64]. Σ-Net introduces a strategy that ensembles the reconstructions from two networks: the sensitivity network, which combines coils explicitly using pre-computed CSM, and the parallel coil network, which implicitly learns coil combination [85].

6.3 Proposed Method

In this section, we will first formulate MRI reconstruction as an inverse problem, which can be solved as an optimization problem using the gradient descent algorithm. We will then show that the gradient descent can be learned by a ConvRNN model. Based on this formulation, we will introduce our proposed model, which features three ConvRNN modules that learn the optimization at multiple scales in a pyramid fashion. Our model can also be considered as learning multi-scale image priors.

6.3.1 MRI reconstruction as an inverse problem

The MRI data acquisition can be formulated as follows:

$$y = Ax + \epsilon \ , \tag{6.1}$$

where $x \in \mathbb{C}^M$ is the image to reconstruct, $y \in \mathbb{C}^N$ is the undersampled k-space, and ϵ is the noise. A is the forward operator and often the multiplication of the Fourier transform matrix $\mathcal{F}$, the binary undersampling matrix D and coil sensitivity matrix S. In our work, we do not use CSM and ignore S in the following chapter. The goal of MRI image reconstruction is to estimate image x from observed k-space y. MRI reconstruction can be considered as an inverse problem, in which the inverse process is ill-posed due to the information loss in the forward process as $N < M$. A regularization term $R(x)$ needs be added to the objective function:

$$\underset{x}{\operatorname{argmin}} \frac{1}{2}||y - Ax||_2^2 + R(x). \tag{6.2}$$

In CS, $R(x)$ takes the form of $||\Psi x||_1$, where Ψ is the transformation matrix. This $\ell 1$ term forces x to be sparse in the transformed domain. In DL, this regularization function $R(x)$ can be learned from data.

6.3.2 Learning optimization

As long as $R(x)$ is differentiable, we can minimize the objective function in Eqn. 6.2 using gradient decent in an iterative fashion,

$$\hat{x}^{(k+1)} = \hat{x}^{(k)} + \alpha[A^T(y - A\hat{x}^{(k)}) + \nabla R(\hat{x}^{(k)})], \tag{6.3}$$

where $k = 0, 1, \dots, K$ is the index for iteration and α is the learning rate.

The hand-crafted iterative optimization process such as the above gradient descent update rule can be learned by deep learning models (for example, LSTM [86], deep reinforcement learning [87] or neural ODE [77] models). Based on this, we utilize a ConvRNN model to learn Eqn.7.2 as

$$\hat{x} = G(\tilde{x}; \theta), \tag{6.4}$$

where $\tilde{x}$ is the input image (e.x., undersampled image) and θ are the model parameters. Note that our proposed ConvRNN are specifically designed for multi-scale learning (see next section), which is different from the previously published ConvRNN model [75]. We omit the parameter notation in the following equations for simplicity.

6.3.3 Learning priors

The network architecture itself introduces inductive bias and can be interpreted as a prior for inverse problems since it captures some statistics of images [88]. The reconstructed image is the output of the ConvRNN model and therefore it lies in the manifold constraint by the model structure [89]. By training the ConvRNN model, we can learn this manifold from data. Therefore the model in Eqn. 6.4 can be interpreted as learning both the optimization and

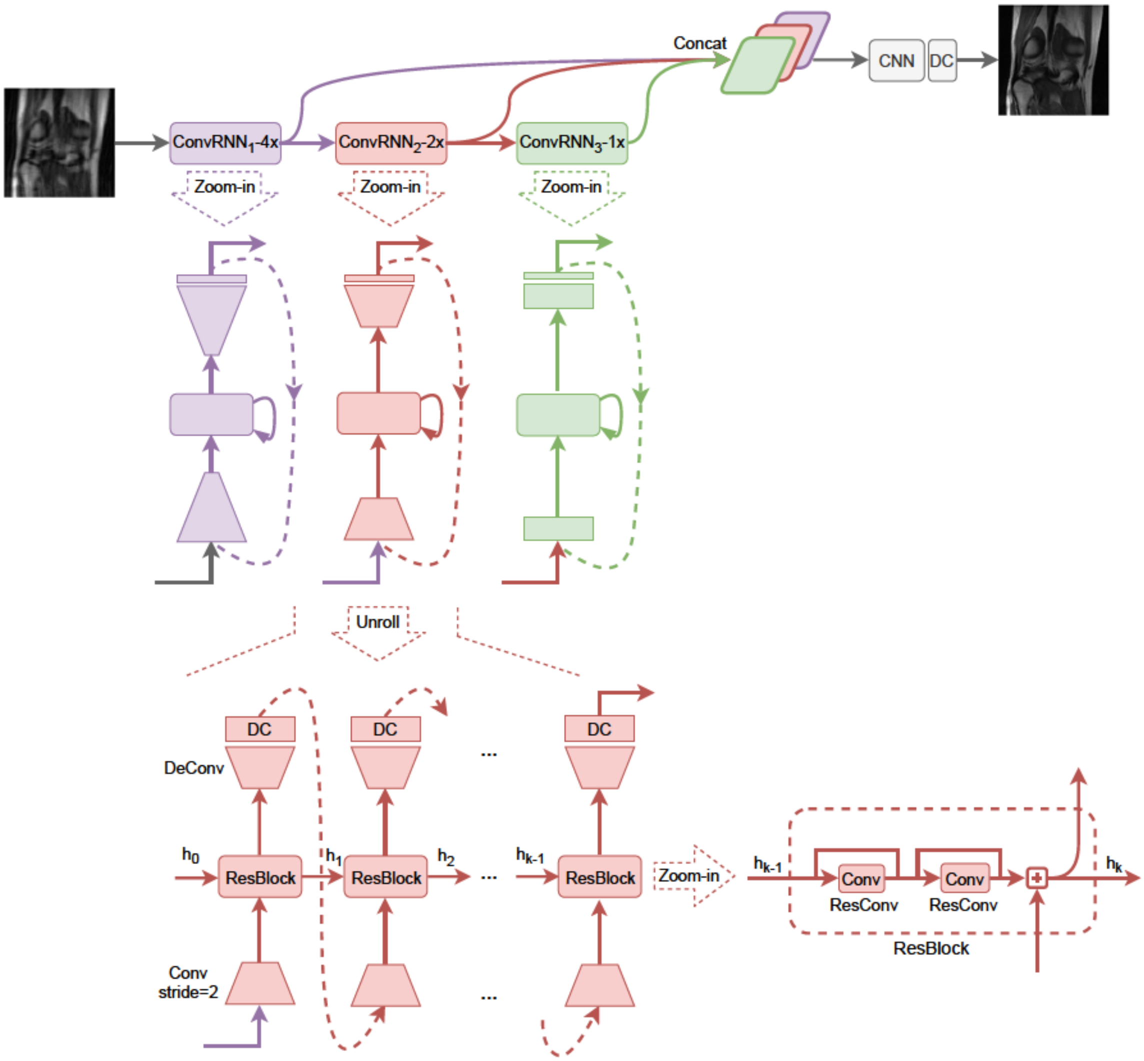


Figure 6.1: The illustration of the proposed Pyramid Convolutional RNN (PC-RNN) model for MRI reconstruction. The model includes three convolutional RNN modules to iteratively reconstruct images at different scales. The reconstructed images are combined by a final CNN module.

the prior.

6.3.4 Pyramid Convolutional RNN

To improve the reconstruction on fine details, we extend the above idea and propose to reconstruct the image at multiple scales. Consider the final reconstructed image $\hat{x}$ as the combination of images $\hat{x}_s$ in different scales.

$$\hat{x} = f(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_S) \ , \tag{6.5}$$

where $\hat{x}_s$ indicates reconstructed image at scale s ($s = 0, 1, \ldots, S$ for total S scales) and f is a function to integrate images from different scales. Large s indicates fine scale.

Then we can reconstruct each $\hat{x}_s$ as:

$$\hat{x}_s = G_s(\hat{x}_{s-1}) \ , \tag{6.6}$$

where $\hat{x}_0 = \tilde{x}$ is the input image and g_s is the function for reconstruction. Each G_s is specialized in modeling the signal at a certain scale s.

The rationale is that we decompose the optimization problem of reconstructing original images into several sub-problems, in which images at different scales are reconstructed separately, from coarse to fine scales, and then combined as the final reconstructed image. In this way, the searching space of the optimization in each sub-problem learned by G_s (Eqn. 6.6) is smaller and it is easier to find better solutions than the original optimization learned by G (Eqn.6.4). Our multi-scale model can also be interpreted as learning separate priors for low, middle and high-frequency information. The explicit high-frequency prior can help reconstruct fine details.

We implement the idea based on Eqn. 6.4,6.5 and 6.6 and propose a novel network Pyramid Convolutional RNN (PC-RNN) as shown in Figure 6.1. It features three ConvRNN modules to model data in different scales, which corresponds to ConvRNN$_1$-4x, ConvRNN$_2$-2x and ConvRNN$_3$-1x in the upper panel of Figure 6.1,

$$\hat{x}_1 = G_1(\hat{x}_0, y, D) \tag{6.7}$$

$$\hat{x}_2 = G_2(\hat{x}_1, y, D) \tag{6.8}$$

$$\hat{x}_3 = G_3(\hat{x}_2, y, D), \tag{6.9}$$

where y and D are included for data consistency.

We apply a CNN module with 4 convolutional layers to combine the three reconstructed images $\hat{x}_1, \hat{x}_2, \hat{x}_3$ and derived the final reconstruction $\hat{x}$:

$$\hat{x} = f(\hat{x}_1, \hat{x}_2, \hat{x}_3, y, D). \tag{6.10}$$

To ensure the data fidelity in the original objective function (Eqn. 6.2) when learning the optimization, the data consistency layers are added into each iteration of ConvRNN modules and the CNN module as

$$\sigma(\hat{x}, y, D) = \mathcal{F}^{-1}[Dy + (1 - D)\mathcal{F}\hat{x})]. \tag{6.11}$$

The details of each ConvRNN module G_s are shown in the middle panel of Figure 6.1, which consists of four components: an encoder g_s^{enc}, an decoder g_s^{dec}, a basic RNN cell (ResBlock) g_s^{res}, and a data consistency layer σ. The output of $(k+1)^{th}$ iteration of ConvRNN module G_s can be derived as follows:

$$\hat{x}_s^{(k+1)} = \sigma(g_s^{dec}(g_s^{res}(h_s^{(k)}) + g_s^{enc}(\hat{x}_s^{(k)}))), \tag{6.12}$$

where $h_s^{(k)} = g_s^{res}(h_s^{(k-1)}) + g_s^{enc}(\hat{x}_s^{(k-1)})$ is the hidden state from previous iteration and $h_s^{(0)} = 0$.

The bottom panel of Figure 6.1 shows the details of each component in ConvRNN. To ensure each ConvRNN extract features at different scales, the spatial sizes of feature maps are downsampled by 4x, 2x, 1x, respectively, using convolutional layers with strides=2 in encoders. The encoder g_1^{enc} in ConvRNN$_1$-4x module includes two convolutional layers with stride=2, which leads to coarse reconstruction at the 4x scale. The ConvRNN$_2$-2x module has one convolutional layer with stride=2 and one with stride=1 in the encoder, which results in reconstruction at the 2x scale. The last ConvRNN$_3$-4x module has two convolutional layers with stride=1 in the encoder and reconstructs images at the 1x scale. The decoders in each ConvRNN use transposed convolutional (deconvolutional) layers to recover the original image size for the final combination. The ResBlock g_s^{res} includes two residual convolutions (ResConv) and acts as a recurrent unit.

6.3.5 Multi-coil data modeling

For multi-coil data, the model takes the stack of all coils as a multi-channel input without using CSM. The network outputs the reconstructions for all coil images. The final reconstruction is obtained by combining multi-coil images using the root sum squared (RSS) method,

$$\hat{x}_{\text{rss}} = \sqrt{\sum_{c=1}^{n_c} |\hat{x}_c|^2}\,, \tag{6.13}$$

where n_c is the number of coils. Since the training loss is added onto the combined image instead of the individual coil image, the model learns to weight each coil in an optimal way such that the final combined image matches the ground truth.

6.3.6 Coil compression

Our PC-RNN model takes inputs with a fixed number of coils. But the coil numbers in the fastMRI brain dataset are different. To solve this problem, we applied the coil compression technique [56] to standardize the coil dimension in the brain dataset. Since the coil numbers are the same in the knee dataset, we did not employ coil compression on the knee dataset although the methods are still readily applicable.

