Efficient Folded Attention for 3D Medical Image Reconstruction and Segmentation

Hang Zhang,[1,2] **Jinwei Zhang,** [1,2] **Rongguang Wang,** [3] **Qihao Zhang,** [1,2]
Pascal Spincemaille, [2] **Thanh D. Nguyen,** [2] **Yi Wang,** [1,2]
[1] Cornell University, Ithaca NY, USA
[2] Weill Cornell Medical College, New York NY, USA
[3] University of Pennsylvania, Philadelphia PA, USA
hz459@cornell.edu

Abstract

Recently, 3D medical image reconstruction (MIR) and segmentation (MIS) based on deep neural networks have been developed with promising results, and attention mechanism has been further designed for performance enhancement. However, the large size of 3D volume images poses a great computational challenge to traditional attention methods. In this paper, we propose a folded attention (FA) approach to improve the computational efficiency of traditional attention methods on 3D medical images. The main idea is that we apply tensor folding and unfolding operations to construct four small sub-affinity matrices to approximate the original affinity matrix. Through four consecutive sub-attention modules of FA, each element in the feature tensor can aggregate spatial-channel information from all other elements. Compared to traditional attention methods, with the moderate improvement of accuracy, FA can substantially reduce the computational complexity and GPU memory consumption. We demonstrate the superiority of our method on two challenging tasks for 3D MIR and MIS, which are quantitative susceptibility mapping and multiple sclerosis lesion segmentation.

Introduction

Recent deep convolutional neural networks (CNNs) are driving advances in various computer vision tasks. These tasks include high-level image recognition (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015), object detection (Ren et al. 2015; Law and Deng 2018), and semantic segmentation (Fu et al. 2019). CNN also significantly improves the performance of several low-level tasks such as super resolution (Dong et al. 2014) and image denoising (Yang et al. 2017), where full functional mapping between source and target images is required. Besides the breakthrough of natural image processing, medical image processing also benefits from CNN in various aspects. CNN based methods surpass traditional methods and achieve the near-radiologist-level performance on MRI brain tumor segmentation (Myronenko 2018), MRI multiple sclerosis segmentation (Zhang et al. 2019), and left atrial segmentation (Zhang et al. 2021a), etc. For full functional mapping task, CNNs (Yoon et al. 2018; Zhang et al. 2020d, 2021b) also outperform traditional optimization-based 3D MRI image reconstruction (Liu et al. 2012) that requires hand-crafted regularizers or priors.

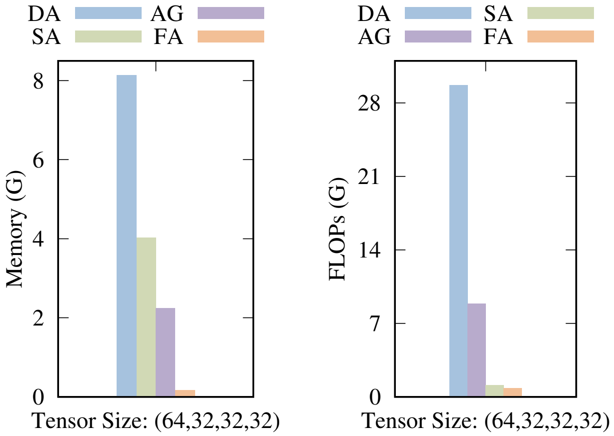


Figure 1: Computational comparison of GPU memory and floating point number operations per second (FLOPs) between four different attention approaches: DA (Fu et al. 2019), AG (Oktay et al. 2018), SA (Wang et al. 2018), and our FA. We get all the numbers from a machine with a single Titan Xp GPU. We test each module using a input feature tensor with size ($64 \times 32 \times 32 \times 32$). Our FA module substantially reduced computational cost compared to DA, AG and SA modules (97.9%, 92.5% and 95.8% of GPU memory reduction, and 88.9%, 63.0% and 25.6% of FLOPs reduction).

These CNN models benefit from capturing contextual information that is essential for many computer vision tasks. Traditional models (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015; He et al. 2016; Huang et al. 2017) stack many layers of convolutional operations to capture the global contextual dependency. However, this stacking procedure has three major drawbacks: 1) Too many convolution layers introduce redundant network parameters that can cause unnecessary memory usage and computational overhead, and makes it prone to overfitting (Simonyan and Zisserman 2015; Peng et al. 2017); 2) Network optimization becomes increasingly difficult as the network depth increases (He et al. 2016; Huang et al. 2017);

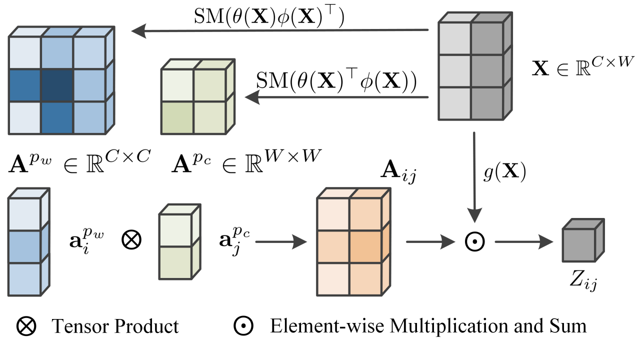

$\mathrm{SM}(\theta(\mathbf{X})\phi(\mathbf{X})^\top)$
$\mathrm{SM}(\theta(\mathbf{X})^\top\phi(\mathbf{X}))$
$\mathbf{X} \in \mathbb{R}^{C\times W}$
$\mathbf{A}^{p_w} \in \mathbb{R}^{C\times C}$
$\mathbf{A}^{p_c} \in \mathbb{R}^{W\times W}$
$\mathbf{A}_{ij}$
$g(\mathbf{X})$
$\mathbf{a}_i^{p_w}$
$\otimes$
$\mathbf{a}_j^{p_c}$
$\odot$
Z_{ij}
$\otimes$ Tensor Product
$\odot$ Element-wise Multiplication and Sum

Figure 2: Example illustration of the regularization of FA.

3) Information propagation among elements with large spatial distances in the feature tensor can be inefficient due to the issue of vanishing gradients (He et al. 2016) and saturated activation units (Ioffe and Szegedy 2015).