To perform the coil compression, a $n_{calib} \times n_{calib}$ central region of the k-space of every coil representing low-spatial-frequency component $y_{calib} \in \mathbb{C}^{n_{calib}^2 \times n_c}$ is used as the calibration data. The calibration data is factorized using singular value decomposition and the first n_{vc} columns of the right-singular vectors are kept to form a compression matrix $M_c \in \mathbb{C}^{n_c \times n_{vc}}$. The acquired n_c-coil data is then compressed to n_{vc} virtual coils through $y_{comp} = yM_c$.

6.4 Experiments

6.4.1 Datasets

We used the knee and brain datasets from fastMRI competition [83]. The knee dataset includes single-coil and multi-coil tasks with 973 volumes (34,742 slices) for training and 199 volumes (7,135) for validation. The brain dataset only includes the multi-coil task with 4,469 volumes (70,748 slices) for training and 1,378 volumes (21,842 slices) for validation. The fully sampled k-space data are available in both training validation data. Only subsampled k-space data (i.e., no fully sampled data) are provided in the challenge dataset and the reconstruction results can be evaluated through the fastMRI website[1]. For the knee dataset, we used the ground truth images provided by fastMRI. However, for the brain dataset, we used the images after the inverse Fourier transform of the k-space as the ground truth since some of the ground truth images are ill-formatted.

6.4.2 Training

We combined the Normalised Mean Square Error (NMSE) loss [83] and the Structural Similarity Index (SSIM) loss [90] on the coil combined images for training,

$$\mathcal{L}(\hat{x}, x) = \mathcal{L}_{\text{NMSE}} + 0.5\mathcal{L}_{\text{SSIM}}. \tag{6.14}$$

For both single-coil and multi-coil tasks, we trained models with the same network architecture except for the input and output dimensions and the number of feature maps of each module. The number of iterations for all ConvRNN is set to 5. All training images were center cropped to 320×320 if possible and normalized by dividing the mean of the undersampled image. The real and imaginary values of input data are split into two channels. We used the looka-

[1] https://fastmri.org/leaderboards/challenge

head version of Adam optimizer [91].The learning rate was set to 10^{-5} for the training warmup and increased to 10^{-4}, which was then reduced by a factor of 2 every 10 epochs. The network was trained for 60 epochs. We applied the random sampling to the knee dataset and the equispaced sampling to the brain dataset [83]. Note that we did not use the validation dataset for additional training as in [53], which may improve the leaderboard results.

For comparison, we also reconstructed the same data using CS and U-Net as in [83]. The code with default parameters from fastMRI github[2] was used. We trained separated models for 4X and 8X acceleration in each task.

6.4.3 Evaluation

We calculated PSNR and SSIM on the validation data for comparison and used the Wilcoxon signed-rank test to calculate pvalues. The results of the knee challenge dataset were evaluated on the fastMRI website. Note that the single-coil 8X task was not included in the competition and therefore results are not evaluated on the challenge leaderboard. The brain challenge leaderboard is not open for submission.

To determine the best reconstruction in the fastMRI competition, the submission results on the challenge dataset were first ranked by SSIM and the top four results were then evaluated by seven expert radiologists based on the

[2] https://github.com/facebookresearch/fastMRI

Table 6.1: Evaluation results of CS, U-Net and PC-RNN on fastMRI knee validation dataset.

			PSNR		SSIM	
Task	Sequence	Method	4X	8X	4X	8X
Knee Single-coil	PD	CS	31.4	27.6	0.645	0.562
		U-Net	33.9	31.2	0.812	0.753
		PC-RNN	**35.2**	**33.3**	**0.835**	**0.787**
	PDFS	CS	27.7	26.4	0.493	0.406
		U-Net	29.9	28.6	0.633	0.554
		PC-RNN	**30.3**	**29.3**	**0.651**	**0.577**
	All	CS	29.5	27.0	0.570	0.484
		U-Net	31.9	29.9	0.723	0.654
		PC-RNN	**32.8**	**31.3**	**0.743**	**0.682**
Knee Multi-coil	PD	CS	32.6	29.6	0.675	0.639
		U-Net	37.3	33.5	0.925	0.881
		PC-RNN	**40.4**	**35.8**	**0.950**	**0.921**
	PDFS	CS	27.5	26.8	0.588	0.549
		U-Net	36.2	33.8	0.863	0.822
		PC-RNN	**37.5**	**37.3**	**0.878**	**0.847**
	All	CS	30.1	28.2	0.632	0.595
		U-Net	36.7	33.7	0.894	0.852
		PC-RNN	**39.0**	**36.5**	**0.914**	**0.884**

following categories: contrast to noise ratio, artifacts, sharpness, diagnostic confidence, and overall image quality. The final rating is the average score from all radiologists.

Table 6.2: Evaluation results of CS, U-Net and PC-RNN on fastMRI brain validation dataset.

			PSNR		SSIM	
Task	Sequence	Method	4X	8X	4X	8X
Brain Multi-coil	T1	CS	31.2	29.0	0.496	0.483
		U-Net	38.1	35.1	0.939	0.915
		PC-RNN	**41.1**	**38.6**	**0.953**	**0.939**
	T2	CS	30.2	26.1	0.551	0.476
		U-Net	35.5	31.9	0.930	0.897
		PC-RNN	**39.3**	**36.3**	**0.954**	**0.935**
	FLAIR	CS	28.8	25.9	0.447	0.380
		U-Net	35.8	32.9	0.894	0.858
		PC-RNN	**39.1**	**36.8**	**0.930**	**0.910**
	T1POST	CS	33.0	29.7	0.543	0.536
		U-Net	38.5	35.5	0.949	0.927
		PC-RNN	**41.7**	**38.8**	**0.966**	**0.950**
	All	CS	30.8	27.2	0.534	0.482
		U-Net	36.5	33.1	0.932	0.902
		PC-RNN	**40.0**	**37.1**	**0.954**	**0.937**

Table 6.3: Top results on fastMRI knee challenge dataset. * indicates the final winning solutions evaluated by radiologists.

		PSNR		SSIM	
Task	Method	4X	8X	4X	8X
Knee Single-coil	**i-RIM [76]***	**33**	NA	**0.754**	NA
	Adaptive-CS-Net [54]	33	NA	0.751	NA
	Masked-DCN	33	NA	0.751	NA
	Inverse spread	32	NA	0.750	NA
	Σ-Net [85]	32	NA	0.750	NA
	PC-RNN	33	NA	0.747	NA
Knee Multi-coil	**Adaptive-CS-Net [54]***	**40**	**37**	**0.927**	**0.902**
	Auto-calibrating deep learning	40	37	0.928	0.901
	Σ-Net [85]	40	37	0.927	0.899
	i-RIM [76]	39	37	0.925	0.899
	PC-RNN*	**40**	**37**	**0.927**	**0.897**
	Dense Head UNet	39	37	0.924	0.897

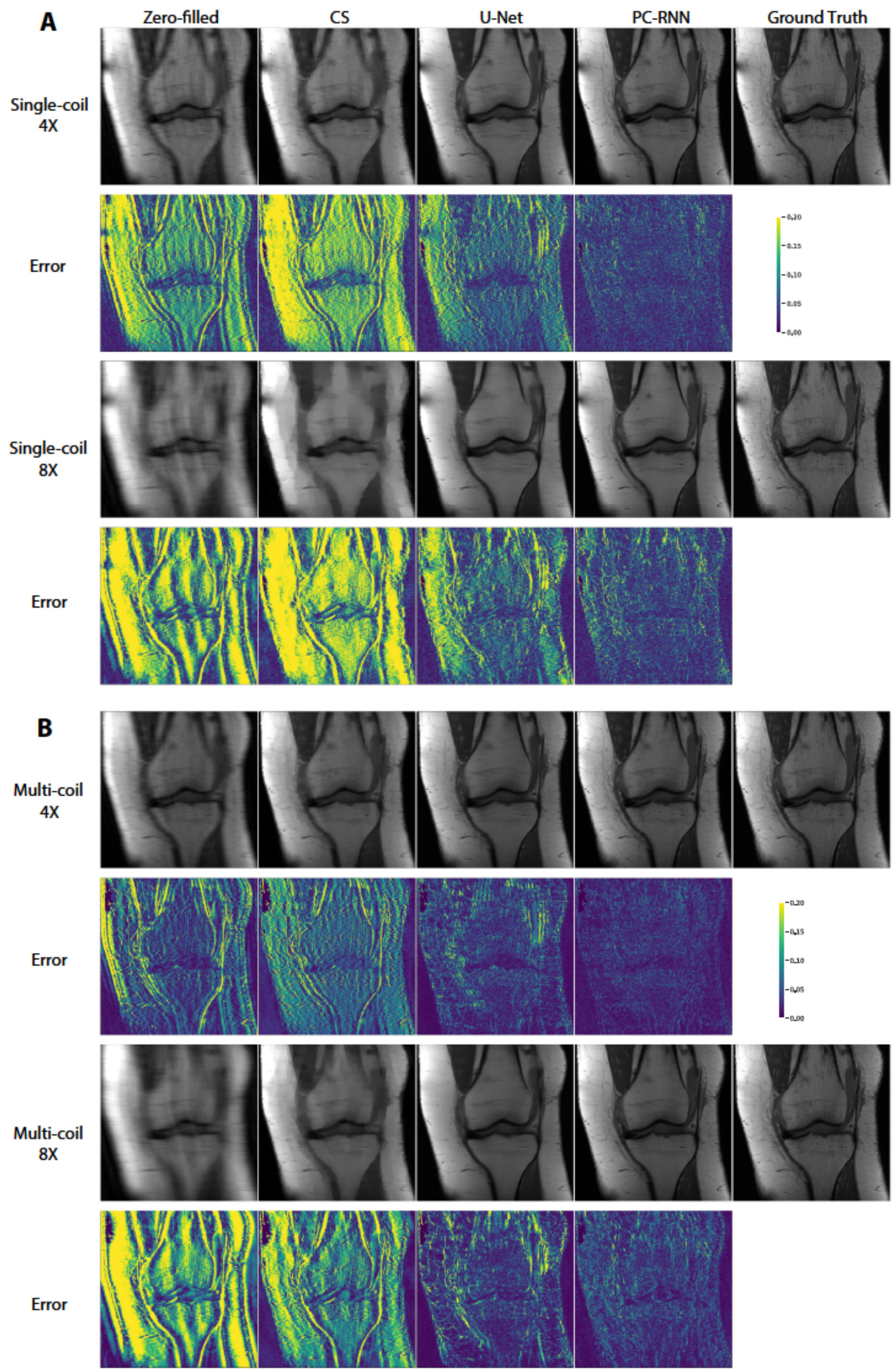


Figure 6.2: Examples of reconstruction and error maps of knee images for (A) the single-coil task and (B) the multi-coil task.

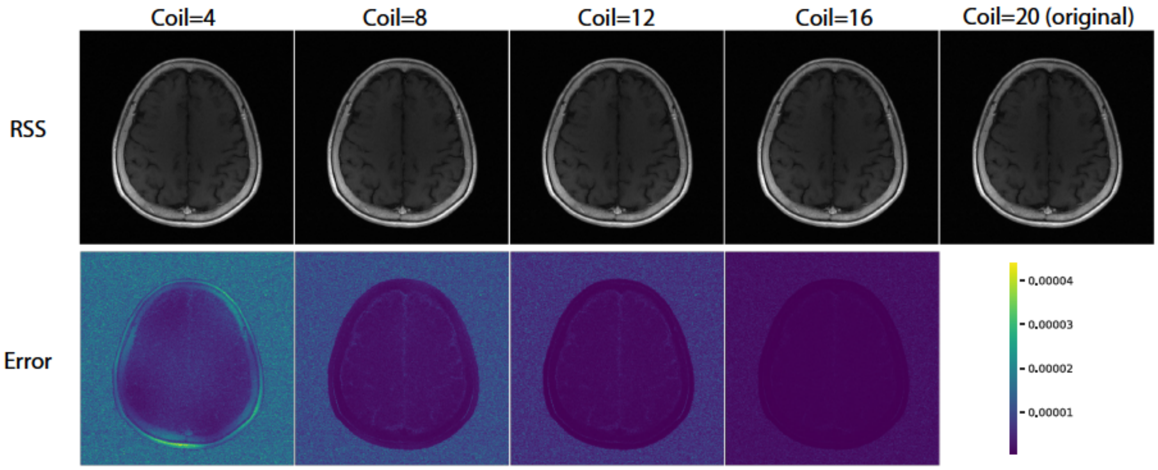


Figure 6.3: An illustration of the coil compression with different numbers of compressed coils. The multi-coil data after coil compression are combined with root-sum-of-square (RSS). The error maps indicate the difference between the coil compressed image and the original image.

6.5 Results

6.5.1 fastMRI knee dataset

We first trained and evaluated our proposed model on the fastMRI knee dataset and compared our model with CS and U-Net. Table 6.1 shows the results evaluated on the validation data. In the single-coil task, PC-RNN outperforms CS by 3.3 and U-Net by 0.9 in PSNR at 4X acceleration. At 8X acceleration, our model improves PSNR by 4.3 and 1.4 compared with CS and U-Net, respectively. In the multi-coil task, the improvement is more significant. Comparing PC-RNN to the other two methods, PSNR is boosted by 8.9 and 2.3 at 4X acceleration as well as 8.3 and 2.8 at 8X acceleration. SSIM also

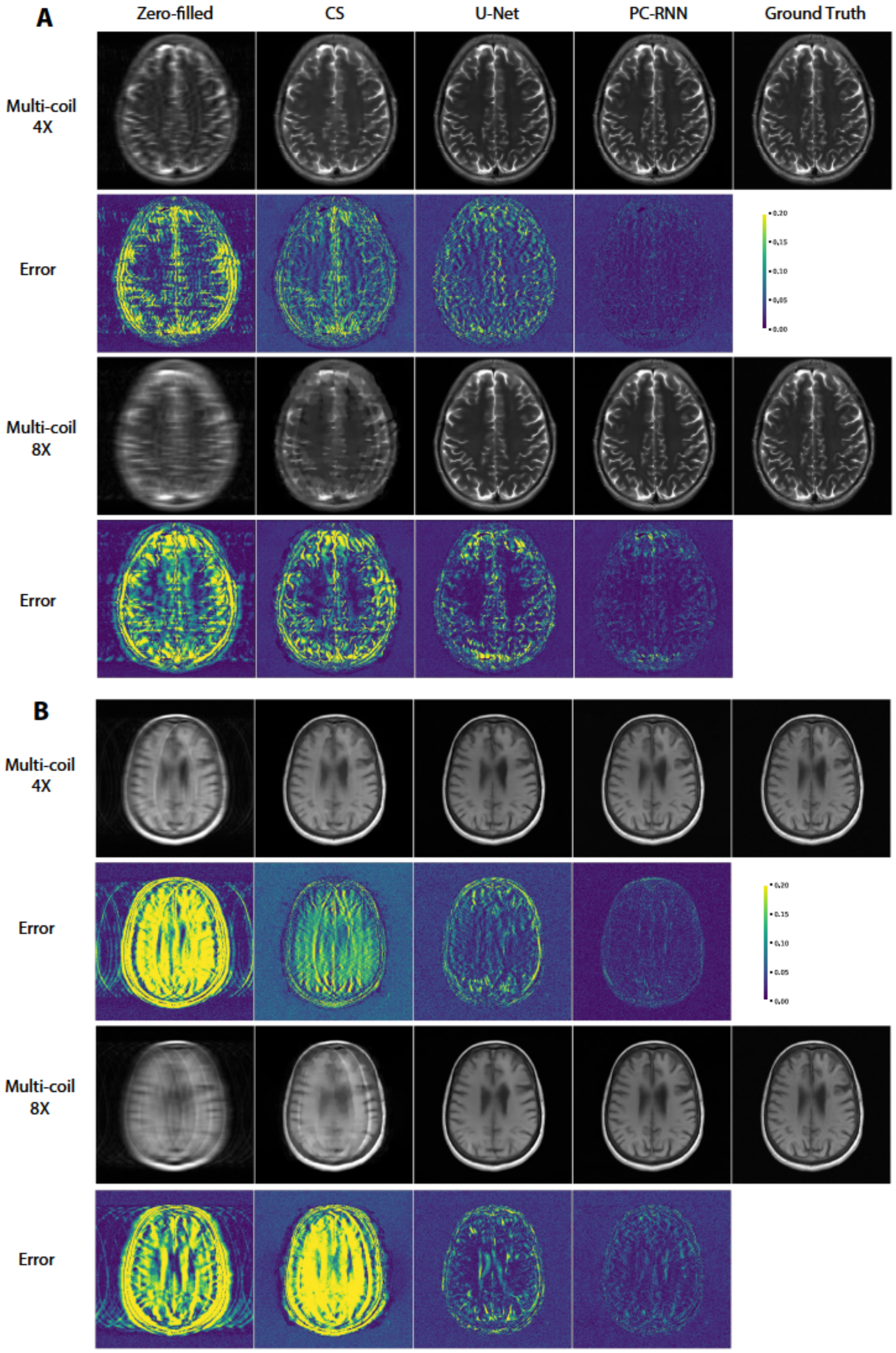


Figure 6.4: Examples of reconstructed brain images and error maps for (A) T1-weighted and (B) T2-weighted MRI. Examples of other sequences are omitted due to the space limit.

shows consistent improvement results ($p < 10^{-5}$).

Figure 6.2 demonstrates examples of reconstructed knee images by PC-RNN and other methods. Our model recovers more details, especially using multi-coil data. At 8X high acceleration, our model can reconstruct considerably more fine details than the other two methods.

Table 6.3 shows the performance of top models on the challenge dataset. The top results have very small differences in terms of PSNR and SSIM. The reconstructed images from top models were then evaluated and scored by seven radiologists from the clinical perspective. Our results ranked as one of the best in the multi-coil 4X task by radiologists' assessment [92].

6.5.2 fastMRI brain dataset

To show the generalizability of our proposed method, we also trained and evaluated the PC-RNN model on the fastMRI brain dataset. Unlike the knee dataset where all data have 15 coils, the brain dataset contains data with various numbers of coils, which ranges from 2 to 28 coils. We applied coil compression to standardize the coil numbers. In order to find the best number of compressed coils, we experimented with different compressed coils and evaluated coil combined images with those from non-compressed coils. When the number of compressed coils is 4, the SSIM of coil-compressed images against non-coil compressed images reduces to 0.931. However, visual examination shows that the information loss is mainly in the background and boundary regions, which is consistent with the previous report that coil compression mainly removes the background noise [56]. Note that the errors showed Figure 6.3 are very small even for 4 compressed coils (the max error is around 4×10^{-5}, which is much smaller than the reconstruction error). Since the brain dataset is much larger than the knee dataset and takes a much longer time to train, we decide to choose 4 compressed coils. The evaluation shows the reconstruction with 4 compressed coils achieves reasonably good results.

Table 6.2 shows the evaluation of reconstruction results on the brain validation dataset. Similarly, PC-RNN shows significant improvements over CS and U-Net ($p < 10^{-5}$). PSNR from PC-RNN is improved by 9.2 and 9.9 over CS as well as 3.5 and 4.0 over U-Net for 4X and 8X, respectively. CS has poor SSIM values for both 4X and 8X. U-Net and PC-RNN have much better SSIM values, while PC-RNN achieves the highest SSIM values. These results indicate that our proposed method can be applied to different MRI datasets and outperforms traditional methods.

6.5.3 Multi-scale learning in PC-RNN

To demonstrate the effectiveness of PC-RNN in learning multi-scale features, we show the intermediate outputs from each ConvRNN as well as the final reconstructed image and compared them with the ground truth in Fig-

Table 6.4: Evaluation of outputs from each ConvRNN module of PC-RNN on fastMRI knee and brain validation dataset

		PSNR		SSIM	
Task	Output	4X	8X	4X	8X
Knee Multi-coil	ConvRNN$_1$-4x	29.9	26.3	0.755	0.641
Knee Multi-coil	ConvRNN$_2$-2x	35.2	31.8	0.868	0.801
Knee Multi-coil	ConvRNN$_3$-1x	25.6	22.3	0.675	0.565
Knee Multi-coil	PC-RNN	**39.0**	**36.5**	**0.914**	**0.884**
Brain Multi-coil	ConvRNN$_1$-4x	30.6	23.9	0.757	0.578
Brain Multi-coil	ConvRNN$_2$-2x	33.1	26.8	0.847	0.681
Brain Multi-coil	ConvRNN$_3$-1x	33.1	28.0	0.866	0.712
Brain Multi-coil	PC-RNN	**40.0**	**37.1**	**0.954**	**0.937**

ure 6.5. From ConvRNN$_1$-4x to ConvRNN$_3$-1x, more and more fine structures are recovered in the reconstructed images. Interestingly, the outputs of ConvRNN$_1$-4x show smoothed reconstructions, while the outputs of ConvRNN$_3$-1x depicts sharpened reconstructions and some of the details are enhanced. The final CNN in PC-RNN combines all three reconstructed images and outputs a clear and sharp MRI image, which is more similar to the ground truth image. Table 6.4 compares the intermediate outputs against the ground truth on the knee and brain validation datasets. It is interesting to observe that each ConvRNN generates a relatively low-quality image but the final reconstructed image based on those intermediate images achieves much better results. This is because each ConvRNN reconstructs images in different scales and the final reconstruction utilizes the multi-scale features.

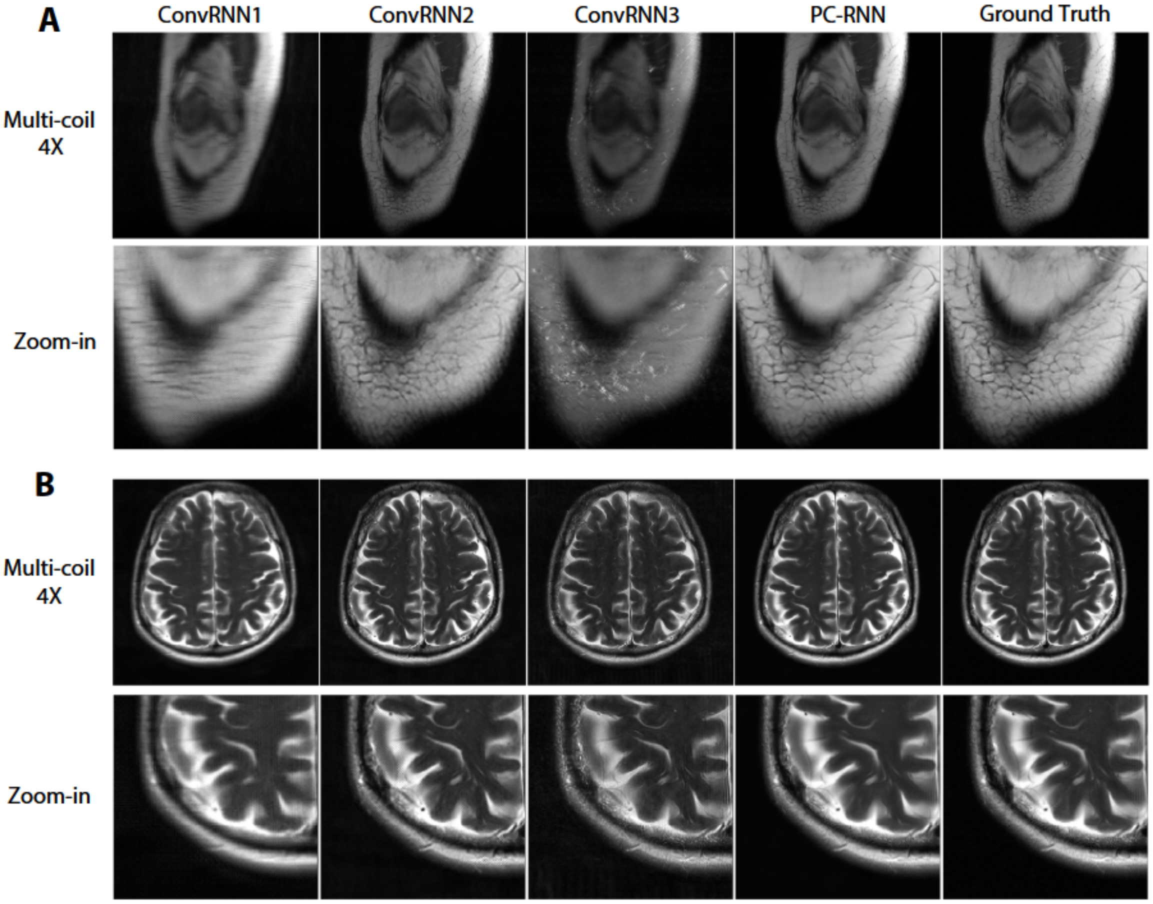


Figure 6.5: Intermediate and final outputs of PC-RNN for (A) knee images and (B) brain images. Outputs from ConvRNN_1, ConvRNN_2 and ConvRNN_3 show that PC-RNN learns to reconstruct images from coarse to fine scales. The details in the output of ConvRNN_3 are enhanced. The final reconstruction that combines the multi-scale reconstructions depicts a clear and sharp image, which is similar to the ground truth.