The recent attention methods have shed light to the above issues. Self-attention methods (Wang et al. 2018; Zhang et al. 2019; Huang et al. 2019; Fu et al. 2019) aim at capturing long-range dependencies by aggregating contextual information of each pixel from all other pixels in the feature map (pixel here indicates a feature vector of that pixel). Another stream of attention methods (Wang et al. 2017a; Hu, Shen, and Sun 2018; Oktay et al. 2018) focus on creating a mask that can implicitly assist CNN to pay more attention to salient areas. Most of these attention methods operate either on spatial dimensions (Zhang et al. 2019; Wang et al. 2018; Oktay et al. 2018; Wang et al. 2017a), or solely on the channel dimension (Hu, Shen, and Sun 2018), which reduces the performance of feature aggregation. Besides, unlike natural images, processing 3D medical images using CNNs usually demands high GPU memory usage, and most of these methods are not satisfactory due to the computation of huge attention maps. We argue that a unified attention approach that considers both the spatial-channel dependency and the efficiency of computation is of great practical value for modern 3D Medical image tasks. In this paper, we present our folded attention (FA) approach, effective and yet efficient, for modeling the global contextual information with negligible computational cost (see Fig. 1 and Fig. 4).

Our FA approach can be considered as the generalization of original self-attention (SA) mechanism (Wang et al. 2018). The original SA ignores channel-wise dependency and only aggregates information from spatial domain, while in our FA, each element in the output feature tensor is the weighted sum of all elements in the input feature tensor (a pixel is denoted as a vector with multiple elements). Channel information does help the network learn better semantic information (Fu et al. 2019; Hu, Shen, and Sun 2018), but directly applying SA to incorporate spatial and channel information will cause unacceptable GPU memory usage (more details in the methodology section). Though DA network (Fu et al. 2019) combines spatial and channel attention by element-wise sum operation, it suffers from the heavy computational cost. (see Fig. 1) FA module resolves the issue by introducing tensor folding and unfolding operations, where the input feature tensor will be broadcast and unfolded to compute four sub-affinity matrices that can approximate the function of original affinity matrix with cascaded aggregation. (see Fig. 3)

Through the approximation, FA can also be considered as the regularization of the SA mechanism. For simplicity and to be visually interpretable, we use a 1D image represented by a 2D feature tensor to illustrate the concept. As shown in Fig. 2, we use two smaller sub-affinity matrices $\mathbf{A}^{p_w}$ and $\mathbf{A}^{p_c}$ to replace the original element-to-element affinity matrix $\mathbf{A} \in \mathbb{R}^{CW\times CW}$. Let $\mathbf{Z}$ denotes the matrix obtained after FA operation to $\mathbf{X}$ and then $\mathbf{Z}$ can be constructed as follows:

$$\mathbf{A}_{ij} = \mathbf{a}_i^{p_w} \otimes \mathbf{a}_j^{p_c}, \tag{1}$$

$$Z_{ij} = \mathbf{A}_{ij} \odot g(\mathbf{X}), \tag{2}$$

where $\otimes$ is the tensor product, $\odot$ is the element-wise multiplication and sum, $\mathbf{A}_{ij}$ is the affinity matrix of element X_{ij} ($A_{ij,pq}$ denotes the entry at pq of matrix$\mathbf{A}_{ij}$, and is also the affinity between element $\mathbf{X}_{ij}$ and $\mathbf{X}_{pq}$), and $\mathbf{a}_i^{p_w}$ and $\mathbf{a}_j^{p_c}$ denote the transpose of i_{th} and j_{th} row of matrix $\mathbf{A}^{p_w}$ and $\mathbf{A}^{p_c}$ respectively. It is obvious that $\mathbf{A}_{ij}$ is a rank-one matrix thus imposing regularization on the original affinity matrix.

Our FA approach can be applied to many other 3D image analysis tasks due to its efficiency and simplicity, and in this paper, we demonstrate the performance on two challenging tasks in 3D Medical images. One task is quantitative susceptibility mapping (QSM) (de Rochefort et al. 2010; Wang and Liu 2015), a functional image mapping task, which enables studying tissue magnetic susceptibility properties (Wang et al. 2017b). This reconstruction problem is challenging due to the ill-posedness of dipole inversion. For deep learning based QSM reconstruction, training data can only be obtained with COSMOS (Liu et al. 2009) that is regarded as a reference standard, but only very limited data samples are available (Yoon et al. 2018; Zhang et al. 2020b). Another task is multiple sclerosis (MS) lesion segmentation, a image semantic segmentation task. Unlike tumor or other organ segmentation problems, MS lesion segmentation is more difficult (Zhang et al. 2020a) as lesions vary enormously in terms of size, shape, and location.

Related Works

The attention concept was first introduced in neural machine translation (Bahdanau, Cho, and Bengio 2014; Luong, Pham, and Manning 2015) to improve the performance of recurrent neural networks (RNN) by capturing dependencies between long-range words in a sentence. Later, RNN was entirely replaced with self-attention operations by transformer (Vaswani et al. 2017). Further, attention mechanism has then been widely adopted in vision tasks, such as image recognition (Wang et al. 2017a; Hu, Shen, and Sun 2018) and image segmentation (Zhang et al. 2019; Fu et al. 2019). In general, most of these methods can be divided into two types: mask-based attention (MA) that learns a salience feature map and self-attention (SA) that learns feature aggregation. MA methods usually generate a mask that emphasizes the importance or saliency on a portion of the feature tensor,