6.6 Discussion

The MRI reconstruction results are often evaluated by PSNR and SSIM, which are commonly used metrics for natural images. However, MRI images are specific for medical diagnosis and the reconstruction requires not only sharpness but also rigid data fidelity. Thus, PSNR and SSIM alone may not be optimal metrics to compare different MRI reconstruction methods. In the fastMRI competition, the final results were evaluated by seven radiologists with clinical standards. Our results are judged by the radiologists as one of the best results for the multi-coil 4X task. This is the most clinically relevant task in the competition since single-coil acquisition is rarely used in clinical settings and 8X acceleration is too high to have the risk of misdiagnosis. This indicates that our reconstruction results not only show superior image quality but also meet the clinical requirements better than other methods from a practical perspective.

The benefits of coil compression are three folds. First, it removes the background noise [56]. Second, it standardizes the coils since after SVD, the compressed coils are the principal components in the coil dimension and the compressed coils are ordered by their corresponding eigenvalues. Thus, the original coil configuration is less relevant and the relationship of compressed coils is easier to model. Third, it speeds up the reconstruction and dramatically reduces data storage. By comparing 15 and 4 compressed coils, the training GPU memory and time decrease from 9.7G and 0.77 sec/sample to 6.8G and 0.53 sec/sample, respectively. The size of the brain dataset is reduced from 1.9T to 0.4T.

The coil compression process can be considered as a pre-learned convolutional layer with a 1×1 kernel. Each column of the compression matrix is a filter and there are total n_{vc} filters, which convert the input n_c-channel data to n_{vc}-channel features. Therefore, the coil compression procedure can be learned and integrated into the deep learning model.

The major difference between our method and other published methods is that our model learns the optimization and priors from different scales. The specifically designed network architectures correspond to good hand-crafted priors [88]. The PC-RNN architecture forces the model to learn the high-frequency prior separately from low-frequency prior, which leads to better recovery of fine details than other methods. In the current model, we only implemented three scales (1x, 2x, and 4x). The reconstruction can be further improved by including more scales in the proposed framework. Unlike Σ-Net that each component needs to be pre-trained separately to avoid large GPU memory consumption [85], our model can be trained end-to-end. PC-RNN has relatively fewer parameters (24M), compared to E2E-VN (30M) [53] and Adaptive-CS-Net (33M) [54]. Therefore, the performance of PC-RNN can be further boosted by increasing the model size.

6.7 Summary

In this chapter, we proposed a multi-scale pyramid convolutional RNN model for MRI image reconstruction. The model reconstructs the image in various scales and then combines the coarse-to-fine images as the final reconstructed image. We evaluated our model on the fastMRI knee and brain datasets. The results show that our model can reconstruct more fine details and outperforms other methods in fastMRI multi-coil 4X tasks. Although we demonstrated our proposed method on MRI image reconstruction, it can be extended to other imaging modalities such as CT image reconstruction. The basic idea of learning multi-scale priors for inverse problems as well as decomposing the optimization search space into several smaller ones can be generalized to other inverse problems such as image super-resolution or inpainting, which will be among our future work.

Chapter 7

Improving Amide Proton Transfer-weighted MRI Reconstruction using T2-weighted Images

7.1 Introduction

Amide Proton Transfer-weighted (APTw) imaging is an emerging molecular MRI method that can generate image contrast unique from the conventional MRI. As a type of chemical exchange saturation transfer (CEST) MRI, APTw signal intensity is based on concentrations of endogenous mobile proteins and peptides or tissue pH. Moreover, APTw MRI does not require any contrast agent administration. Previous studies in animals and humans have demonstrated that APT imaging is capable of detecting brain tumors [93] and ischemic stroke [94]. In a recent preclinical study [95], APT imaging was shown to accurately detect intracerebral hemorrhage and distinctly differentiate hyperacute hemorrhage from cerebral ischemia. Notably, the capability and uniqueness of APT imaging for the detection of primary and secondary brain injuries in experimental Controlled Cortical Impact (CCI) Traumatic Brain Injury (TBI) models have recently been explored with promising results [96].

However, relatively long acquisition times due to the use of multiple RF saturation frequencies and multiple acquisitions to increase the signal-to-noise ratio (SNR) hinders the wide spread clinical use of APTw imaging. A typical CEST MRI acquisition currently requires long scan times in the range of 5 to 10 minutes. Recently, several methods have been developed to accelerate CEST/APT acquisitions. These can be classified into conventional fast imaging techniques (e.g. turbospin-echo [97]) and reduced k-space acquisition techniques (including spectroscopy with linear algebraic modeling [98] and CS [99]) that require more advanced data processing. Due to recent advances in deep learning, deep learning-based methods have shown to provide much better generic MRI image reconstruction results from undersampled k-space data than conventional CS-based methods. The combination of convolutional autoencoder and generative adversarial networks can perform faster and more accurate reconstruction [100]. In [101], a pyramid convolutional RNN was designed to iteratively refine reconstructed image in three different feature scales.

Despite the success of deep learning-based MR image reconstruction methods for single contrast/modality imaging, multi-contrast reconstruction still remains a challenge. In multiple-contrast MR imaging it is beneficial to utilize fully sampled images acquired at one contrast for the reconstruction of undersampled MR images in another contrast [102]. For instance, information pertaining to undersampled T_1w images and undersampled T_2w images can be mutually beneficial when reconstructing both images. A joint reconstruction network of T_1, T_2 and PD images was proposed in [103] and was shown to outperform single-contrast models. Furthermore, undersampled T_2w image scan be reconstructed more accurately using the information from fully sampled high-resolution T_1w images [104]. To this end, Y-net was proposed in [104] by modifying U-net which takes two inputs and produces a single output. Features extracted from two independent encoders are concatenated together to generate the final output reconstruction. However, these methods are only evaluated on structural MR images and can be affected by slice mismatch between different scans. To deal with this issue, additional registration process between the images might be required.

Current 2D APTw imaging protocol starts with a high-resolution 3D T_2w scan that is used to locate the slice of interest (usually contains lesion region) by examination. After setting the interested slice, to reduce the effect of B_0 field inhomogeneity on APT imaging, high-order localized slab shimming is performed around the lesion. The final APTw image is defined as the difference of ±3.5 ppm image normalized by unsaturated image. While one can accelerate APTw imaging by reducing the raw k-space measurement data and apply reconstruction using off-the-shelf algorithms, no existing methods take 3D T_2w scan into reconstruction process as the idea of multi-contrast MR reconstruction suggests.

In this chapter, in order to leverage the structural information of T_2w images, we present a Recurrent Feature Sharing Reconstruction network (RFS-Rec) that has two convolutional RNNs (CRNN). These two CRNNs are connected by the proposed recurrent feature sharing approach to encourage bidirectional flow of information. In addition, we propose a sparse representation-based slice matching algorithm to find the corresponding slice in T_2w volume given the undersampled APT k-space data. As a result, input T_2w and APT raw images are aligned and mutual information can be maximized.

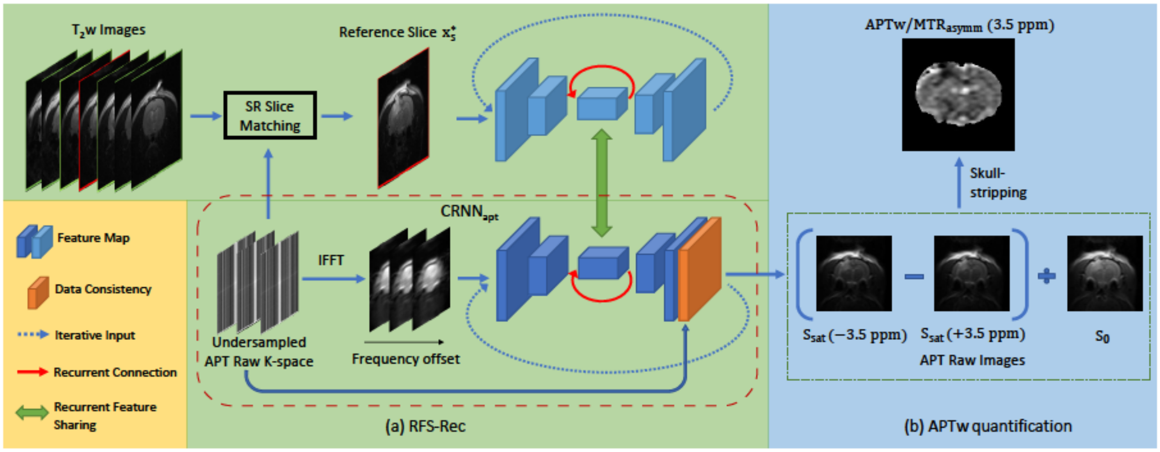


Figure 7.1: An overview of the proposed framework.

7.2 Methodology

In this section, we first give a brief introduction of APTw imaging. Then we describe our recurrent feature sharing reconstruction network and sparse representation (SR) based slice matching algorithm. As shown in Fig.7.1, the slice matching step in the proposed framework takes T_2w images and undersampled k-space as input and selects out a reference T_2w slice. The APT raw images are reconstructed by RFS-Rec using both reference T_2w slice and undersampled APT k-space data.

7.2.1 APTw Imaging

CEST effects are usually analyzed using Z-spectrum, in which the intensity of the water signal during saturation at a frequency offset from water, $S_{\text{sat}}(\Delta\omega)$, normalized by the signal without saturation S_0, is displayed as a function of irradiation frequency using the water frequency as a zero-frequency reference. The sum of all saturation effects at a certain offset is called the magnetization transfer ratio (MTR), defined as follows

$$MTR(\Delta\omega) = 1 - Z(\Delta\omega) = 1 - \frac{S_{\mathrm{sat}}(\Delta\omega)}{S_0}, \tag{7.1}$$

where $Z = S_{\mathrm{sat}}/S_0$ is the signal intensity in the Z-spectrum. As a type of CEST, APTw imaging is designed to detect the exchangeable amide protons in the backbone of mobile proteins and are assessed using magnetization transfer ratio asymmetry at 3.5ppm, namely $MTR_{\mathrm{asymm}}(3.5\mathrm{ppm})$ as APTw signal

$$\begin{aligned} APTw &= MTR_{\mathrm{asymm}}(3.5\mathrm{ppm}) \\ &= MTR(+3.5\mathrm{ppm}) - MTR(-3.5\mathrm{ppm}) \\ &= \frac{S_{\mathrm{sat}}(-3.5\mathrm{ppm}) - S_{\mathrm{sat}}(+3.5\mathrm{ppm})}{S_0}. \end{aligned} \tag{7.2}$$

Hence, the quality of APTw image solely depends on the above three images at different frequency offsets. An example of APTw quantification is shown in the right part of Fig. 7.1. For visualization purpose, skull-stripping procedure is usually performed on APTw image. In the rest of chapter, we refer $S_{\mathrm{sat}}(\pm 3.5\mathrm{ppm})$ and S_0 as APT raw images and $MTR_{\mathrm{asymm}}(3.5\mathrm{ppm})$ as an APTw image.

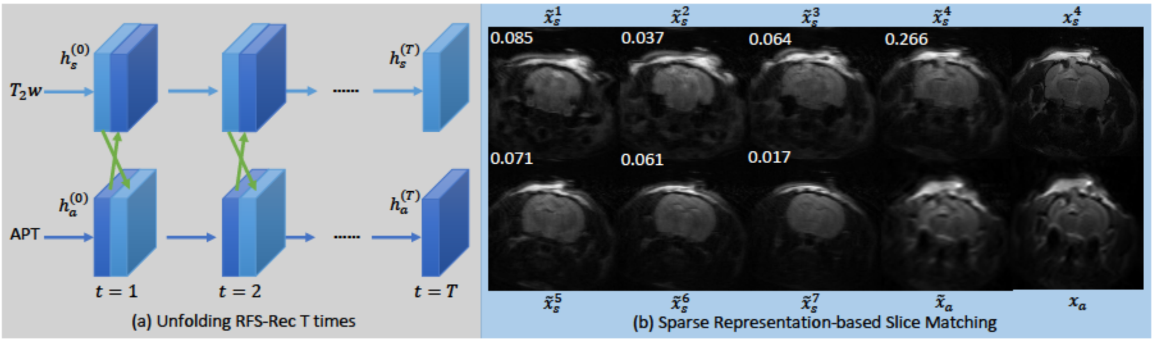


Figure 7.2: (a) The proposed recurrent neural network, RFS-Rec, can be unfolded T times. Hidden states of T_2w and APT RNN are connected by two-way feature sharing. (b) Absolute weights in sparse vector w_a are shown at the top left corner of the corresponding T_2w slices. x_a is the average of fully sampled APT raw images.

7.2.2 Recurrent Feature Sharing Reconstruction

The data acquisition process of accelerated MRI can be formulated as follows

$$y = F_D x + \epsilon \,, \tag{7.3}$$

where $x \in \mathbb{C}^M$ is the fully sampled image, $y \in \mathbb{C}^N$ is the observed k-space, and ϵ is the noise. Both x and y are image data represented in vector forms. F_D is the undersampling Fourier encoding matrix which is defined as the multiplication of the Fourier transform matrix F and the binary undersampling matrix D. We define the acceleration factor R as the ratio of the amount of k-space data required for a fully sampled image to the amount collected in an accelerated acquisition. The goal of MRI image reconstruction is to estimate image x from the observed k-space y. MRI reconstruction problem is an ill-posed problem due to the information loss in the forward process as $N \ll M$.

We solve the MRI image reconstruction problem in an iterative manner using CRNN as the base reconstruction network in RFS-Rec. A single contrast CRNN can be divided into four parts: 1) encoder f_{enc}, 2) decoder f_{dec}, 3) hidden state transition f_{res} consisting of two residual convolution blocks (ResBlock), and 4) data consistency (DC) layer. f_{enc} and f_{dec} are constructed using strided and transposed convolutions. Input images to CRNN are zero-filled undersampled complex APT raw images $x^{(0)} = F_D^H y$ with real and imaginary values as two channels. The output of the $(t+1)^{th}$ iteration of a single CRNN_{apt} model can be described as follows:

$$\begin{aligned} x^{(t+1)} &= \text{DC}(f(x^{(t)}, h^{(t)}, y, D)), \\ &= F^{-1}[Dy + (1-D)F f_{dec}(f_{res}(h^{(t)}) + f_{enc}(x^{(t)}))], \end{aligned} \tag{7.4}$$

where $h^{(t)} = f_{res}(h^{(t-1)}) + f_{enc}(x^{(t-1)})$ is the hidden state from the previous iteration and $h^{(0)} = 0$.

As discussed above, by using the information from other contrast, one can more accurately reconstruct an image of another contrast. This approach is known as multi-contrast MRI reconstruction [105]. Information or feature sharing has been shown to be the key for multi-contrast MR image reconstruction [104] [103].

Note that CRNNs have been proposed for MRI reconstruction [75] and it has been demonstrated that they can outperform cascaded models and U-net [5]. However, feature sharing in CRNN has not been studied in the literature for MRI reconstruction. We present a novel recurrent feature sharing method that exchanges intermediate hidden state features of two CRNNs (see Fig.7.2(a)). This allows us to use CRNNs for multi-contrast MR image reconstruction in a more efficient way.

The proposed RFS-Rec consists of two CRNNs, $\mathrm{CRNN}_{\mathrm{apt}}$ and $\mathrm{CRNN}_{\mathrm{t2w}}$. $\mathrm{CRNN}_{\mathrm{t2w}}$ for $\mathrm{T_2w}$ images are constructed similar to $\mathrm{CRNN}_{\mathrm{apt}}$ which is defined in Eq.7.4 but without the DC layer. $\mathrm{CRNN}_{\mathrm{t2w}}$ takes the reference slice x_s^* which is assumed to be aligned with underlying full sampled x.

To enable two-way information flow between APT features h_a and structural $\mathrm{T_2w}$ features h_s, we add bi-directional skip connection links (Fig.7.2(a)) in each iteration, which is inspired by the one-time feature concatenation in Y-net [104]. Thus, the overall dynamics of our proposed RFS-Rec is given as follows

$$\begin{aligned} h_a^{(t)} &= f_{res}(h_a^{(t-1)} \oplus h_s^{(t-1)}) + f_{enc}(x^{(t-1)}), \text{ and} \\ h_s^{(t)} &= f_{res}(h_s^{(t-1)} \oplus h_a^{(t-1)}) + f_{enc}(x_s^{(t-1)}), \end{aligned} \tag{7.5}$$

where $\oplus$ stands for channel-wise concatenation. We refer to this hidden state exchange design as recurrent feature sharing.

In terms of the loss function, we use a combination of the Normalised Mean

Square Error (NMSE) loss and the Structural Similarity Index (SSIM) loss as our training loss. The overall loss function we use to train the network is defined as follows

$$\begin{aligned} \mathcal{L}(\hat{x}, x) &= \mathcal{L}_{\text{NMSE}} + \beta \mathcal{L}_{\text{SSIM}}, \\ &= \frac{\|\hat{x} - x\|_2^2}{\|x\|_2^2} + \beta \frac{(2\mu_{\hat{x}}\mu_x + c_1)(2\sigma_{\hat{x}x} + c_2)}{(\mu_{\hat{x}}^2 + \mu_x^2 + c_1)(\sigma_{\hat{x}}^2 + \sigma_x^2 + c_2)}, \end{aligned} \tag{7.6}$$

where $\mu_{\hat{x}}$ and μ_x are the average pixel intensities in $\hat{x}$ and x, respectively, $\sigma_{\hat{x}}^2$ and σ_x^2 are their variances, $\sigma_{\hat{x}x}$ is the covariance between $\hat{x}$ and x, and $c_1 = (k_1 L)^2$, $c_2 = (k_2 L)^2$. In this chapter, we choose a window size of 7×7, and set $k_1 = 0.01$, $k_2 = 0.03$, and define L as the maximum magnitude value of the target image x, i.e. $L = \max(|x|)$. We use $\beta = 0.5$ to balance the two loss functions.

7.2.3 Sparse Representation-based Slice Matching

As mentioned earlier, a 3D T_2w scan is normally acquired prior to 2D APTw imaging. In order to fully leverage the T_2w scan, it is important to identify the matching slices between T_2w and the undersampled APT raw image. We propose a simple yet effective sparse representation-based slice matching algorithm that can find the closest slice in T_2w scan in terms of location given undersampled APT raw images.

Sparse representation-based approach, first described in [106], exploits the discriminative nature of sparsity. The average undersampled APT raw image can be represented by a set of T$_2$w images as a linear combination of all elements. This representation is naturally sparse and can be recovered efficiently via ℓ_1-minimization, seeking the sparsest representation of the APT raw image. Let $X_s = [\tilde{x}_s^1,\ \tilde{x}_s^2, \ldots,\ \tilde{x}_s^n]$ be the matrix that contains all n undersampled structural T$_2$w slices, $\tilde{x}_s = F_D^H F_D x_s$ and $\tilde{x}_a = \sum_{i=1}^{3} F_D^H y_i/3$ be the average of the undersampled APT raw images. The sparsest vector w_a that represents $\tilde{x}_a$ in X_s and gives small reconstruction error $\|\tilde{x}_a - X_s w_a\|_2$ can be found by solving the following l_1-minimization problem

$$w_a = \underset{w}{\operatorname{argmin}} \|w\|_1 \quad \text{s.t.} \quad \|\tilde{x}_a - X_s w\|_2 \leq \sigma. \tag{7.7}$$

After solving the optimization problem, the matching T$_2$w slice x_s^* is determined by the slice index $i = \operatorname{argmax} |w_a^i|$ (i.e. the slice with the largest absolute weight). From an example of SR slice matching ($\sigma = 0.1$) shown in Fig.7.2(b), x_s^4 which has the largest absolute weight $w_a^4 = 0.266$ is the one matched to the APT raw image $\tilde{x}_a$ and will be used as the reference T$_2$w image in the reconstruction phase.

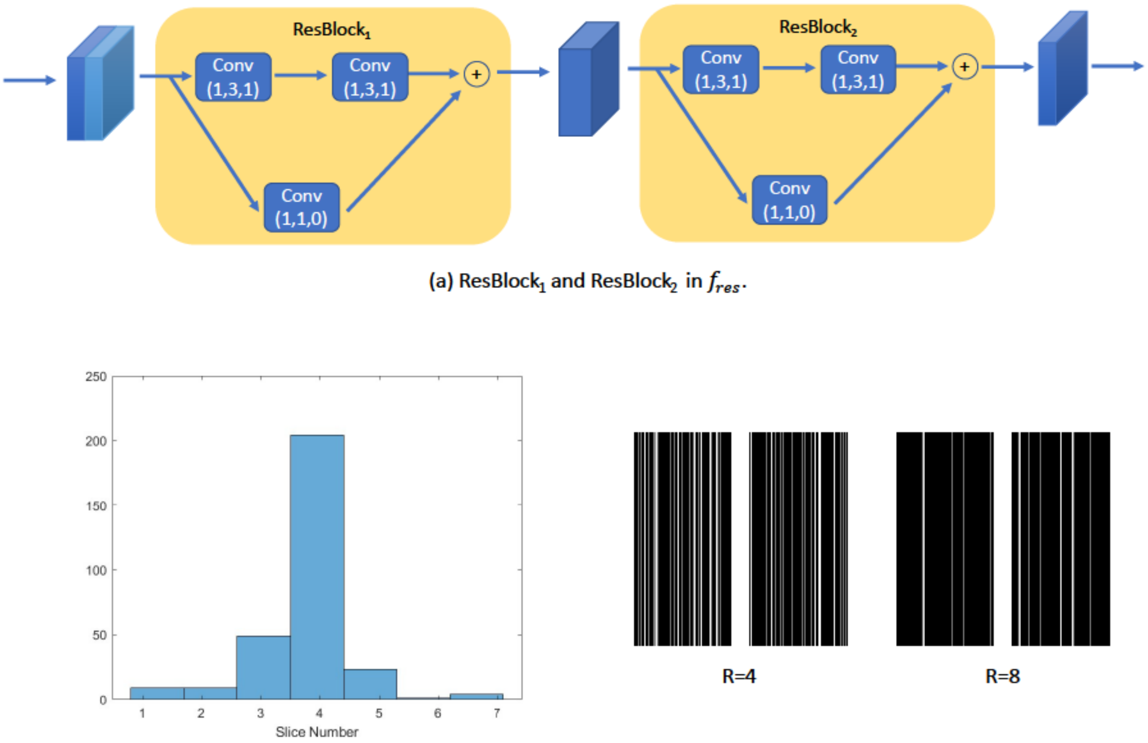
Table 7.1: Details of proposed RFS-Rec network architecture for implementation on Rat TBI dataset. $x^{(0)}$ is the concatenation of all three undersampled complex APT raw images and x_s^* is the reference T$_2$w slice. Conv(2,4,1) represents a 2D convolution layer with stride=2, kernel size=3×3 and zero-padding size=1×1. Rectified Linear units (ReLU) are omitted for simplicity.

Module	Block	Type	Output size
CRNN$_{\text{apt}}$	Input $x^{(0)}$	–	6×64×64
	Encoder f_{enc}	Conv(2,4,1) Conv(2,4,1)	128×32×32 256×16×16
	ResBlocks f_{res}	ResBlock$_1$ ResBlock$_2$	512×16×16 256×16×16
	Decoder f_{dec}	DeConv(2,4,1) DeConv(2,4,1)	128×32×32 6×64×64
	Data Consistency	–	6×64×64
CRNN$_{\text{t2w}}$	Input x_s^*	–	1×64×64
	Encoder f_{enc}	Conv(2,4,1) Conv(2,4,1)	128×32×32 256×16×16
	ResBlocks f_{res}	ResBlock$_1$ ResBlock$_2$	512×16×16 256×16×16
	Decoder f_{dec}	DeConv(2,4,1) DeConv(2,4,1)	128×32×32 1×320×320

	ResBlock$_1$	Input
		Conv(1,3,1)
		Conv(1,3,1)
		Conv(1,1,0)
			C×M×N
			C×M×N
			C×M×N
			C×M×N

	ResBlock$_2$	Input
		Conv(1,3,1)
		Conv(1,3,1)
		Conv(1,1,0)
			C×M×N
			C×M×N
			C/2×M×N
			C/2×M×N

Output $\hat{x}$	–	–	6×64×64



(a) ResBlock$_1$ and ResBlock$_2$ in f_{res}.