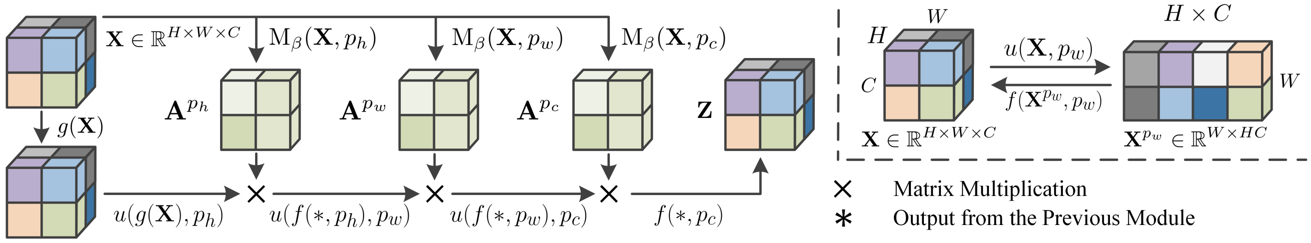

$\mathbf{X} \in \mathbb{R}^{H\times W\times C}$
$\mathrm{M}_\beta(\mathbf{X}, p_h)$
$\mathrm{M}_\beta(\mathbf{X}, p_w)$
$\mathrm{M}_\beta(\mathbf{X}, p_c)$
$g(\mathbf{X})$
$\mathbf{A}^{p_h}$
$\mathbf{A}^{p_w}$
$\mathbf{A}^{p_c}$
$\mathbf{Z}$
$u(g(\mathbf{X}), p_h)$
$u(f(*, p_h), p_w)$
$u(f(*, p_w), p_c)$
$f(*, p_c)$
W
H
C
$\mathbf{X} \in \mathbb{R}^{H\times W\times C}$
$u(\mathbf{X}, p_w)$
$f(\mathbf{X}^{p_w}, p_w)$
$H \times C$
W
$\mathbf{X}^{p_w} \in \mathbb{R}^{W\times HC}$
× Matrix Multiplication
∗ Output from the Previous Module

Figure 3: The left panel is the overall FA pipeline and the upper right panel is the visualization of fold and unfold operations. For simplicity and visualization-friendly, we use a 3D input tensor to illustrate, but 4D or higher dimensional tensors can be easily extended. In the left panel, $P_h = (0, 1, 2)$, $P_w = (1, 0, 2)$ and $P_c = (3, 0, 1)$; we first compute three sub-affinity matrices $\mathbf{A}^{p_h}$, $\mathbf{A}^{p_w}$, and $\mathbf{A}^{p_c}$; With these sub-affinity matrices, we then use three consecutive unfolding-and-folding steps to perform the feature aggregation and get output $\mathbf{Z}$. In the right panel, we show how function f and u works; each element of the tensor is marked with a different color, and the color remains its position after folding or unfolding operations. (best view in color)

either spatial-wise (Wang et al. 2017a; Oktay et al. 2018) or channel-wise (Hu, Shen, and Sun 2018). Though AG-Net (Oktay et al. 2018) improves by using grid-based gating scheme, MA methods is not suitable for image-to-image functional mapping tasks as any pixel matters and salient area is unnecessary. SA methods produce a function that pass through a feature map without any modification of the input size, and features either from spatial locations (Zhang et al. 2019; Wang et al. 2018) or channel maps (Fu et al. 2019) are aggregated during the pass, where each element is replaced with a weighted sum of features from some of other elements. SA methods raise memory issue in 3D medical images as it needs to compute huge attention maps (See DA and SA in Fig. 1). Though RSA-Net (Zhang et al. 2019) solves the memory issue by iterative feature aggregation, it ignores the channel information aggregation.

Contributions

In this paper, we propose a novel FA approach that can efficiently capture global contextual dependencies with negligible computational cost. We exploit the superiority of FA in QSM reconstruction and MS lesion segmentation tasks, and the contributions of FA can be summarized as follows:

- We propose a folded attention approach that can improve the performance of general 3D medical image tasks by global contextual information aggregation, and our method can tremendously reduce the computational cost of GPU memory (at least 95.8%) and FLOPs (at least 25.6%) compared to most existing attention approaches.
- Extensive experimental results from both a semantic segmentation task and a functional image mapping task on 3D medical images show the effectiveness and the efficiency of our method. By insertion of our FA module, with negligible cost, we outperform all other attention methods, and improve the baseline Dice metric of MS lesion segmentation by 3% and the baseline RMSE metric of QSM reconstruction by 3%.

Methodology

In this section, we will present details of the proposed folded attention (FA) approach. We will first review traditional SA mechanism and its simple generalization to channel dimension. We then illustrate how our FA approach can generalize and regularize the SA mechanism. Complexity analysis of memory and computational cost on FA will be discussed.

Self Attention Mechanism

In this paper, we adopt a widely used instantiation of SA as the rest shares similar performance (Wang et al. 2018). The adopted embedded Gaussian SA can be described as follows:

$$\mathbf{A} = \mathrm{SM}(\theta(\mathbf{X})\phi(\mathbf{X})^\top), \tag{3}$$

$$\mathbf{Z} = \mathbf{A}g(\mathbf{X}), \tag{4}$$

where SM is the Softmax function along each matrix row, $\mathbf{X} \in \mathbb{R}^{N\times C}$ is the input feature tensor, $\mathbf{A} \in \mathbb{R}^{N\times N}$ is the affinity matrix, $\mathbf{Z} \in \mathbb{R}^{N\times C}$ is the output feature tensor of SA, and $N = HWD$ is the number of pixels in the image. Function θ and ϕ are single-layer perceptrons that can linearly transform features of $\mathbf{X}$ to facilitate the computation of affinity matrix. The inner product between θ and ϕ computes the pixel-to-pixel affinity. Function g is also a single-layer perceptron that can help the network to learn a better feature embedding.

Generalization of SA to Channel Dimension

DA-Net (Fu et al. 2019) uses the element-wise sum of the outputs of spatial attention and channel attention to approximate spatial-channel attention. However, separate operations on spatial and channel dimensions are prone to be sub-optimal. One natural idea to generalize the original SA is to replace $\mathbf{X} \in \mathbb{R}^{N\times C}$ as $\hat{\mathbf{X}} \in \mathbb{R}^{NC\times 1}$, where element-to-element instead of pixel-to-pixel affinity matrix can be obtained by $\hat{\mathbf{A}} \in \mathbb{R}^{NC\times NC}$. Unfortunately, the matrix $\hat{\mathbf{A}}$ is too huge for modern commercial GPU to process. (According to our experiments, $\hat{\mathbf{A}}$ may consume several hundred Gigabytes memory on our 3D image tasks.) It is obvious that direct computation of such huge matrix $\hat{\mathbf{A}}$ is not realistic, thus we propose our FA to ease the problem.

Folded Attention (FA)

The FA approach can relieve the above issue by considering spatial-channel attention in a single module with negligible additional computational resources. Next, we will introduce modules of our FA approach, including the folding and unfolding operations and sub-affinity matrix computation, and feature aggregation.

Fold and Unfold Operations Let $\mathbf{X} \in \mathbb{R}^{H\times W\times D\times C}$, where H, W, D are sizes of three spatial dimensions of the feature tensor $\mathbf{X}$ and C is the number of channels. We define an unfold function u, where $u(\mathbf{X}, p) = \mathbf{X}^p$, and P is a permutation that indicates how to unfold the tensor. Here we use an example to illustrate the function u. Let $p = (1, 0, 2, 3)$, we can get $u(\mathbf{X}, p) = \mathbf{X}^p \in \mathbb{R}^{W\times HDC}$, where u first permutes the four dimensions of $\mathbf{X}$ according to p, and then f unfolds the last three dimensions into one dimension, resulting in a 2D matrix $\mathbf{X}^p$. Also, we define a function f as the inverse operation of u, where $f(\mathbf{X}^p, p) = \mathbf{X}$. For simplicity, we further set four permutation vectors as $p_h = (0, 1, 2, 3)$, $p_w = (1, 0, 2, 3)$, $p_d = (2, 0, 1, 3)$, and $p_c = (3, 0, 1, 2)$.

Sub-Affinity Matrix The generalization of SA with channel attention requires the computation of a huge affinity matrix $\hat{\mathbf{A}}$, which suffers from heavy memory cost. In our proposed FA approach, we use four sub-affinity matrices to replace the huge one. We denote the four matrices as $\mathbf{A}^{p_h}$, $\mathbf{A}^{p_w}$, $\mathbf{A}^{p_d}$, and $\mathbf{A}^{p_c}$, where $p_h, ..., p_c$ are the permutation vectors defined in the last section. The sub-affinity matrix can be computed as follows:

$$\mathbf{A}^p = \mathrm{SM}(u(\theta(\mathbf{X}), p)u(\phi(\mathbf{X}), p)^\top). \tag{5}$$

The size of each sub-affinity matrix $\mathbf{A}^p$ is much smaller than the original affinity matrix $\hat{\mathbf{A}}$. Even the sum of the sizes of all four sub-affinity matrices is several orders of magnitude smaller than $\hat{\mathbf{A}}$. (see more details in complexity analysis)

Feature Aggregation The next step after obtaining affinity matrix is to aggregate features from the original feature tensor $\mathbf{X}$. Suppose we have obtained a sub-affinity matrix $\mathbf{A}^p$ from Eq. (5), the feature aggregation based on the sub-affinity matrix can be described as follows:

$$\mathbf{Z} = f(\mathbf{A}^p u(g(\mathbf{X}), p), p). \tag{6}$$

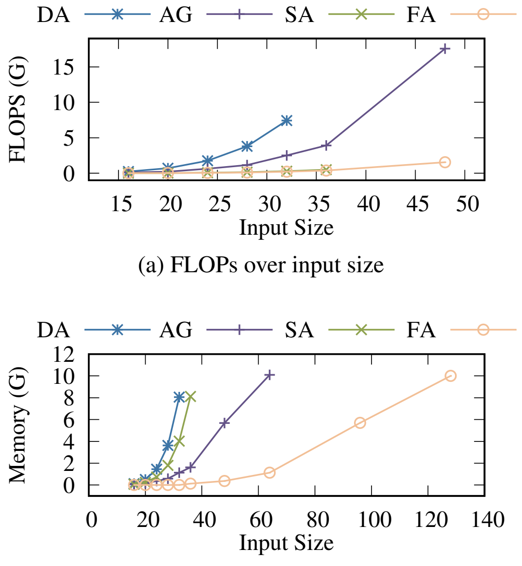
For simplicity, we denote Eq. (6) as: $\mathbf{Z} = \mathrm{U}_\gamma(g(\mathbf{X}), p)$, where γ represents the parameters of the function g. Also, Eq. (5) can be simplified as $\mathrm{M}_\beta(\mathbf{X}, p)$, where β denotes the parameters of function θ and ϕ. We can then get our four sub-affinity matrices by $\mathbf{A}^{p_h} = \mathrm{M}_\beta(\mathbf{X}, p_h)$, $\mathbf{A}^{p_w} = \mathrm{M}_\beta(\mathbf{X}, p_w)$, $\mathbf{A}^{p_d} = \mathrm{M}_\beta(\mathbf{X}, p_d)$, and $\mathbf{A}^{p_c} = \mathrm{M}_\beta(\mathbf{X}, p_c)$. Now our proposed FA operation is derived as follows:

$$\mathbf{Z} = \mathrm{U}_\gamma(\mathrm{U}_\gamma(\mathrm{U}_\gamma(\mathrm{U}_\gamma(\mathbf{X}, p_h), p_w), p_d), p_c) \tag{7}$$

Rank-One Constraint For any input feature tensor $\mathbf{X}$, we can compute four sub-affinity matrices $\mathbf{A}^{p_h} \in \mathbb{R}^{H\times H}$, $\mathbf{A}^{p_w} \in \mathbb{R}^{W\times W}$, $\mathbf{A}^{p_d} \in \mathbb{R}^{D\times D}$, and $\mathbf{A}^{p_c} \in \mathbb{R}^{C\times C}$. Let $\mathbf{A}^p_i$ denotes the transpose of the i_{th} row of the matrix $\mathbf{A}^p$. We further define $\mathbf{A}_v \in \mathbb{R}^{H\times W\times D\times C}$ as follows:

$$\mathbf{A}_v = \mathbf{A}^{p_h}_i \otimes \mathbf{A}^{p_w}_j \otimes \mathbf{A}^{p_d}_k \otimes \mathbf{A}^{p_c}_q, \tag{8}$$

where $v = (i, j, k, q)$ denotes the position of an element in the feature tensor and $\otimes$ is the tensor product; $\mathbf{A}_v$ is the affinity tensor of element $\mathbf{x}_v$ that shares the same size as input feature tensor, and all elements of the input feature tensor have their own affinity tensors. The original affinity matrix $\hat{\mathbf{A}} \in \mathbb{R}^{NC\times NC}$ can be reconstructed using $\{\mathbf{A}_v | v \in \Omega\}$, where Ω enumerates all possible element positions. We can further derive $\mathbf{Z}_v$ as follows:



(a) FLOPs over input size

(b) Memory over input size

Figure 4: The x-axis denotes the height, width, depth or number of channels in a tensor, and we assume all these four scalars have equal values. The y-axis represents the computational cost measured by FLOPs and GPU memory consumption for (a) and (b) respectively. All numbers are obtained with a Titan XP GPU.