(b) Histogram of proposed SR-based slice matching results.

(c) Example of random sampling masks in Kspace.

Figure 7.3: (a) Details of ResBlock$_1$ and ResBlock$_2$. (b) Note that slice #4, the center slice in RAT TBI T$_2$w volume, is the most frequent output of our SR-based slice matching algorithm. This result is expected for real-world APTw MR imaging protocol.

Table 7.2: Quantitative results of APT raw image reconstruction under the acceleration factors $R = 4$ and $R = 8$. T_2w indicates whether T_2w image is used during reconstruction. SM denotes the use of the proposed SR slice matching instead of always using the center T_2w slice. Note that, for Human brain dataset, the slice matching does not apply because T_2w and APT volume are already well co-registered.

				R=4			R=8		
Dataset	Method	T_2w	SM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM
Rat	U-net [5]			0.144	34.29	0.920	0.242	31.96	0.878
	Y-net [104]	✓		0.111	35.35	0.932	0.218	32.29	0.889
	CRNN_{apt}			0.087	36.41	0.939	0.217	32.31	0.889
	CRNN	✓		0.085	36.43	0.940	0.219	32.28	0.893
	CRNN	✓	✓	0.084	36.56	0.941	0.212	32.37	0.893
	RFS-Rec	✓	✓	**0.076**	**36.94**	**0.950**	**0.187**	**33.11**	**0.906**
Human	U-net [5]		N/A	0.022	37.19	0.910	0.045	33.76	0.872
	Y-net [104]	✓	N/A	0.014	39.27	0.938	0.037	34.65	0.889
	CRNN_{apt}		N/A	0.014	39.64	0.943	0.041	34.30	0.887
	CRNN	✓	N/A	0.012	40.35	0.950	0.038	34.84	0.896
	RFS-Rec	✓	N/A	**0.010**	**40.99**	**0.956**	**0.034**	**35.27**	**0.903**

7.3 Experiments

7.3.0.1 Datasets

We evaluate the proposed image reconstruction framework on two datasets. **Rat TBI Data**: 300 MRI scans are performed on 65 open-skull rats with controlled cortical impact model of TBI at different time point after TBI. Each MRI scan includes high-resolution T_2w imaging with a fast spin echo sequence in coronal plane (number of slices= 7; matrix= 256×256; field of view (FOV) = 32×32 mm^2; slice thickness = 1.5 mm) and 2D APT (frequency labeling offsets of ±3.5 ppm; matrix= 64×64; FOV = 32×32 mm^2; single slice; slice thickness = 1.5 mm). An unsaturated image S_0 in the absence of radio-frequency saturation was also acquired for APT imaging signal intensity normalization. All animal experiments were conducted according to the National Institute of Health (NIH) guidelines for the care and use of laboratory animals, and approved by the Johns Hopkins University Animal Care and Use Committee (ACUC).

Human Brain Tumor Data: 144 3D T_2w and APTw MRI volumes were collected from 90 patients with pathologically proven primary malignant glioma. Imaging parameters for APTw can be summarized as follows: FOV = 212×212×66 mm3; resolution = 0.82×0.82×4.4 mm^3; size in voxel = 256×256×15. T_2w sequences were acquired with the following imaging parameters: FOV = 212×212× 165 mm3; resolution, 0.41×0.41×1.1 mm3 ; size in voxel, 512×512×150. Coregistration between APTw and T_2w sequences [107], and MRI standardization [108] were performed. After preprocessing, the final volume size of each sequence is 256×256×15. Data collection and processing are approved by the Institutional Review Board.

Training Details: We simulated undersampled k-space measurements of APT raw images using the Cartesian sampling method with a fixed 0.08% center frequency sampled and random sampling in other frequencies uniformly. Training and testing subsets are randomly selected with 80/20% split. We conducted model training under the acceleration factors R=4 and 8. All models are implemented in Pytorch and trained on NVIDIA GPUs. Hyperparameters are set as follows: learning rate of 10^{-3} with decreasing rate of 0.9 for every 5 epochs, 60 maximum epochs, batch size of 6. Adam optimizer is used in training all the networks. For CRNN and RFS-Rec, the number of iterations T is set equal to 7.

We compare our proposed RFS-Rec against U-net [5], Y-net [104], single contrast CRNN $_{\text{apt}}$, CRNN with concatenation of center T_2w slice and undersampled APT raw images as input and CRNN using the proposed SR slice matching to select the reference slice. Regarding the U-net implementation, a DC layer was added at the end of the network. The quantitative metrics, including NMSE, PSNR and SSIM, are computed between fully sampled APT raw images ($S_{\text{sat}}(\pm 3.5\text{ppm})$ and S_0) and their reconstructions. Detailed quanti-

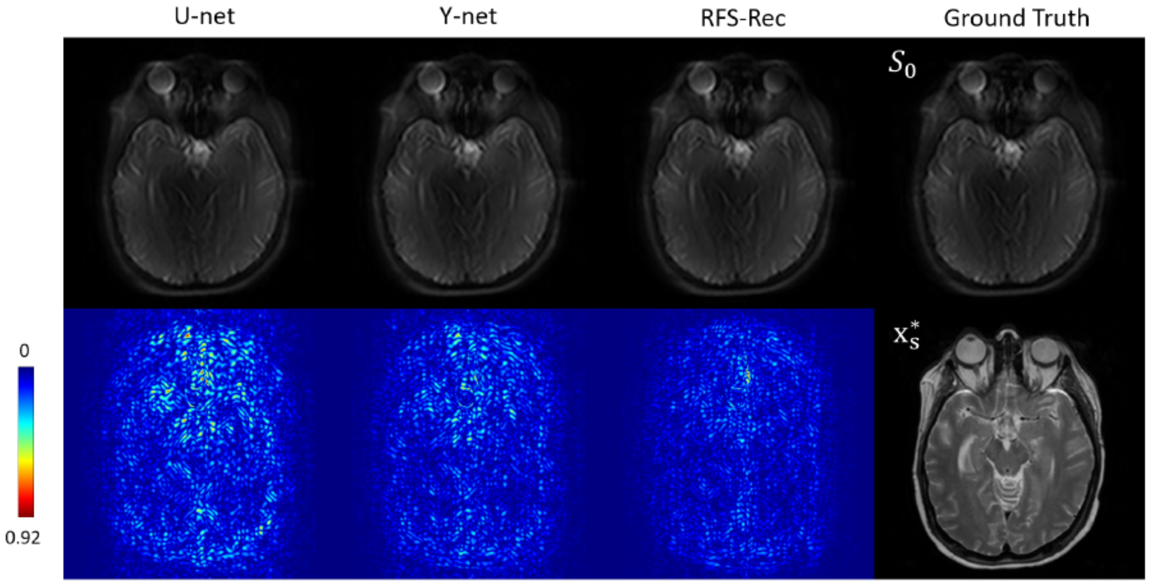
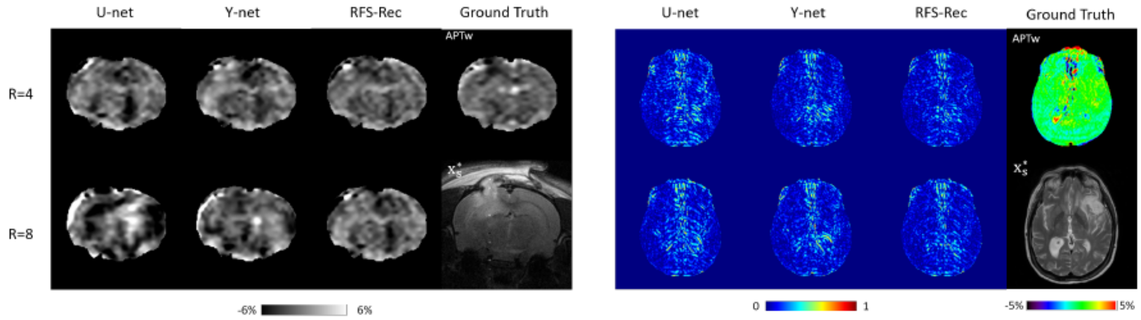


Figure 7.4: S_0 reconstructions at $R = 4$ and the corresponding error maps.

tative experimental results are shown in Tab.7.2. It can be seen from the table that the proposed RFS-Rec approach outperforms all the other compared methods on both datasets. Furthermore, the individual contribution of the modules in the proposed method (SR-based slice matching and RFS), are demonstrated by an ablation study (i.e. CRNN with/without SM and RFS-Rec). One interesting observation from Tab. 7.2 is that the difference between CRNN_{apt} and Y-net, when $R = 8$, on the human dataset is inverse of what we observe on the rat dataset. This may be caused by the good registration of T_2w and APT in the human dataset. The issue of shape inconsistency of the APT raw image and T_2w image in the rat dataset can also be observed by comparing CRNN_{apt} and CRNN with T_2w.

Results of reconstructed S_0 and APTw images compared to the ground truth



(a) Reconstructed APTw image

(b) Error map of reconstructed APTw image

Figure 7.5: Results of APTw images derived from the reconstructed APT raw images using Eq.7.2. Skull-stripping is performed for better visualization. Reference T_2w slice x_s^* used for reconstruction are also shown.

in Fig.7.4 and Fig.7.5 show consistent findings as quantitative results suggest. Our method yields not only better $S_{\text{sat}}(\pm 3.5\text{ppm})$ and S_0 reconstruction but also more accurate APTw images.

7.4 Summary

We proposed an APTw image reconstruction network RFS-Rec for accelerating APTw imaging, which can more accurately reconstruct APT raw images by using the information of fully sampled T_2w images. We achieved this goal by incorporating a novel recurrent feature sharing mechanism into two CRNNs which enable two-way information flow between APT and T_2w features. In addition, to maximize the effectiveness of RFS-Rec, we use a sparse representation-based slice matching algorithm to locate reference T_2w slice. Extensive experiments on two real datasets consisting of brain data from rats and humans showed the significance of the proposed work.

Chapter 8

Summary and Future Work

In this thesis, we addressed two problems in medical image processing and analysis (i.e. segmentation and reconstruction). For each problem, deep learning-based solutions were presented and the effectiveness of the methods was assessed with extensive experiments and analysis.

In order to make US an essential imaging modality in orthopedics clinically acceptable accuracy and robustness of guidance system needs to be ensured. Therefore, complete, accurate and robust bone segmentation is of paramount importance for US-based orthopedic surgical and non-surgical procedures where automatically extracted bone surfaces are used for continuous real-time guidance.

We first proposed a unified multi-feature guided CNN framework that enables robust and accurate bone surfaces classification and segmentation for

US-based computer assisted orthopedic procedures. Although experiments demonstrated the effectiveness of using filtered image features including local phase-based enhanced bone images, and signal transmission-based bone shadow enhanced image. It lacks the integration of computing filtered images and deep learning framework, thus prohibit the real-time application.

In Chapter 4, we proposed a novel local phase tensor-guided CNN architecture for bone surface segmentation from US data of various quality. In order to improve the computation time of previous-proposed multi-feature guided CNN [18], our proposed framework accommodates a Local Phase Tensor (LPT) network that is trained to capture contrast and noise invariant local phase information. To further improve the robustness and suppress non-bone responses of LPT network, a Global Context Tensor (GCT) network that focuses on learning global context is proposed. The main novelty of this work lies in (1) the integration of learning Local Phase Tensor (LPT) and Global Context Tensor (GCT) into a single network, (2) the design of fusion method of LPT and GCT to improve cross-machine segmentation performance of various bone imaging quality, and (3) the first systematic design of a fully automatic real-time framework for robust multi-machine LPT-guided bone surface segmentation from US images. It is critical for an automatic US segmentation algorithm to maintain robust performance on various US machines without any modification. As a segmentation task, the base segmentation network plays a crucial part in performance. Therefore, we tried to improve bone segmentation by proposing a novel multi-scale pyramid pooling decoder-based CNN. Furthermore, we applied the idea of using filtered image as pseudo ground truth to bone shadow segmentation by replacing LPT image with Bone Shadow Enhanced (BSE) image and introducing an novel subtask of estimating horizontal bone interval mask.

For the second problem, we focus on applying deep learning techniques to Magnetic Resonance Imaging (NRI) reconstruction. We proposed a multi-scale pyramid Convolutional RNN model for MRI image reconstruction. The model reconstructs MRI image in various scales and then combines the coarse-to-fine images as the final reconstructed image. We evaluated our model on the fastMRI knee and brain datasets. The results show that our model can reconstruct more fine details and outperforms other methods in fastMRI multi-coil 4X tasks. Finally, we investigated the Amide Proton Transfer-weighted (APTw) imaging, an emerging molecular MRI, reconstruction problem by using two convolutional RNNs as base network. These two CRNNs are connected by the proposed recurrent feature sharing approach to encourage bi-directional flow of information. In addition, we proposed a sparse representation-based slice matching algorithm to find the corresponding slice in T_2w volume given the undersampled APT k-space data. As a result, input T_2w and APT raw images are aligned and mutual information can be maximized. There it can more effectively leverage the structural information embedding in the T_2w slice to better reconstruct the APTw slice. However, there are many problems still remain to be addressed. In what follows, we outline a few open problems.

8.1 US Image Segmentation Future Work

- One limitation of our study is that the uncertainty present in the gold standard labels (manual expert segmentation), resulting from intra- and inter-user variability, was not investigated. These errors can have direct impact on the accuracy of the developed deep neural networks. Computing and analyzing uncertainty map of segmentation result is one of the future work for US image segmentation.

- Another limitation is that only two US machines were used to collect the data. In order to fully investigate the true generalization of our method, more US machines should be used for data collection. We also would like to mention that one of the ongoing limitation in US bone segmentation research is that there are still no publicly available large data sets on which different algorithms could be trained and evaluated on. One solution to this could be joint effort in order to construct a publicly available database. Such an effort was most recently proposed in [109]. Our future work will include the extensive evaluation of our proposed method on this publicly available dataset.

- We will also investigate the incorporation of the extracted bone surfaces into a registration method and extension to 3D data for processing volumetric US scans in our future work.

8.2 MRI Reconstruction Future Work

- In the future, we hope to address challenges that we face in practical implementation of MR image reconstruction with more sampling patterns and real undersampled k-space data.
- Although we demonstrated our proposed method on MR image reconstruction, it can be extended to other imaging modalities such as CT image reconstruction.
- The problem of how deep learning-based reconstructed MR images affect the downstream tasks such as classification and segmentation still needs to be investigated.
- Further, deployment of deep learning-based MRI reconstruction networks across different sites also worth investigating as well. Current deep learning-based methods require large amounts of data which is difficult to collect and share due to the high cost of acquisition and medical data privacy reg-

ulations. In order to overcome this challenge, a federated learning (FL) based solution in which it can take advantage of the MR data available at different institutions while preserving patients privacy can be studied.

Bibliography

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2015, pp. 234–241.

[3] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1125–1134.

[4] I. Hacihaliloglu, "Enhancement of bone shadow region using local phase-based ultrasound transmission maps," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 6, pp. 951–960, 2017.

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[6] N. Baka, S. Leenstra, and T. van Walsum, “Ultrasound aided vertebral level localization for lumbar surgery,” *IEEE transactions on medical imaging*, vol. 36, no. 10, pp. 2138–2147, 2017.

[7] S. Robinson, “Neonatal posthemorrhagic hydrocephalus from prematurity: pathophysiology and current treatment concepts: a review,” *Journal of Neurosurgery: Pediatrics*, vol. 9, no. 3, pp. 242–258, 2012.

[8] W. Qiu, J. Yuan, J. Kishimoto, J. McLeod, Y. Chen, S. de Ribaupierre, and A. Fenster, “User-guided segmentation of preterm neonate ventricular system from 3-d ultrasound images using convex optimization,” *Ultrasound in medicine & biology*, vol. 41, no. 2, pp. 542–556, 2015.

[9] W. Qiu, Y. Chen, J. Kishimoto, S. de Ribaupierre, B. Chiu, A. Fenster, and J. Yuan, “Automatic segmentation approach to extracting neonatal cerebral ventricles from 3d ultrasound images,” *Medical image analysis*, vol. 35, pp. 181–191, 2017.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural*

Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[12] H. Greenspan, B. van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[13] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[14] I. Hacihaliloglu, "Ultrasound imaging and segmentation of bone surfaces: A review," *Technology*, pp. 1–7, 2017.

[15] M. Yamauchi, R. Kawaguchi, S. Sugino, M. Yamakage, E. Honma, and A. Namiki, "Ultrasound-aided unilateral epidural block for single lower-extremity pain," *Journal of anesthesia*, vol. 23, no. 4, pp. 605–608, 2009.

[16] A. Seitel, S. Sojoudi, J. Osborn, A. Rasoulian, S. Nouranian, V. A. Lessoway, R. N. Rohling, and P. Abolmaesumi, "Ultrasound-guided spine anesthesia: feasibility study of a guidance system," *Ultrasound in medicine & biology*, vol. 42, no. 12, pp. 3043–3049, 2016.

[17] E. M. A. Anas, A. Seitel, A. Rasoulian, P. S. John, D. Pichora, K. Darras, D. Wilson, V. A. Lessoway, I. Hacihaliloglu, P. Mousavi, R. ROhling, and P. Abolmaesumi, "Bone enhancement in ultrasound using local spectrum variations for guiding percutaneous scaphoid fracture fixation procedures," *International journal of computer assisted radiology and surgery*, vol. 10, no. 6, pp. 959–969, 2015.

[18] P. Wang, V. M. Patel, and I. Hacihaliloglu, "Simultaneous segmentation and classification of bone surfaces from ultrasound using a multi-feature guided cnn," in *International Conference on Medical Image Computing and Computer-Assisted Intervention.* Springer, 2018, pp. 134–142.

[19] A. Z. Alsinan, V. M. Patel, and I. Hacihaliloglu, "Automatic segmentation of bone surfaces from ultrasound using a filter-layer-guided cnn," *International journal of computer assisted radiology and surgery*, pp. 1–9, 2019.

[20] M. Villa, G. Dardenne, M. Nasan, H. Letissier, C. Hamitouche, and E. Stindel, "Fcn-based approach for the automatic segmentation of bone

surfaces in ultrasound images," *International journal of computer assisted radiology and surgery*, vol. 13, no. 11, pp. 1707–1716, 2018.

[21] S. Schumann, "State of the art of ultrasound-based registration in computer assisted orthopedic interventions," in *Computational Radiology for Orthopaedic Interventions*. Springer, 2016, pp. 271–297.

[22] I. Hacihaliloglu, P. Guy, A. J. Hodgson, and R. Abugharbieh, "Automatic extraction of bone surfaces from 3d ultrasound images in orthopaedic trauma cases," *International journal of computer assisted radiology and surgery*, vol. 10, no. 8, pp. 1279–1287, 2015.

[23] J. M. Edmund and T. Nyholm, "A review of substitute ct generation for mri-only radiation therapy," *Radiation Oncology*, vol. 12, no. 1, p. 28, 2017.

[24] M. Lecchi, P. Fossati, F. Elisei, R. Orecchia, and G. Lucignani, "Current concepts on imaging in radiotherapy," *European journal of nuclear medicine and molecular imaging*, vol. 35, no. 4, pp. 821–837, 2008.

[25] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[26] V. M. Patel, G. R. Easley, D. M. Healy Jr, and R. Chellappa, "Compressed synthetic aperture radar," *IEEE Journal of selected topics in signal processing*, vol. 4, no. 2, pp. 244–254, 2010.

[27] S. Ma, W. Yin, Y. Zhang, and A. Chakraborty, "An efficient algorithm for compressed mr imaging using total variation and wavelets," in *CVPR*. IEEE, 2008, pp. 1–8.

[28] J. C. Ye, "Compressed sensing mri: a review from signal processing perspective," *BMC Biomedical Engineering*, vol. 1, no. 1, p. 8, 2019.

[29] S. Ravishankar and Y. Bresler, "Mr image reconstruction from highly undersampled k-space data by dictionary learning," *IEEE TMI*, vol. 30, no. 5, pp. 1028–1041, 2010.

[30] Y. Huang, J. Paisley, Q. Lin, X. Ding, X. Fu, and X.-P. Zhang, "Bayesian nonparametric dictionary learning for compressed sensing mri," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5007–5019, 2014.

[31] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse mri: The application of compressed sensing for rapid mr imaging," *MRM*, vol. 58, no. 6, pp. 1182–1195, 2007.

[32] J. A. Fessler, "Optimization methods for mr image reconstruction," *arXiv:1903.03510*, 2019.

[33] K. P. Pruessmann, M. Weiger, M. B. Scheidegger, and P. Boesiger, “Sense: sensitivity encoding for fast mri,” *MRM*, vol. 42, no. 5, pp. 952–962, 1999.

[34] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase, “Generalized autocalibrating partially parallel acquisitions (grappa),” *MRM*, vol. 47, no. 6, pp. 1202–1210, 2002.

[35] A. Deshmane, V. Gulani, M. A. Griswold, and N. Seiberlich, “Parallel mr imaging,” *Journal of Magnetic Resonance Imaging*, vol. 36, no. 1, pp. 55–72, 2012.

[36] A. Selvikvg Lundervold and A. Lundervold, “An overview of deep learning in medical imaging focusing on mri,” *arXiv:1811.10052*, 2018.

[37] Z. He, J. Zhou, D. Liang, Y. Wang, and Q. Liu, “Learning priors in high-frequency domain for inverse imaging reconstruction,” *arXiv:1910.11148*, 2019.

[38] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[39] R. K. Srivastava, K. Greff, and J. Schmidhuber, “Highway networks,” *arXiv preprint arXiv:1505.00387*, 2015.

[40] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," *arXiv preprint arXiv:1608.06993*, 2016.

[41] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[42] M. Salehi, R. Prevost, J.-L. Moctezuma, N. Navab, and W. Wein, "Precise ultrasound bone registration with learning-based segmentation and speed of sound calibration," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 682–690.

[43] I. Hacihaliloglu, A. Rasoulian, R. N. Rohling, and P. Abolmaesumi, "Local phase tensor features for 3-d ultrasound to statistical shape+ pose spine model registration," *IEEE transactions on medical imaging*, vol. 33, no. 11, pp. 2167–2179, 2014.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 448–456.

[45] T. Kurmann, P. M. Neila, X. Du, P. Fua, D. Stoyanov, S. Wolf, and R. Sznitman, "Simultaneous recognition and pose estimation of in-

struments in minimally invasive surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 505–513.

[46] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[47] I. Hacihaliloglu, "Localization of bone surfaces from ultrasound data using local phase information and signal transmission maps," in *International Workshop and Challenge on Computational Methods and Clinical Applications in Musculoskeletal Imaging*. Springer, 2017, pp. 1–11.

[48] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[49] P.-A. Ganaye, M. Sdika, and H. Benoit-Cattin, "Semi-supervised learning for segmentation under semantic constraint," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 595–602.

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[51] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2881–2890.

[52] A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” *arXiv preprint arXiv:1511.06434*, 2015.

[53] A. Sriram, J. Zbontar, T. Murrell, A. Defazio, C. L. Zitnick, N. Yakubova, F. Knoll, and P. Johnson, “End-to-end variational networks for accelerated mri reconstruction,” *arXiv:2004.06688*, 2020.

[54] N. Pezzotti, E. de Weerdt, S. Yousefi, M. S. Elmahdy, J. van Gemert, C. Schülke, M. Doneva, T. Nielsen, S. Kastryulin, B. P. Lelieveldt *et al.*, “Adaptive-cs-net: Fastmri with adaptive intelligence,” *arXiv:1912.12259*, 2019.

[55] H. C. M. Holme, S. Rosenzweig, F. Ong, R. N. Wilke, M. Lustig, and M. Uecker, “Enlive: an efficient nonlinear method for calibrationless and robust parallel imaging,” *Scientific reports*, vol. 9, no. 1, pp. 1–13, 2019.

[56] T. Zhang, J. M. Pauly, S. S. Vasanawala, and M. Lustig, “Coil compression for accelerated imaging with cartesian sampling,” *MRM*, vol. 69, no. 2, pp. 571–582, 2013.

[57] K. C. Tezcan, C. F. Baumgartner, R. Luechinger, K. P. Pruessmann, and E. Konukoglu, “Mr image reconstruction using deep density priors,” *IEEE TMI*, vol. 38, no. 7, pp. 1633–1642, 2018.

[58] H. K. Aggarwal, M. P. Mani, and M. Jacob, “Modl: Model-based deep learning architecture for inverse problems,” *IEEE TMI*, vol. 38, no. 2, pp. 394–405, 2018.

[59] M. Mardani, E. Gong, J. Y. Cheng, S. Vasanawala, G. Zaharchuk, M. Alley, N. Thakur, S. Han, W. Dally, J. M. Pauly *et al.*, “Deep generative adversarial networks for compressed sensing automates mri,” *arXiv:1706.00051*, 2017.