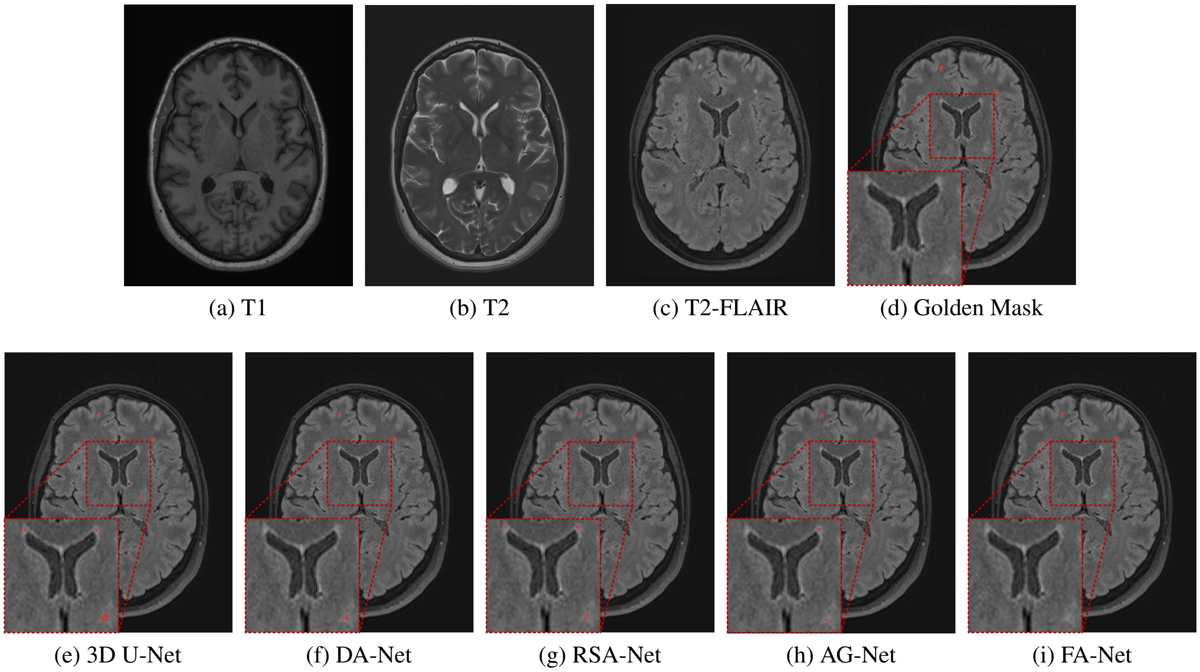
$$\mathbf{Z}_v = \mathbf{A}_v \odot g(\mathbf{X}) \tag{9}$$

We have derived our full FA operation in Eq. 7, It is easy to understand that enumerating all element positions using Eq. 8 and Eq. 9 can get the same result as using Eq. (7). However, using Eq. 7 with a cascaded process can tremendously save the computational cost of GPU memory by taking advantages of replacing the original dense affinity matrix with four smaller sub-affinity matrices. Since $\mathbf{A}_v$ is a rank-one tensor, FA can be considered as imposing an explicit low-rank constraint on the affinity tensor of each element.

Complexity Analysis

Given the input feature tensor of size (H, W, D, C), we analyze the computational complexity of the proposed FA approach. Let N, M denotes the product and the sum of H, W, D, C, we can obtain the complexity of our FA as $\mathcal{O}(NC + NM)$.

We then numerically compare the computational complexity and GPU memory consumption of FA with other three approaches in Fig. 4. It can be seen that, comparing with DA (Fu et al. 2019) that considers both spatial and channel attention, our FA is much more computational-efficient and GPU memory-friendly. Though numerically,


(a) T1 (b) T2 (c) T2-FLAIR (d) Golden Mask

(e) 3D U-Net (f) DA-Net (g) RSA-Net (h) AG-Net (i) FA-Net

Figure 5: Example MS lesion segmentation results. T1, T2, T2-FLAIR images and the corresponding golden mask are shown in the first line. Segmentations from 3D U-Net, DA-Net, RSA-Net, AG-Net, and FA-Net are shown in the second line.

FA has little improvement of FLOPs over SA (Wang et al. 2018), it reduces the GPU memory usage dramatically and in the meanwhile incorporates channel attention. Comparing with FA, AG (Oktay et al. 2018) has its limitation in 3D medical image applications as it still requires large memory for computation and suffers from scaling.

Experimental Results

We use PyTorch (Paszke et al. 2019) for all of our implementations. We compare our models with several recent state-of-the-art attention approaches, including baseline 3D U-Net (Çiçek et al. 2016), dual attention (DA) (Fu et al. 2019), recurrent slice-wise attention (RSA) (Zhang et al. 2019) and attention gated (AG) net (Oktay et al. 2018). For fairness, we adopt methods from their open-source implementations and do our best to adjust their parameters to achieve the best performance. Particularly, DA is originally designed for 2D images, so it is modified and adjusted to be capable of processing 3D MR images. All models in the experiments are trained in a machine with a Titan Xp GPU. Our implementation is made publicly available [1].

Multiple Sclerosis (MS) Lesion Segmentation

We conduct our first experiment on MS lesion segmentation, a high-level segmentation task. MS is a chronic, inflammatory demyelinating disease of central nervous system in the brain. Precise lesion tracing can provide important biomarkers for clinical diagnosis and disease progress assessment. However, MS lesion segmentation is challenging as lesions vary vastly in terms of location, appearance, shape, and conspicuity (see Fig. 5 for more details).

We use a dataset with 30 MR images acquired from a 3.0 T GE scanner. Images from T1, T2, and T2-FLAIR sequences are collected, and each voxel size is $0.7 \times 0.7 \times 3.0mm^3$. Golden masks are traced by a neural radiologist with over 8 years' lesion tracing experience. Images are linearly co-registered using FLIRT at FSL (Jenkinson et al. 2012) neuroimaging toolbox. All images are normalized to zero-mean with a unit-variance during the pre-processing step.

Implementation Details We perform five random splits on the dataset, where each split contains 15, 5, and 10 subjects for training, validation, and testing. A Model that achieves the minimum loss on the validation set will be used for testing. We perform random crop with fixed cropping size (128 × 160 × 32), and use elastic deformation, intensity shifting for data augmentation. We adopt the sum of weighted cross entropy and soft dice (Dice 1945) as our loss function. Adam (Kingma and Ba 2014) with the initial learning rate of $1e{-}3$ and a multi-step learning rate scheduler with milestones at 50%, 70% and 90% of the total epochs are used for optimal convergence. A batch size of four is used for training, and training would stop after 120 epochs.

Dice score (DSC), lesion-wise true positive rate (LTPR), lesion-wise positive predicted value (LPPV), and lesion-wise F1 score (L-F1) are used for evaluations. LTPR and LPPV are defined as $\text{LTPR} = \frac{\text{TPR}}{\text{GL}}, \text{LPPV} = \frac{\text{TPR}}{\text{PL}}$, where TPR denotes the number of lesions in the Golden segmentation that overlaps with a lesion in the produced segmen-

[1] https://github.com/tinymilky/FANet

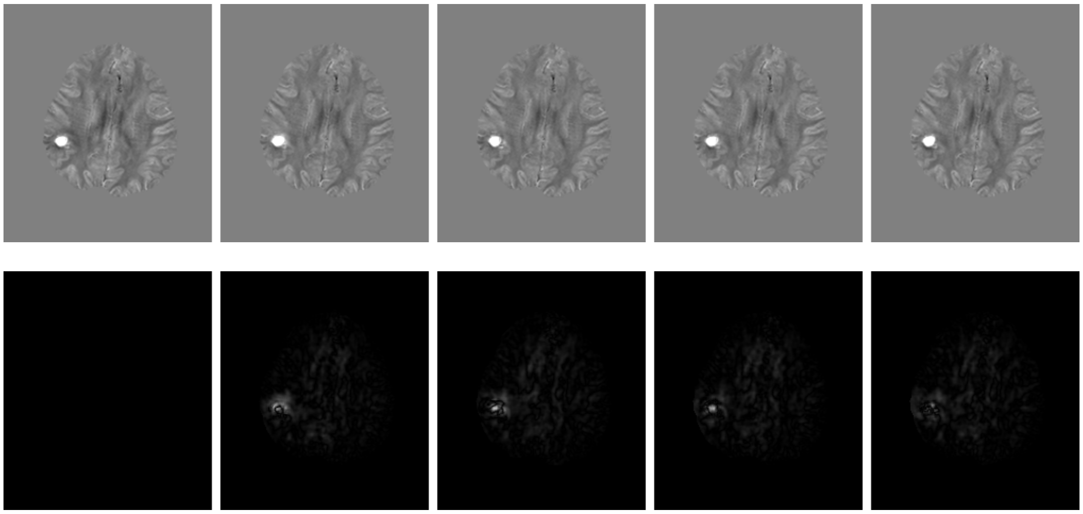


Figure 6: Example QSM reconstruction results (upper line with window level: [-0.15, 0.15] ppb) and absolute error maps (lower line with window level: [0, 0.05] ppb) on one test case with COSMOS label. From left to right are COSMOS (golden ground-truth), predictions of QSMnet, DA-Net, RSA-Net, and FA-Net.

tation, and GL, PL is the number of lesions in ground-truth segmentation and produced segmentation respectively. L-F1 can be obtained from LTPR and LPPV as L-F1 $= 2\frac{\text{LTPR} \cdot \text{LPPV}}{\text{LTPR} + \text{LPPV}}$.

Quantitative Results We use DA-Net, RSA-Net, and FA-Net to denote a backbone 3D U-Net with the corresponding attention module inserted at the bottom layer of 3D U-Net. Specifically, AG-Net inserts three attention modules according to the literature (Oktay et al. 2018). As shown in Table 1, all attention methods outperform 3D U-Net backbone network in all metrics by a significant margin. RSA-Net and AG-Net have no clues about dependencies or salience of channels; thus, we can see from the table that our FA-Net outperform them in both DSC and L-F1 metrics; Though RSA-Net obtains similar LTPR as our FA-Net, it falls behind a lot in LPPV. Though DA-Net considers both spatial and channel attention and our FA-Net has only marginal improvement compared to DA-Net, incorporation of our FA module consumes negligible additional GPU memory and FLOPs (See Fig. 1 and Fig. 4).

Qualitative Results We showcase one slice from a testing subject, and compare the qualitative results of different models with the golden mask. We can see from Fig. 5 that besides MS lesions, there still exists many other concurrent hyper-intensities in the T2-FLAIR image. Particularly, the hyper-intensities near the lateral ventricles are prone to be over-segmented. This is because some hyper-intensities near ventricles are MS lesions, but some are not, depending on their anatomical and surrounding structures. We can see that all attention models help ease the over-segmenting problem in some degree. DA-Net and our FA-Net perform the best as these two models both consider the dependencies of spatial and channel dimensions.

Method	DSC	LPPV	LTPR	L-F1
3D U-Net	0.667	0.682	0.838	0.752
DA-Net	0.682	0.689	**0.871**	0.770
RSA-Net	0.677	0.678	0.870	0.762
AG-Net	0.682	0.702	0.830	0.761
FA-Net (ours)	**0.684**	**0.703**	0.867	**0.776**

Table 1: Quantitative comparison of MS lesion segmentation with different approaches.

Quantitative Susceptibility Mapping (QSM)

We conduct our second experiment on a challenging image reconstruction problem in MRI: quantitative susceptibility mapping (QSM) (de Rochefort et al. 2010; Wang and Liu 2015). QSM can measure the underlying tissue apparent magnetic susceptibility, which can be used to quantify specific bio-markers such as iron that is independent of imaging parameters (Stüber, Pitt, and Wang 2016; Kirui et al. 2013), and filed strength (Deh et al. 2015). The forward model of generating magnetic field from susceptibility map with additive noise is a 3D spatial convolutional process and can be described as following:

$$b = \chi * d + n, \tag{10}$$

where b is the magnetic field, χ is the tissue susceptibility, d is the dipole convolution kernel, and n is the additive measurement noise. The aim of QSM is to solve the deconvolutional problem from measured noisy magnetic field b to tissue susceptibility χ. This is intrinsically an ill-posed inverse problem due to the zero cone surfaces of the dipole kernel in k-space (Wang and Liu 2015). To tackle the ill-posedness, COSMOS (Calculation Of Susceptibility through Multiple Orientation Sampling) (Liu et al. 2009) reconstruction is proposed to eliminate all zeros in the k-space cone surface by multiple orientation scans, thereby serving as the reference standard for further susceptibility analysis.