[60] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, F. Liu, S. Arridge, J. Keegan, Y. Guo *et al.*, “Dagan: deep de-aliasing generative adversarial networks for fast compressed sensing mri reconstruction,” *IEEE TMI*, vol. 37, no. 6, pp. 1310–1321, 2017.

[61] J. Sun, H. Li, Z. Xu *et al.*, “Deep admm-net for compressive sensing mri,” in *NIPS*, 2016, pp. 10–18.

[62] J. Zhang and B. Ghanem, “Ista-net: Interpretable optimization-inspired deep network for image compressive sensing,” in *CVPR*, 2018, pp. 1828–1837.

[63] J. Duan, J. Schlemper, C. Qin, C. Ouyang, W. Bai, C. Biffi, G. Bello, B. Statton, D. P. ORegan, and D. Rueckert, "Vs-net: Variable splitting network for accelerated parallel mri reconstruction," in *MICCAI*. Springer, 2019, pp. 713–722.

[64] Y. Chen, T. Xiao, C. Li, Q. Liu, and S. Wang, "Model-based convolutional de-aliasing network learning for parallel mr imaging," in *MICCAI*. Springer, 2019, pp. 30–38.

[65] K. Hammernik, T. Klatzer, E. Kobler, M. P. Recht, D. K. Sodickson, T. Pock, and F. Knoll, "Learning a variational network for reconstruction of accelerated mri data," *MRM*, vol. 79, no. 6, pp. 3055–3071, 2018.

[66] J. Schlemper, J. Caballero, J. V. Hajnal, A. N. Price, and D. Rueckert, "A deep cascade of convolutional neural networks for dynamic mr image reconstruction," *IEEE TMI*, vol. 37, no. 2, pp. 491–503, 2017.

[67] Q. Huang, D. Yang, P. Wu, H. Qu, J. Yi, and D. Metaxas, "Mri reconstruction via cascaded channel-wise attention network," in *ISBI*. IEEE, 2019, pp. 1622–1626.

[68] R. Souza and R. Frayne, "A hybrid frequency-domain/image-domain deep network for magnetic resonance image reconstruction," in *SIBGRAPI*. IEEE, 2019, pp. 257–264.

[69] R. Souza, M. Bento, N. Nogovitsyn, K. J. Chung, R. M. Lebel, and R. Frayne, "Dual-domain cascade of u-nets for multi-channel magnetic resonance image reconstruction," *arXiv:1911.01458*, 2019.

[70] T. Eo, Y. Jun, T. Kim, J. Jang, H.-J. Lee, and D. Hwang, "Kiki-net: cross-domain convolutional neural networks for reconstructing undersampled magnetic resonance images," *MRM*, vol. 80, no. 5, pp. 2188–2201, 2018.

[71] S. Wang, Z. Ke, H. Cheng, S. Jia, Y. Leslie, H. Zheng, and D. Liang, "Dimension: Dynamic mr imaging with both k-space and spatial prior knowledge obtained via multi-supervised network training," *arXiv:1810.00302*, 2018.

[72] L. Bao, F. Ye, C. Cai, J. Wu, K. Zeng, P. C. van Zijl, and Z. Chen, "Undersampled mr image reconstruction using an enhanced recursive residual network," *Journal of Magnetic Resonance*, vol. 305, pp. 232–246, 2019.

[73] D. Gilton, G. Ongie, and R. Willett, "Neumann networks for inverse problems in imaging," *arXiv:1901.03707*, 2019.

[74] B. Zhou and S. K. Zhou, "Dudornet: Learning a dual-domain recurrent network for fast mri reconstruction with deep t1 prior," in *CVPR*, 2020, pp. 4273–4282.

[75] C. Qin, J. Schlemper, J. Caballero, A. N. Price, J. V. Hajnal, and D. Rueck-

ert, “Convolutional recurrent neural networks for dynamic mr image reconstruction,” *IEEE TMI*, vol. 38, no. 1, pp. 280–290, 2018.

[76] P. Putzky and M. Welling, “Invert to learn to invert,” in *NIPS*, 2019, pp. 444–454.

[77] E. Z. Chen, T. Chen, and S. Sun, “Mri image reconstruction via learning optimization using neural odes,” *arXiv:2006.13825*, 2020.

[78] D. Lee, J. Yoo, S. Tak, and J. C. Ye, “Deep residual learning for accelerated mri using magnitude and phase networks,” *IEEE T BIO-MED ENG*, vol. 65, no. 9, pp. 1985–1995, 2018.

[79] Y. Han, L. Sunwoo, and J. C. Ye, “k-space deep learning for accelerated mri,” *IEEE TMI*, 2019.

[80] M. Akçakaya, S. Moeller, S. Weingärtner, and K. Uğurbil, “Scan-specific robust artificial-neural-networks for k-space interpolation (raki) reconstruction: Database-free deep learning for fast imaging,” *MRM*, vol. 81, no. 1, pp. 439–453, 2019.

[81] B. Zhu, J. Z. Liu, S. F. Cauley, B. R. Rosen, and M. S. Rosen, “Image reconstruction by domain-transform manifold learning,” *Nature*, vol. 555, no. 7697, p. 487, 2018.

[82] J. Schlemper, I. Oksuz, J. R. Clough, J. Duan, A. P. King, J. A. Schnabel, J. V. Hajnal, and D. Rueckert, “dautomap: decomposing automap to achieve scalability and enhance performance,” *arXiv:1909.10995*, 2019.

[83] J. Zbontar, F. Knoll, A. Sriram, M. J. Muckley, M. Bruno, A. Defazio, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana *et al.*, “fastmri: An open dataset and benchmarks for accelerated mri,” *arXiv:1811.08839*, 2018.

[84] N. Meng, Y. Yang, Z. Xu, and J. Sun, “A prior learning network for joint image and sensitivity estimation in parallel mr imaging,” in *MICCAI*. Springer, 2019, pp. 732–740.

[85] K. Hammernik, J. Schlemper, C. Qin, J. Duan, R. M. Summers, and D. Rueckert, “Sigma-net: Systematic evaluation of iterative deep neural networks for fast parallel mr image reconstruction,” *arXiv:1912.09278*, 2019.

[86] M. Andrychowicz, M. Denil, S. Gomez, M. W. Hoffman, D. Pfau, T. Schaul, B. Shillingford, and N. De Freitas, “Learning to learn by gradient descent by gradient descent,” in *NIPS*, 2016, pp. 3981–3989.

[87] K. Li and J. Malik, “Learning to optimize,” *arXiv:1606.01885*, 2016.

[88] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *CVPR*, 2018, pp. 9446–9454.

[89] D. Van Veen, A. Jalal, M. Soltanolkotabi, E. Price, S. Vishwanath, and A. G. Dimakis, “Compressed sensing with deep image prior and learned regularization,” *arXiv preprint arXiv:1806.06438*, 2018.

[90] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, “Loss functions for neural networks for image processing,” *arXiv*, 2015.

[91] M. R. Zhang, J. Lucas, G. Hinton, and J. Ba, “Lookahead optimizer: k steps forward, 1 step back,” *arXiv:1907.08610*, 2019.

[92] F. Knoll, T. Murrell, A. Sriram, N. Yakubova, J. Zbontar, M. Rabbat, A. Defazio, M. Muckley, D. Sodickson, C. Zitnick *et al.*, “Advancing machine learning for mr image reconstruction with an open competition: Overview of the 2019 fastmri challenge.” *MRM*, 2020.

[93] J. Zhou, E. Tryggestad, Z. Wen, B. Lal, T. Zhou, R. Grossman, S. Wang, K. Yan, D.-X. Fu, E. Ford *et al.*, “Differentiation between glioma and radiation necrosis using molecular magnetic resonance imaging of endogenous proteins and peptides,” *Nature medicine*, vol. 17, no. 1, pp. 130–134, 2011.

[94] P. Z. Sun, J. S. Cheung, E. Wang, and E. H. Lo, “Association between ph-

weighted endogenous amide proton chemical exchange saturation transfer mri and tissue lactic acidosis during acute ischemic stroke," *Journal of Cerebral Blood Flow & Metabolism*, vol. 31, no. 8, pp. 1743–1750, 2011.

[95] M. Wang, X. Hong, C.-F. Chang, Q. Li, B. Ma, H. Zhang, S. Xiang, H.-Y. Heo, Y. Zhang, D.-H. Lee *et al.*, "Simultaneous detection and separation of hyperacute intracerebral hemorrhage and cerebral ischemia using amide proton transfer mri," *Magnetic resonance in medicine*, vol. 74, no. 1, pp. 42–50, 2015.

[96] H. Zhang, W. Wang, S. Jiang, Y. Zhang, H.-Y. Heo, X. Wang, Y. Peng, J. Wang, and J. Zhou, "Amide proton transfer-weighted mri detection of traumatic brain injury in rats," *Journal of Cerebral Blood Flow & Metabolism*, vol. 37, no. 10, pp. 3422–3432, 2017.

[97] X. Zhao, Z. Wen, G. Zhang, F. Huang, S. Lu, X. Wang, S. Hu, M. Chen, and J. Zhou, "Three-dimensional turbo-spin-echo amide proton transfer mr imaging at 3-tesla and its application to high-grade human brain tumors," *Molecular imaging and biology*, vol. 15, no. 1, pp. 114–122, 2013.

[98] Y. Zhang, H.-Y. Heo, S. Jiang, D.-H. Lee, P. A. Bottomley, and J. Zhou, "Highly accelerated chemical exchange saturation transfer (cest) measurements with linear algebraic modeling," *Magnetic resonance in medicine*, vol. 76, no. 1, pp. 136–144, 2016.

[99] H.-Y. Heo, Y. Zhang, D.-H. Lee, S. Jiang, X. Zhao, and J. Zhou, "Accelerating chemical exchange saturation transfer (cest) mri by combining compressed sensing and sensitivity encoding techniques," *Magnetic resonance in medicine*, vol. 77, no. 2, pp. 779–786, 2017.

[100] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, "Compressed sensing mri reconstruction using a generative adversarial network with a cyclic loss," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1488–1497, 2018.

[101] P. Wang, E. Z. Chen, T. Chen, V. M. Patel, and S. Sun, "Pyramid convolutional rnn for mri reconstruction," *arXiv:1912.00543*, 2019.

[102] J. Huang, C. Chen, and L. Axel, "Fast multi-contrast mri reconstruction," *Magnetic resonance imaging*, vol. 32, no. 10, pp. 1344–1352, 2014.

[103] L. Sun, Z. Fan, X. Fu, Y. Huang, X. Ding, and J. Paisley, "A deep information sharing network for multi-contrast compressed sensing mri reconstruction," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6141–6153, 2019.

[104] W.-J. Do, S. Seo, Y. Han, J. C. Ye, S. Hong Choi, and S.-H. Park, "Reconstruction of multi-contrast mr images through deep learning," *Medical Physics*, 2019.

[105] B. Bilgic, V. K. Goyal, and E. Adalsteinsson, “Multi-contrast reconstruction with bayesian compressed sensing,” *Magnetic resonance in medicine*, vol. 66, no. 6, pp. 1601–1615, 2011.

[106] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, “Robust face recognition via sparse representation,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–227, 2008.

[107] Y. Zhang, H.-Y. Heo, D.-H. Lee, X. Zhao, S. Jiang, K. Zhang, H. Li, and J. Zhou, “Selecting the reference image for registration of cest series,” *Journal of Magnetic Resonance Imaging*, vol. 43, no. 3, pp. 756–761, 2016.

[108] L. G. Nyúl, J. K. Udupa, and X. Zhang, “New variants of a method of mri scale standardization,” *IEEE transactions on medical imaging*, vol. 19, no. 2, pp. 143–150, 2000.

[109] P. Pandey, H. Patel, P. Guy, I. Hacihaliloglu, and A. J. Hodgson, “Preliminary planning for a multi-institutional database for ultrasound bone segmentation,” *EPiC Series in Health Sciences*, vol. 3, pp. 297–300, 2019.

Vita

Puyang Wang received the B.S. degree in electrical engineering from University of Electronic Science and Technology of China in 2016. Currently, he is pursuing the Ph.D. degree from The Johns Hopkins University in the Department of Electrical and Computer Engineering. His research focuses include but are not limited to Deep Learning (DL) and its application to the medical image analysis and processing, novel DNN/CNN network architecture and and self-supervised learning and their applications to image segmentation and classification.