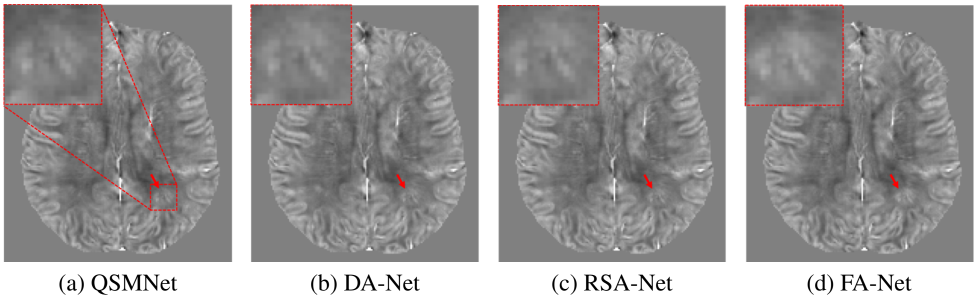

(a) QSMNet (b) DA-Net (c) RSA-Net (d) FA-Net

Figure 7: Example QSM reconstruction results (window level: [-0.15, 0.15] ppb) on a subject with MS lesions (patient data without ground-truth COSMOS). Hyperintense MS lesions are pointed out by red arrows.

Recently, several deep learning based QSM reconstruction methods (Yoon et al. 2018; Zhang et al. 2020b,c; Chen et al. 2020) have been developed with promising results. They use 3D U-Net as the backbone network to perform the functional mapping from the magnetic field input to susceptibility output. In this experiment, we follow previous work and use COSMOS data to train our deep networks. To acquire and reconstruct COSMOS data, 6 healthy subjects were recruited to do MRI scan with 5 brain orientations using a 3.0T GE scanner (Please note that COSMOS technique cannot be applied to patients as it needs four additional head orientations that are very difficult or impossible for patients to perform). Acquisition matrix was $256 \times 256 \times 48$ and voxel size was $1 \times 1 \times 3$ mm^3. Reference tissue was reconstructed from five orientations of each subject with local magnetic field estimated from phase data (Liu et al. 2011).

Implementation Details We perform six splits on the dataset, where each split contains 4, 1, and 1 subjects(s) for training, validation, and testing, and each subject contains 5 volumes. During training, we cropped each volume into 3D patches in size ($64 \times 64 \times 32$) and use in-plane rotation of $\pm 15^\circ$ for data augmentation. Loss function from QSMnet (Yoon et al. 2018) is adopted. Adam (Kingma and Ba 2014) optimizer is used for training with the same hyperparameters as MS lesion segmentation experiment. Training is performed with a batch size of 16 and training would stop after 60 epochs . During testing, a model with the best validation loss is used to evaluate the performance. In addition, a patient subject with MS lesion is also used to qualitatively verify the performance of our networks. (Note that a patient subject does not have the COSMOS ground-truth) Different from MS lesion segmentation, we use QSM-Net (Yoon et al. 2018), a modified U-Net, as our backbone network. We use DA-Net, RSA-Net, and FA-Net to denote a QSM-Net with the corresponding attention module inserted at its bottom layer. AG-Net is excluded in the QSM experiment as it is unfair to compare MA based methods with SA based methods in a full functional image mapping task.

Method	RMSE	HFEN	SSIM	PSNR
QSMnet	31.99	33.37	0.9824	48.86
DA-Net	32.15	33.84	0.9826	48.78
RSA-Net	31.65	33.18	0.9830	48.91
FA-Net (Ours)	**31.18**	**32.49**	**0.9833**	**49.06**

Table 2: Quantitative comparison of QSM.

Quantitative Results We use root mean square error (RMSE), peak signal-to-noise ratio (PSNR) (measures general reconstruction error), high-frequency error norm (HFEN) (measures the similarity at high spatial frequencies), and structural similarity index (SSIM) (quantifies image contrast, intensity, structural similarity between image pairs (Wang et al. 2004)) to quantify the reconstruction accuracy. Quantitative results averaged among six splits are shown in Table 2, and we can see that our FA-Net shows the best reconstruction results in all four metrics.

Qualitative Results We choose one slice from the testing image of one split, and the chosen subject is diagnosed as cerebral hemorrhage (hyper-intensity tissue area in Fig. 6); however, the hemorrhage situation is not covered in the training data. As we can see from Fig. 6, the error map from our FA-Net achieves the minimum intensity which shows the robustness of our FA-Net compared to others.

We use an additional MS lesion subject without ground-truth COSMOS to compare the reconstruction performance among four trained networks in Fig. 7. As can be seen from Fig. 7, on one hand, our FA-Net generated the most hyperintense lesions, and on the other hand, the lesion shows clearer boundary in FA-Net produced image compared to others. The superiority of our FA-Net is that it aggregates features from both spatial and channel dimensions, and in the meanwhile, it regularizes the dense affinity matrix with rank-one constraint and thus generalizes better to unseen situations.

Conclusions

We presented a novel folded attention module. Our FA module exploits the spatial-channel correlations in an efficient and effective way. FA not only achieves the highest accuracy on MS lesion segmentation and QSM reconstruction among all state-of-the-art attention methods, but also reduces tremendously the computational overhead and memory usage. Our method can be easily plugged into any existing CNN model with negligible cost, thereby serving as a new baseline for general 3D MR image processing.

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Chen, Y.; Jakary, A.; Avadiappan, S.; Hess, C. P.; and Lupo, J. M. 2020. Qsmgan: improved quantitative susceptibility mapping using 3d generative adversarial networks with increased receptive field. *NeuroImage* 207: 116389.

Çiçek, Ö.; Abdulkadir, A.; Lienkamp, S. S.; Brox, T.; and Ronneberger, O. 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, 424–432. Springer.

de Rochefort, L.; Liu, T.; Kressler, B.; Liu, J.; Spincemaille, P.; Lebon, V.; Wu, J.; and Wang, Y. 2010. Quantitative susceptibility map reconstruction from MR phase data using bayesian regularization: validation and application to brain imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 63(1): 194–206.

Deh, K.; Nguyen, T. D.; Eskreis-Winkler, S.; Prince, M. R.; Spincemaille, P.; Gauthier, S.; Kovanlikaya, I.; Zhang, Y.; and Wang, Y. 2015. Reproducibility of quantitative susceptibility mapping in the brain at two field strengths from two vendors. *Journal of magnetic resonance imaging* 42(6): 1592–1600.

Dice, L. R. 1945. Measures of the amount of ecologic association between species. *Ecology* 26(3): 297–302.

Dong, C.; Loy, C. C.; He, K.; and Tang, X. 2014. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, 184–199. Springer.

Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; and Lu, H. 2019. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3146–3154.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; and Liu, W. 2019. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, 603–612.

Ioffe, S.; and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167* .

Jenkinson, M.; Beckmann, C. F.; Behrens, T. E.; Woolrich, M. W.; and Smith, S. M. 2012. Fsl. *Neuroimage* 62(2): 782–790.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kirui, D. K.; Khalidov, I.; Wang, Y.; and Batt, C. A. 2013. Targeted near-IR hybrid magnetic nanoparticles for in vivo cancer therapy and imaging. *Nanomedicine: Nanotechnology, Biology and Medicine* 9(5): 702–711.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Law, H.; and Deng, J. 2018. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 734–750.

Liu, J.; Liu, T.; de Rochefort, L.; Ledoux, J.; Khalidov, I.; Chen, W.; Tsiouris, A. J.; Wisnieff, C.; Spincemaille, P.; Prince, M. R.; et al. 2012. Morphology enabled dipole inversion for quantitative susceptibility mapping using structural consistency between the magnitude image and the susceptibility map. *Neuroimage* 59(3): 2560–2568.

Liu, T.; Khalidov, I.; de Rochefort, L.; Spincemaille, P.; Liu, J.; Tsiouris, A. J.; and Wang, Y. 2011. A novel background field removal method for MRI using projection onto dipole fields. *NMR in Biomedicine* 24(9): 1129–1136.

Liu, T.; Spincemaille, P.; De Rochefort, L.; Kressler, B.; and Wang, Y. 2009. Calculation of susceptibility through multiple orientation sampling (COSMOS): a method for conditioning the inverse problem from measured magnetic field map to susceptibility source image in MRI. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 61(1): 196–204.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* .

Myronenko, A. 2018. 3D MRI brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, 311–320. Springer.

Oktay, O.; Schlemper, J.; Folgoc, L. L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N. Y.; Kainz, B.; et al. 2018. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999* .

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 8024–8035.

Peng, C.; Zhang, X.; Yu, G.; Luo, G.; and Sun, J. 2017. Large kernel matters–improve semantic segmentation by global convolutional network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4353–4361.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 91–99.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Stüber, C.; Pitt, D.; and Wang, Y. 2016. Iron in multiple sclerosis and its noninvasive imaging with quantitative susceptibility mapping. *International journal of molecular sciences* 17(1): 100.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; and Tang, X. 2017a. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3156–3164.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Nonlocal neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wang, Y.; and Liu, T. 2015. Quantitative susceptibility mapping (QSM): decoding MRI data for a tissue magnetic biomarker. *Magnetic resonance in medicine* 73(1): 82–101.

Wang, Y.; Spincemaille, P.; Liu, Z.; Dimov, A.; Deh, K.; Li, J.; Zhang, Y.; Yao, Y.; Gillen, K. M.; Wilman, A. H.; et al. 2017b. Clinical quantitative susceptibility mapping (QSM): Biometal imaging and its emerging roles in patient care. *Journal of magnetic resonance imaging* 46(4): 951–971.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4): 600–612.

Yang, W.; Tan, R. T.; Feng, J.; Liu, J.; Guo, Z.; and Yan, S. 2017. Deep joint rain detection and removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1357–1366.

Yoon, J.; Gong, E.; Chatnuntawech, I.; Bilgic, B.; Lee, J.; Jung, W.; Ko, J.; Jung, H.; Setsompop, K.; Zaharchuk, G.; et al. 2018. Quantitative susceptibility mapping using deep neural network: QSMnet. *Neuroimage* 179: 199–206.

Zhang, H.; Wang, R.; Zhang, J.; Li, C.; Yang, G.; Spincemaille, P.; Nguyen, T.; and Wang, Y. 2021a. NeRD: Neural Representation of Distribution for Medical Image Segmentation. *arXiv preprint arXiv:2103.04020* .

Zhang, H.; Zhang, J.; Wang, R.; Zhang, Q.; Gauthier, S. A.; Spincemaille, P.; Nguyen, T. D.; and Wang, Y. 2020a. Geometric Loss for Deep Multiple Sclerosis lesion Segmentation. *arXiv preprint arXiv:2009.13755* .

Zhang, H.; Zhang, J.; Zhang, Q.; Kim, J.; Zhang, S.; Gauthier, S. A.; Spincemaille, P.; Nguyen, T. D.; Sabuncu, M.; and Wang, Y. 2019. RSANet: Recurrent Slice-Wise Attention Network for Multiple Sclerosis Lesion Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 411–419. Springer.

Zhang, J.; Liu, Z.; Zhang, S.; Zhang, H.; Spincemaille, P.; Nguyen, T. D.; Sabuncu, M. R.; and Wang, Y. 2020b. Fidelity imposed network edit (FINE) for solving ill-posed image reconstruction. *NeuroImage* 116579.

Zhang, J.; Zhang, H.; Li, C.; Spincemaille, P.; Sabuncu, M.; Nguyen, T. D.; and Wang, Y. 2021b. Temporal Feature Fusion with Sampling Pattern Optimization for Multi-echo Gradient Echo Acquisition and Image Reconstruction. *arXiv preprint arXiv:2103.05878* .

Zhang, J.; Zhang, H.; Sabuncu, M.; Spincemaille, P.; Nguyen, T.; and Wang, Y. 2020c. Bayesian Learning of Probabilistic Dipole Inversion for Quantitative Susceptibility Mapping. *arXiv preprint arXiv:2004.12573* .

Zhang, J.; Zhang, H.; Wang, A.; Zhang, Q.; Sabuncu, M.; Spincemaille, P.; Nguyen, T. D.; and Wang, Y. 2020d. Extending LOUPE for K-space Under-sampling Pattern Optimization in Multi-coil MRI. *arXiv preprint arXiv:2007.14450* .